南京信息工程大学

Nanjing University of Information Science and Technology

Nanjing, Jiangsu, 210044, China

Deep Learning Final Report

on

"Deep Learning for Pediatric Wrist Fracture Detection Using the YOLOv8 Algorithm"

Name: HAQUE MD IMAMUL
ID: 202352620008
Major: Artificial Intelligence

Submitted To

TANG Ling
Professor
School of Artificial Intelligence

03/06/2024

Abstract:

Wrist fractures are common, particularly among children who experience a high incidence of such injuries. Before surgery, X-ray imaging is typically required, with treatment plans informed by radiologist analysis. The rapid advancement of neural networks has led to the widespread use of the You Only Look Once (YOLO) series in fracture detection as part of computer-assisted diagnosis (CAD) systems. In 2023, Ultralytics released the newest YOLO model variant, designed for detecting fractures in various body parts. A key innovation in enhancing model performance is the use of attention mechanisms. This study introduces the YOLOv8-AM, which integrates attention mechanisms into the YOLOv8 framework. Four specific attention modules—Convolutional Block Attention Module (CBAM), Global Attention Mechanism (GAM), Efficient Channel Attention (ECA), and Shuffle Attention (SA)—are incorporated to develop improved versions of the model. These versions were trained on the GRAZPEDWRI-DX dataset. Results show that the YOLOv8-AM model with ResCBAM achieved a state-of-the-art mean Average Precision at IoU 50 (mAP 50) of 65.8%, up from 63.6%. Meanwhile, the model with GAM achieved only a marginal increase of 64.2%, leading to the development of ResGAM, which improved mAP 50 to 65.0%.

Keywords: Deep Learning, Fracture Detection, Object Detection, Medical Image Processing, Radiology, X-ray Imaging.

摘要:

手腕骨折很常见，尤其是在此类损伤发生率很高的儿童中。手术前，通常需要进行 X 射线成像，并通过放射科医生的分析告知治疗计划。神经网络的快速发展导致 YOLO 系列作为计算机辅助诊断系统的一部分在骨折检测中广泛使用。2023 年，Ultralytics 发布了最新的 YOLO 模型变体，旨在检测身体各个部位的骨折。提高模型性能的一个关键创新是使用注意力机制。本研究介绍了 YOLOv8 AM，它将注意力机制集成到 YOLOv8AM 框架中。四个特定的注意力模块——卷积块注意力模块（CBAM）、全局注意力机制（GAM）、有效通道注意力（ECA）和无序注意力（SA）——被纳入到该模型的改进版本中。这些版本是在 GRAZPEDWRI-DX 数据集上训练的。结果显示，具有 ResCBAM 的 YOLOv8 AM 模型在 IoU 50（mAP 50）时实现了最先进的平均精度 65.8%，高于 63.6%。同时，带有 GAM 的模型仅实现了 64.2%的边际增长，导致了 ResGAM 的开发，将 mAP50 提高到 65.0%。

关键词：深度学习，骨折检测，目标检测，医学图像处理，放射学，X 射线成像。

Table of Contents

Abstract: i
1 Introduction 1
2 Related Work 3
2.1 Fracture Detection 3
2.2 Attention Module 4
3 Methodology 5
3.1 Baseline Model 5
3.2 Proposed Method 6
3.3 Attention Modules 7
4 Experiment Results and Analysis 10
4.1 Dataset 10
4.2 Preprocessing and Data Augmentation 11
4.3 Evaluation Metric 11
4.4 Experiment Setup 12
4.5 Experimental Results 12
5 Discussion 16
6 Conclusion 17
6.1 Future Work 17
References: 18

List of Figures:

Figure 1 Sample Input and Ground Truth Images from the Study. 2
Figure 2 Overview of the YOLOv8-AM Model Architecture with.....2
Figure 3 Architecture of YOLOv8-AM showing attention modules: Shuffle Attention (SA), Efficient Channel Attention (ECA), Global Attention Mechanism (GAM), and ResBlock + Convolutional Block Attention Module (ResCBAM). 4
Figure 4 CBAM and ResCBAM architectures. 7
Figure 5 Architectures of Shuffle Attention (SA) and Efficient Channel Attention (ECA).....8
Figure 6 GAM and ResGAM Architectures. 10
Figure 7 Pediatric wrist fracture detection using YOLOv8-AM models: (a) manually labeled, (b) ResCBAM, (c) ECA, (d) SA, (e) GAM, (f) ResGAM.....15
Figure 8 Precision-Recall Curves for YOLOv8-AM models across different categories in the GRAZPEDWRI-DX dataset. 15

List of Tables:

Table 1 Experimental results for fracture detection on the GRAZPEDWRI-DX dataset using YOLOv8-AM models with three attention modules. 'Inference' denotes the total prediction time per X-ray image, including preprocessing, inference, and post-processing stages. 13

Table 2 Results for fracture detection on the GRAZPEDWRI-DX dataset using YOLOv8-AM with GAM and ResGAM. 'Inference' denotes the total prediction time per X-ray image, including preprocessing and post-processing 14

Table 3 Performance comparison (F1 Score/mAP) for fracture detection using YOLOv8 and YOLOv8-AM on the GRAZPEDWRI-DX dataset. 'Inference' denotes the total prediction time per X-ray image, including all processing stages. 14

Acknowledgement:

I am profoundly grateful to **Professor TAN Ling** whose expert guidance and invaluable insights in the field of deep learning immensely enriched my understanding and directly contributed to the completion of this final report. Her resourceful classes and consistent support have been fundamental to my research and academic growth.I would also like to extend my heartfelt thanks to my parents, whose unwavering care and support have been a constant source of strength and motivation for me. Their encouragement has been vital to my academic journey.

Lastly, I am deeply appreciative of my classmates and friends for their camaraderie and support.

Thank you all for your invaluable support and encouragement.

1 Introduction

Wrist fractures are one of the most frequently occurring injuries, especially common among older adults and children. These fractures typically happen within the last two centimeters of the radius near the joint. Without prompt and appropriate treatment, complications such as deformities, limited joint mobility, and chronic pain may arise. In pediatric patients, incorrect diagnosis can have long-lasting consequences.

For pediatric wrist fractures, doctors often review the circumstances of the injury and perform a thorough preoperative assessment. Currently, fracture assessments largely depend on imaging technologies like Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and X-rays, with X-rays being preferred due to their affordability. In advanced medical facilities, these images must adhere to the Health Level 7 (HL7) and Digital Imaging and Communications in Medicine (DICOM) standards. However, the shortage of radiologists, particularly in underdeveloped areas, poses a significant barrier to delivering timely patient care, with error rates in emergency imaging analyses as high as 26%.

The development of computer-assisted diagnosis (CAD) systems has provided a valuable tool for healthcare professionals, aiding in diagnostic decisions. The continual advancements in deep learning and medical image processing have encouraged more researchers to apply neural network technologies to CAD tasks, including fracture detection. The YOLO (You Only Look Once) model, known for its efficacy in object detection, has been enhanced in its latest version, YOLOv8, introduced by Ultralytics in 2023. This version has seen wide applications in object detection tasks, including fracture detection in a dataset of over 20,000 pediatric wrist X-ray images.

The application of attention mechanisms, which finely tune the focus on critical information within the input, has become common in various neural network architectures. Currently, the primary types of attention mechanisms are spatial and channel attention, which target pixel-level relationships and channel dependencies, respectively. Research has shown that embedding these mechanisms into convolutional networks can significantly boost performance.

Accordingly, this paper introduces the YOLOv8-AM model, which incorporates four different attention modules: Convolutional Block Attention Module (CBAM), Global Attention Mechanism (GAM), Efficient Channel Attention (ECA), and Shuffle Attention (SA) into the YOLOv8 framework, as shown in Fig.2. Experimental analyses demonstrate enhanced

performance when combining Residual Block (Res Block) with CBAM, surpassing the effects of CBAM alone.

Figure 1 Sample Input and Ground Truth Images from the Study.

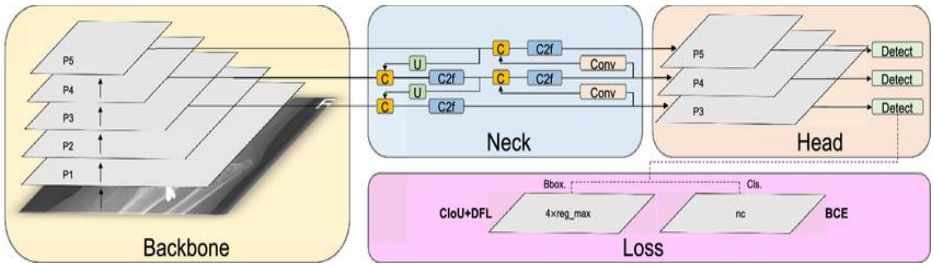


Figure 2 Overview of the YOLOv8-AM Model Architecture with Attention Modules for Pediatric Wrist Fracture Detection.

Our study finds that the GAM module has a lesser impact on model enhancement, which led to the development of a new model variant, ResGAM, combining ResBlock and GAM. The significant contributions of this work include applying these attention mechanisms to the YOLOv8 architecture, proposing new model configurations based on their performance, and demonstrating the improved diagnostic capabilities on the GRAZPEDWRI-DX dataset. The following sections detail the research on using deep learning for fracture detection, the architecture of the YOLOv8-AM model, and a comparative analysis of different model performances.

Main Contributions:

1. Integration of Attention Mechanisms: We have integrated four distinct attention modules—Convolutional Block Attention Module (CBAM), Global Attention Mechanism (GAM), Efficient Channel Attention (ECA), and Shuffle Attention (SA)—into the YOLOv8 architecture. This approach has significantly enhanced the model's ability to focus on relevant details for more accurate fracture detection.

2. Optimized Model Design: Our experiments reveal that the combination of Residual Block (Res Block) and CBAM (ResCBAM) significantly outperforms the standalone CBAM in terms of detection accuracy, leading us to integrate ResCBAM into the YOLOv8 architecture.

3. Evaluation of GAM: We evaluated the performance impact of the GAM module and found it to be less effective. This led to the development of ResGAM, which combines ResBlock with GAM, to optimize the effectiveness of the attention module within the YOLOv8 architecture.

4. Comparative Performance Analysis: The paper demonstrates that all YOLOv8-AM models, regardless of the specific attention module used, exhibit improved performance over the standard YOLOv8 model on the GRAZPEDWRI-DX dataset. This showcases the value of integrating attention mechanisms into deep learning models for medical imaging.

5. Evidence-Based Enhancements: Our findings are supported by rigorous testing and comparative analysis, establishing the YOLOv8-AM model based on ResCBAM (Residual Block + CBAM) as achieving state-of-the-art performance on the pediatric wrist fracture detection task.

2 Related Work

2.1 Fracture Detection

Fracture detection is a prominent topic in medical image processing (MIP). Researchers often use various neural networks for prediction, including the YOLO series models. Burkow et al. utilized the YOLOv5 model to recognize rib fractures in 704 pediatric Chest X-ray (CXR) images. Tsai et al. performed data augmentation on CXR images and used the YOLOv5 model for fracture detection. Warin et al. categorized maxillofacial fractures into four types (frontal, midfacial, mandibular, and no fracture) and predicted them using the YOLOv5 model on 3,407 CT images. Additionally, Warin et al. used the YOLOv5 model to detect fractures in mandible X-ray images. Yuan et al. incorporated external attention and 3D feature fusion methods into the YOLOv5 model for fracture detection in skull CT images. Furthermore, vertebral localization is essential for recognizing vertebral deformities and fractures. Mushtaq et al. utilized YOLOv5 for lumbar vertebrae localization, achieving a mean Average Precision (mAP) value of 0.957. While the YOLOv5 model is widely used in fracture detection, the utilization of the YOLOv8 model is relatively rare.

2.2 Attention Module

SENet initially proposed a mechanism to efficiently learn channel attention by applying Global Average Pooling (GAP) to each channel independently. Subsequently, channel weights were generated using the Fully Connected layer and the Sigmoid function, leading to improved model performance. Following the introduction of feature aggregation and feature recalibration in SENet, some studies attempted to enhance the SE block by capturing more sophisticated channel-wise dependencies. Woo et al. combined the channel attention module with the spatial attention module, introducing the CBAM to improve the representation capabilities of Convolutional Neural Networks (CNNs). To reduce information loss and enhance global dimension-interactive features, Liu et al. introduced modifications to CBAM and presented GAM, reconfiguring submodules to highlight important cross-dimension receptive regions. Although these methods have achieved better accuracy, they often result in higher model complexity and increased computational burden. Therefore, Wang et al. proposed the ECA module, which captures local cross-channel interaction by considering every channel and its k neighbors, resulting in significant performance improvement with fewer parameters. Differing from the ECA module, Zhang et al. introduced the SA module, which groups channel dimensions into multiple sub-features and uses the Shuffle Unit to integrate complementary sub-features with the spatial attention module for each sub-feature, achieving excellent performance with low model complexity. Each of these attention modules can be applied to different neural network architectures to enhance model performance.

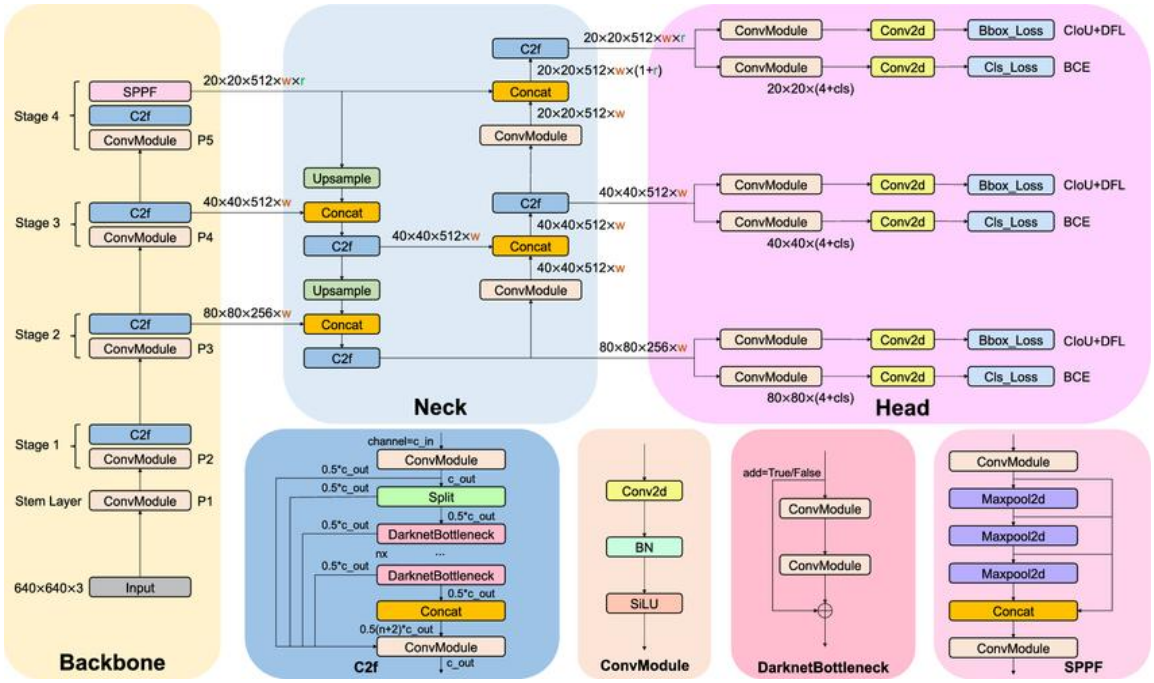


Figure 3 Architecture of YOLOv8-AM showing attention modules: Shuffle Attention (SA), Efficient Channel Attention (ECA), Global Attention Mechanism (GAM), and ResBlock + Convolutional Block Attention Module (ResCBAM).

3 Methodology

3.1 Baseline Model

The YOLOv8 architecture consists of four main components: Backbone, Neck, Head, and Loss Function. The Backbone incorporates the Cross Stage Partial (CSP) concept, which reduces computational load while enhancing the learning capability of CNNs. As illustrated in Fig. 3, YOLOv8 differs from YOLOv5 by using the C2f module instead of the C3 module. The C2f module combines the C3 module with the Extended ELAN concept from YOLOv7. Specifically, the C3 module includes three convolutional modules and multiple bottlenecks, whereas the C2f module consists of two convolutional modules concatenated with multiple bottlenecks. The convolutional module is structured as Convolution-Batch Normalization-SiLU (CBS).

In the Neck part, YOLOv5 employs the Feature Pyramid Network (FPN) architecture for top-down sampling to enrich the lower feature map with more information. Simultaneously, the Path Aggregation Network (PAN) structure is applied for bottom-up sampling to enhance the top feature map with precise location information. This combination ensures accurate prediction across varying image dimensions. YOLOv8 follows the FPN and PAN frameworks but removes the convolution operation during the up-sampling stage, as shown in Fig. 3.

Unlike YOLOv5, which uses a coupled head, YOLOv8 adopts a decoupled head, separating the classification and detection heads. YOLOv8 eliminates the objectness branch, retaining only the classification and regression branches. Additionally, it switches from an anchor-based method to an anchor-free approach, where the target's location is determined by its center, and predictions estimate the distance from the center to the boundary.

In YOLOv8, the classification branch uses the Binary Cross-Entropy (BCE) Loss, expressed by the equation:

$$LossBCE = -w[yn \log xn + (1 - yn)\log(1 - xn)] \quad (\mathbf{1})$$

where w denotes the weight, y_n represents the labeled value, and x_n signifies the predicted value generated by the model. For the regression branch, YOLOv8 incorporates Distribute Focal Loss (DFL) and Complete Intersection over Union (CIoU) Loss.

The DFL function emphasizes the expansion of probability values around object y. Its equation is:

$$Loss_{DF} = -[(y_{n+1} - y)\log\frac{y_{n+1} - y_n}{y_{n+1} - y_n} + (y - y_n)\log\frac{y - y_n}{y_{n+1} - y_n}] \quad (2)$$

The CIoU Loss introduces an influence factor to the Distance Intersection over Union (DIoU) Loss by considering the aspect ratio of the predicted bounding box and the ground truth bounding box. The corresponding equation is:

$$Loss_{CIoU} = 1 - IoU + \frac{d^2}{c^2} + \frac{v^2}{(1 - IoU) + v} \quad (3)$$

where IoU measures the overlap between the predicted and ground truth bounding boxes, d is the Euclidean distance between the center points of the predicted and ground truth bounding boxes, and c is the diagonal length of the smallest enclosing box that contains both predicted and ground truth bounding boxes. Additionally, v represents the parameter quantifying the consistency of the aspect ratio, defined by:

$$v = \frac{4}{\pi^2}(\arctan\frac{w_{gt}}{h_{gt}} - \arctan\frac{w_p}{h_p})^2 \quad (4)$$

where w denotes the weight of the bounding box, h represents the height of the bounding box, gt refers to the ground truth, and p refers to the prediction.

3.2 Proposed Method

In recent years, attention mechanisms have achieved outstanding results in the field of object detection. By integrating attention mechanisms, models can effectively identify and focus on the most relevant information in input images while filtering out irrelevant data

This study integrates attention modules into the Neck part of YOLOv8 to improve the identification of crucial features and reduce noise. As shown in Fig. 3, attention modules such as CBAM, GAM, ECA, and SA are individually applied after each of the four C2f modules.

A detailed explanation of these four attention modules is provided in Section 3.3.

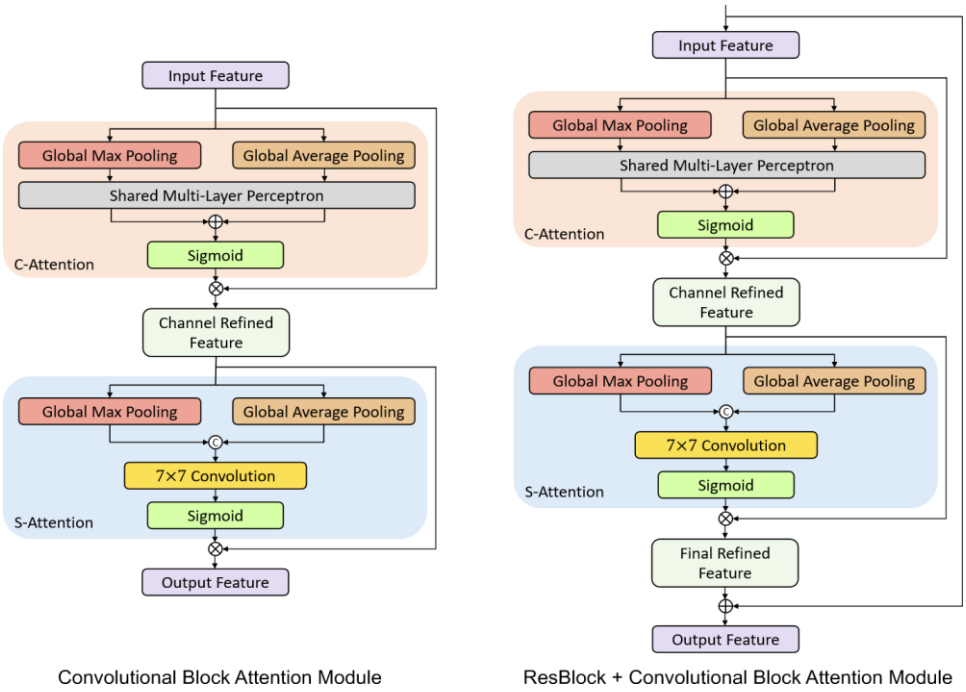


Figure 4 CBAM and ResCBAM architectures.

3.3 Attention Modules

Convolutional Block Attention Module (CBAM):

CBAM comprises both channel attention (C-Attention) and spatial attention (S-Attention), as shown on the left of Fig. 4. Given an intermediate feature map denoted as Finput ∈ RCHW, CBAM sequentially infers a 1D channel attention map MC ∈ RC×1×1 and a 2D spatial attention map MS ∈ R1×H×W through the following equation:

$$FCR = MC(Finput) \otimes Finput \quad (5)$$

$$FFR = MS(FCR) \otimes FCR \quad (6)$$

,

where ⊗ is the element-wise multiplication; F_{CR} is the Channel Refined Feature, and F_{FR} is the Final Refined Feature. For CBAM, F_{output} is F_{FR} as shown in the following equation:

$$Foutput = FFR \quad (7)$$

It can be seen from the right of Fig. 4, for ResBlock + CBAM (ResCBAM), F_{output} is the element-wise summation of F_{input} and F_{FR} as shown in the following equation:

$$Foutput = Finput + FFR \quad (8)$$

Based on the previous studies [68,44], CBAM employs both Global Average Pooling (GAP) and Global Max Pooling (GMP) to aggregate the spatial information of a feature map, which generates two different spatial contextual descriptors. Subsequently, these two descriptors share the same Multi-Layer Perceptron (MLP) with one hidden layer. Finally, the output feature vectors from the element-wise summation are input to the sigmoid function (σ). The specific channel attention equation is as follows:

$$MC(F) = \sigma\left[MLP\left(GAP(F)\right) + MLP\left(GMP(F)\right)\right] \quad (9)$$

Efficient Channel Attention (ECA):

Efficient Channel Attention (ECA) primarily focuses on cross-channel interactions and utilizes a 1D convolution with an adaptive kernel, as illustrated on the right side of Fig. 5. This cross-channel interaction introduces a novel method for feature combination, enhancing the representation of feature-specific semantics. The process starts with the input feature map $F_{input} \in R^{C\times H\times W}$, which, after applying Global Average Pooling (GAP) and cross-channel interaction, results in the aggregated feature F_a. The transformation is expressed by the equation:

$$Fa = C\left(GAP(Finput)\right) \quad (10)$$

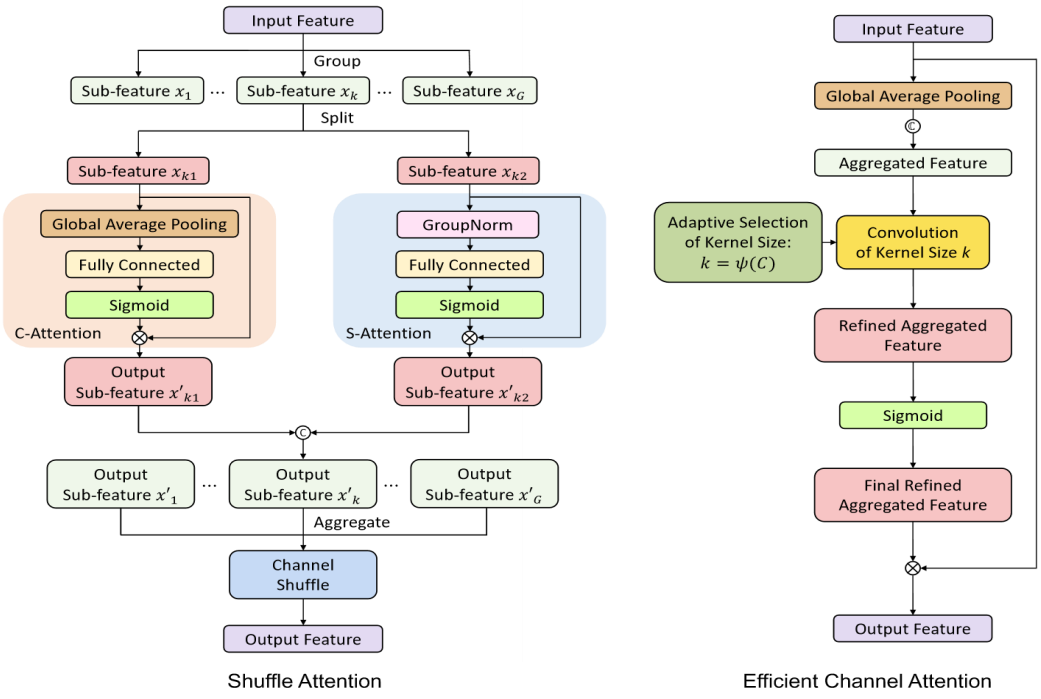


Figure 5 Architectures of Shuffle Attention (SA) and Efficient Channel Attention (ECA).

Shuffle Attention (SA):

Shuffle Attention (SA) splits the input feature maps into various groups and uses the Shuffle Unit to incorporate both channel and spatial attention within a single block for each group, as depicted on the left side of Fig. 5. Following this, the sub-features are combined, and the "Channel Shuffle" operator, similar to that used in ShuffleNetV2, is implemented to enhance information exchange among the different sub-features.

For channel attention, SA utilizes Global Average Pooling (GAP) to capture and integrate global information for the sub-feature x_{k1}. It also employs a straightforward gating mechanism with sigmoid functions, allowing for precise and adaptable feature selection. The resultant output of the channel attention process is given by the equation below:

$$x_{k1}{}' = \sigma[FC(GAP(x_{k1}))] \otimes x_{k1} \quad (\mathbf{11})$$

Global Attention Mechanism (GAM):

Global Attention Mechanism (GAM) incorporates the core architecture of CBAM, featuring both channel and spatial attention, but modifies the submodules as depicted in Fig. 6. Additionally, a Shortcut Connection is introduced between the layers within GAM to facilitate faster input propagation, as expressed in the following equation:

$$F_{output} = F_{input} + [M_S(M_C(F_{input}) \otimes F_{input}) \otimes (M_C(F_{input}) \otimes F_{input})]. \quad (\mathbf{12})$$

For channel attention, GAM first applies a 3D permutation to preserve three-dimensional information. It then uses a two-layer MLP to enhance the channel-spatial interdependencies across dimensions. The process is summarized by the equation provided.

$$MC(F) = \sigma\left[ReversePermutation\left(MLP\left(Permutation(F)\right)\right)\right]. \quad (\mathbf{13})$$

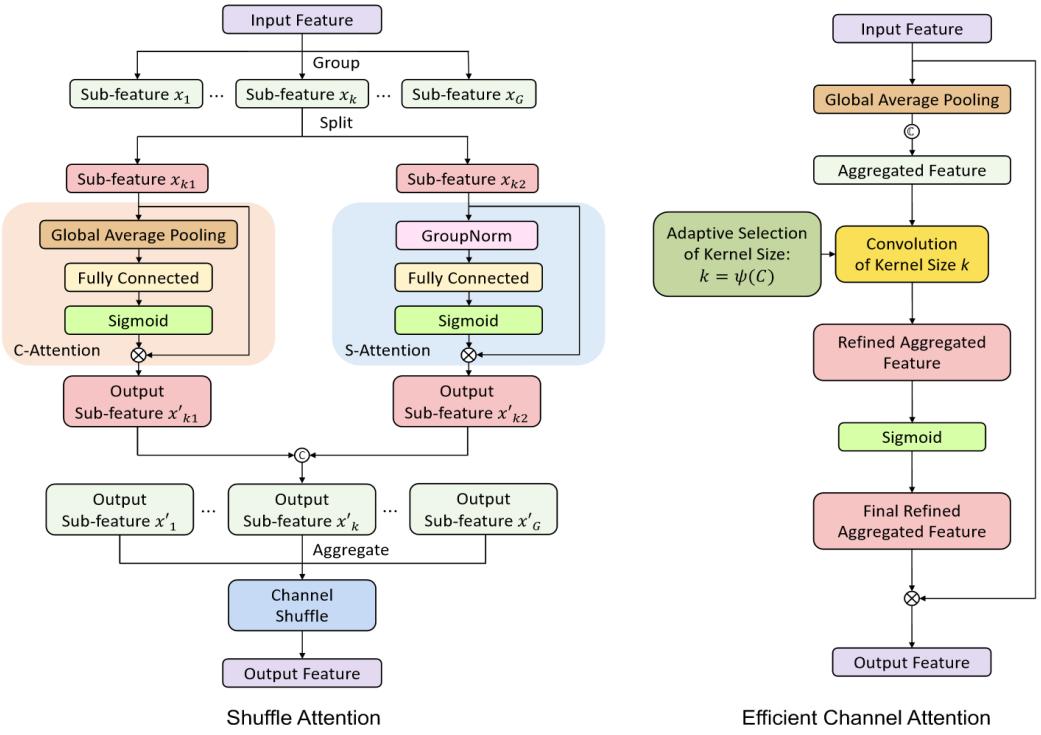


Figure 6 GAM and ResGAM Architectures.

For spatial attention, GAM uses two 7 × 7 convolution layers to integrate spatial information. It also adopts a reduction rate *r* consistent with the approach in BAM [44]. The corresponding equation is presented as follows:

$$MS(F) = \sigma\left[BN\left(f7\times7\left(BN + ReLU\left(f7\times7(F)\right)\right)\right)\right]. \quad (\mathbf{14})$$

In contrast to CBAM, the authors of GAM considered that max-pooling would reduce the amount of information and have a negative effect, so pooling was eliminated to further preserve the feature map.

4 Experiment Results and Analysis

4.1 Dataset

The GRAZPEDWRI-DX is a publicly available dataset from the University of Medicine of Graz, consisting of 20,327 pediatric wrist trauma X-ray images. These images were acquired by pediatric radiologists at the Department for Pediatric Surgery of the University Hospital Graz from 2008 to 2018, encompassing 6,091 patients across 10,643 studies. The dataset includes 74,459 image annotations, with a total of 67,771 labeled objects.

4.2 Preprocessing and Data Augmentation

The GRAZPEDWRI-DX dataset does not come with predefined splits for training, validation, and testing; hence, we randomly divided the dataset into 70% for training (14,204 images), 20% for validation (4,094 images), and 10% for testing (2,029 images).

Given the limited variation in brightness within the dataset's X-ray images, training a model solely on these images could result in poor performance on other X-ray images. To improve the model's robustness, we utilized data augmentation techniques to enrich the training set. Specifically, we adjusted the contrast and brightness of the images using the `addWeighted` function from OpenCV, an open-source computer vision library.

4.3 Evaluation Metric

Parameters (Params): The number of parameters in a model is influenced by its architectural complexity, including the number of layers and neurons per layer, among other factors. Typically, a model with more parameters has a larger size, which often correlates with improved performance. However, larger models require more data and computational resources for training. In practical applications, finding a balance between model complexity and computational costs is crucial.

Floating Point Operations (FLOPs): Floating Point Operations per Second is a measure used to evaluate the performance of computing systems and is commonly applied to gauge the computational complexity of neural network models. FLOPs indicate the number of floating-point operations executed per second, serving as an important measure of computational performance and speed. In environments with limited resources, models with fewer FLOPs may be more appropriate, whereas models with higher FLOPs typically demand more robust hardware.

Mean Average Precision (mAP): Mean Average Precision is widely used to assess the effectiveness of object detection models. This metric evaluates a model's ability to correctly identify and locate objects in an image. Precision reflects the accuracy of detected objects compared to actual objects, while recall represents the percentage of actual objects correctly identified by the model. These metrics are combined in mAP, which involves calculating the area under the Precision-Recall curve for each category (Average Precision), and then averaging these values across all categories to determine the overall mAP.

F1 Score: The F1 Score is derived from the harmonic mean of precision and recall, ranging from 0 to 1, with values closer to 1 indicating a better balance between precision and recall. A skewed precision or recall value towards 0 results in a lower F1 Score, indicating suboptimal model performance. The F1 Score is particularly useful in evaluating the model's accuracy in predicting positive categories and its sensitivity, especially in scenarios involving imbalanced category classification, where relying solely on accuracy might introduce bias.

4.4 Experiment Setup

We train the YOLOv8 and various YOLOv8-AM models on the dataset. Although Ultralytics recommends training YOLOv8 for 300 epochs, experimental evidence suggests optimal performance can be achieved within 60 to 70 epochs; hence, we set training for all models at 100 epochs.

Regarding hyperparameters, we opt for the SGD optimizer over Adam, following insights from ablation studies. The optimizer's weight decay is set at 5e-4, with a momentum of 0.937, and an initial learning rate of 1e-2. We explore the impact of different input image sizes, setting them at 640 and 1024 for various tests. All models are trained using Python 3.9 on PyTorch 1.13.1, and we recommend using at least Python 3.7 and PyTorch 1.7 for training. For practical execution, we utilize an NVIDIA GeForce RTX 3090 GPU with a batch size of 16 to manage GPU memory limits.

4.5 Experimental Results

To assess the impact of various input image sizes on the YOLOv8-AM model's performance in fracture detection, we conducted training sessions with image sizes of 640 and 1024. After training, we evaluated the model's performance as presented in the YOLOv8-AM model based on different attention modules on the test set with the corresponding image sizes.

Table 1 Experimental results for fracture detection on the GRAZPEDWRI-DX dataset using YOLOv8-AM models with three attention modules. 'Inference' denotes the total prediction time per X-ray image, including preprocessing, inference, and post-processing stages.

Table 1c: Efficient Channel Attention

Model Size	Input Size	mAPval 50	mAPval 50-95	Params (M)	FLOPs (B)	Inference (ms)
Small	640	61.4	37.4	11.14	28.67	1.9
Medium	640	62.1	38.7	25.86	79.1	2.5
Large	640	62.6	40.2	43.64	165.45	3.6
Small	1024	62.1	38.7	11.14	28.67	2.7
Medium	1024	62.4	40.1	25.86	79.1	5.2

Table 1b: Shuffle Attention

Model Size	Input Size	mAPval 50	mAPval 50-95	Params (M)	FLOPs (B)	Inference (ms)
Small	640	62.7	39.0	11.14	28.67	1.7
Medium	640	63.3	40.1	25.86	79.1	2.5
Large	640	64.0	41.5	43.64	165.44	3.9
Small	1024	63.5	39.8	11.14	28.67	2.9
Medium	1024	64.1	40.3	25.86	79.1	5.2
Large	1024	64.3	41.6	43.64	165.44	8.0

Table 1a: ResBlock + Convolutional Block Attention Module

Model Size	Input Size	mAPval 50	mAPval 50-95	Params (M)	FLOPs (B)	Inference (ms)
Small	640	61.6	38.9	16.06	38.27	1.9
Medium	640	62.8	39.8	33.84	98.19	2.9
Large	640	62.9	40.1	53.87	196.2	4.1
Small	1024	63.2	39.0	16.06	38.27	3.0
Medium	1024	64.3	41.5	33.84	98.19	5.7
Large	1024	65.8	42.2	53.87	196.2	8.7

As shown in Tables 1 and 2, models trained with an input image size of 1024 outperform those trained with an input image size of 640. However, this performance improvement comes with an increase in inference time. For example, the ResCBAM-based YOLOv8-AM model with a large model size achieves a mean Average Precision at IoU 50 (mAP 50) of 42.2% with an input image size of 1024, which is 5.24% higher than the 40.1% achieved with an input image size of 640. Nevertheless, the inference time increases from 4.1ms to 8.7ms due to the larger model size.

Table 2 Results for fracture detection on the GRAZPEDWRI-DX dataset using YOLOv8-AM with GAM and ResGAM. 'Inference' denotes the total prediction time per X-ray image, including preprocessing and post-processing

Table 2: Global Attention Mechanism

Model Size	Input Size	mAPval 50	mAPval 50-95	Params (M)	FLOPs (B)	Inference (ms)
Small	640	61.4	38.6	13.86	34.24	2.7
Medium	640	62.8	40.5	30.27	90.26	3.9
Large	640	64.0	41.2	49.29	183.53	9.4
Small	1024	64.8	41.2	13.86	34.24	4.4
Medium	1024	64.9	41.3	30.27	90.26	12.4
Large	1024	65.0	41.8	49.29	183.53	18.1

Table 1: ResBlock + Convolutional Block Attention Module

Model Size	Input Size	mAPval 50	mAPval 50-95	Params (M)	FLOPs (B)	Inference (ms)
Small	640	0.625	0.397	13.86	34.24	2.2
Medium	640	0.628	0.398	30.27	90.26	3.6
Large	640	0.633	0.407	49.29	183.53	8.7
Small	1024	0.635	0.4	13.86	34.24	4.3
Medium	1024	0.637	0.405	30.27	90.26	8.9
Large	1024	0.642	0.41	49.29	183.53	12.7

Table 1 displays the performance of the YOLOv8-AM model using three different attention modules: ResCBAM, SA, and ECA, tested across various model and input sizes. In Table 2, the effectiveness of the YOLOv8-AM is evaluated using GAM alone. Furthermore, we introduce an innovative method, ResGAM, which combines ResBlock with GAM to improve the model's performance. Notably, with a medium model size and an input size of 1024, ResGAM significantly improves mAP, increasing it from 63.7% to 64.9%, indicating a clear benefit of our ResGAM approach in enhancing model efficacy.

Table 3 Performance comparison (F1 Score/mAP) for fracture detection using YOLOv8 and YOLOv8-AM on the GRAZPEDWRI-DX dataset. 'Inference' denotes the total prediction time per X-ray image, including all processing stages.

Table 3: Performance Comparison of Fracture Detection

Module	Params (M)	FLOPs (B)	F1 Score	mAPval 50	mAPval 50-95	Inference
N/A	43.61M	164.9B	0.62	63.6%	40.4%	7.7ms
ResCBAM	53.87M	196.2B	0.64	65.8%	42.2%	8.7ms
SA	43.64M	165.4B	0.63	64.3%	41.6%	8.0ms
ECA	43.64M	165.5B	0.65	64.2%	41.9%	7.7ms
GAM	49.29M	183.5B	0.65	64.2%	41.0%	12.7ms
ResGAM	49.29M	183.5B	0.64	65.0%	41.8%	18.1ms

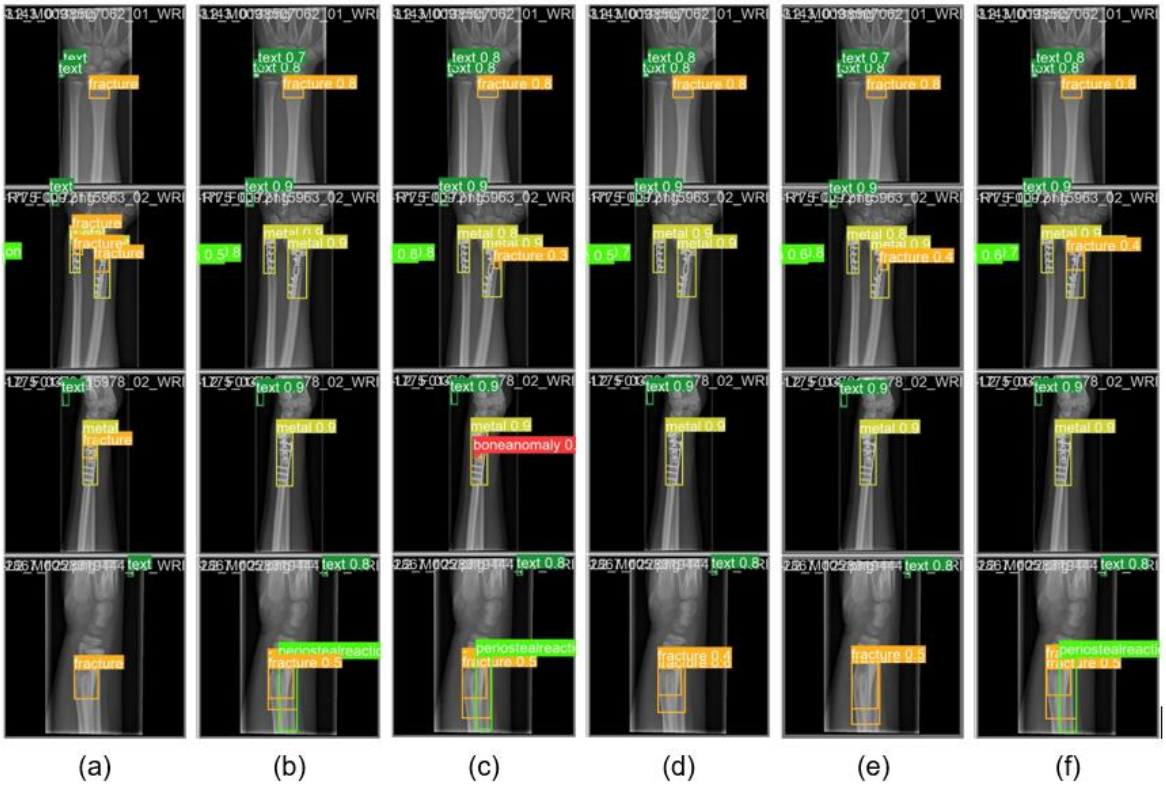
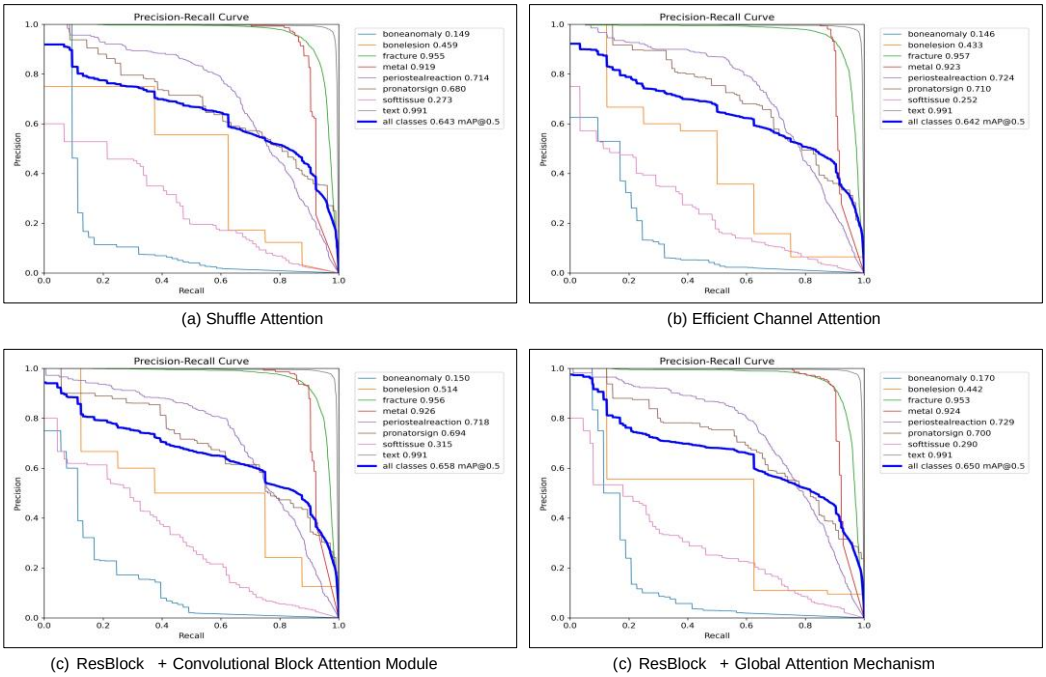


Figure 7 Pediatric wrist fracture detection using YOLOv8-AM models: (a) manually labeled, (b) ResCBAM, (c) ECA, (d) SA, (e) GAM, (f) ResGAM.



(a) Shuffle Attention

(b) Efficient Channel Attention

(c) ResBlock + Convolutional Block Attention Module

(c) ResBlock + Global Attention Mechanism

Figure 8 Precision-Recall Curves for YOLOv8-AM models across different categories in the GRAZPEDWRI-DX dataset.

To assess the impact of various attention modules on model performance, we used experimental data with an input image size of 1024 and a large model size, as shown in Table 3. The F1 score, mAP 50-95, and mAP 50 values for all YOLOv8-AM models exceed those of the YOLOv8 model. Notably, the YOLOv8-AM models with SA and ECA attention modules achieved mAP 50 values of 64.3% and 64.2%, respectively, slightly higher than the 63.6% by the YOLOv8 model. The inference times are comparable. The ResCBAM model, in particular, reached an mAP 50 of 65.8%, setting a new benchmark for performance. However, the performance gains with the GAM module were unsatisfactory, leading us to integrate ResGAM into the YOLOv8-AM framework.

This study also evaluates the effectiveness of these models as computer-assisted diagnostic tools for fracture prediction in a clinical setting. Four X-ray images were randomly selected to demonstrate the predictive capabilities of the various YOLOv8-AM models (Fig. 7). These models significantly aid radiologists and surgeons by accurately identifying fractures and detecting metallic punctures in specific scenarios, although their performance may diminish with dense fractures.

Fig. 8 presents the Precision-Recall Curves for different YOLOv8-AM models across various categories. These models generally perform well, with average accuracies exceeding 90% for detecting fractures, metal, and text. However, their effectiveness is lower in detecting bone anomalies and soft tissue, with accuracies around 45% and 30%, respectively. These lower accuracies impact the overall mAP 50 values. The limited presence of bone anomaly and soft tissue objects in the GRAZPEDWRI-DX dataset hinders performance improvements. To address these limitations, incorporating additional data is essential to enhance model effectiveness.

5 Discussion

Table 3 shows that the improvement in model performance due to the GAM in the YOLOv8-AM is minimal on the GRAZPEDWRI-DX dataset. To try to enhance the YOLOv8-AM model's performance with GAM, we introduced ResGAM. However, its performance boost still falls short compared to that provided by ResCBAM. This may be attributed to GAM's elimination of pooling, which is crucial for extracting key features from each channel and focusing on important aspects according to theory [55].

Research [39] has shown significant improvements with CBAM on larger datasets such as CIFAR100 [30] and ImageNet-1K [11], which allow the neural network to learn from a diverse range of images and features across all dimensions. However, these datasets differ significantly from ours; CIFAR100 has 50,000 images and ImageNet-1K contains over 1.28 million images. In contrast, our dataset consists of only 14,204 X-ray images focused on specific features like bone fractures and lesions. This discrepancy suggests that our model's learning scope is narrower, and it primarily needs to identify relevant features specific to X-ray images, deviating from the broader learning applications suggested by [39].

6 Conclusion

Since the launch of the YOLOv8 model by Ultralytics in 2023, it has been widely applied in fracture detection across various body parts. Although the YOLOv8 model shows notable performance on the GRAZPEDWRI-DX dataset, it does not reach state-of-the-art (SOTA) levels. To overcome this, we introduced four attention modules—CBAM, ECA, SA, and GAM—into the YOLOv8 architecture to boost performance. Furthermore, by integrating ResBlock with CBAM and GAM, we developed ResCBAM and ResGAM. Remarkably, the mAP 50 for the YOLOv8-AM model using ResGAM increased from 64.2% to 65.0%, without additional parameters and FLOPs. The YOLOv8-AM model with ResCBAM achieved an even higher performance, reaching a mAP of 65.8%, and setting a new benchmark.

6.1 Future Work

Moving forward, we aim to explore the scalability of these enhancements across other datasets and expand the application of YOLOv8-AM models to other diagnostic scenarios beyond fractures. This includes refining the models to handle more complex imaging conditions and diverse medical imaging datasets. Moreover, we plan to integrate more advanced attention mechanisms and possibly combine different attention modules to further refine the model's accuracy and efficiency. Additionally, assessing the real-world applicability of these models in clinical settings will be crucial to ensuring their effectiveness in practical medical diagnostics.

References:

1. Adams, S.J., Henderson, R.D., Yi, X., Babyn, P.: Artificial intelligence solutions for analysis of x-ray images. Canadian Association of Radiologists Journal 72(1), 60–72 (2021)
2. Bamford, R., Walker, D.M.: A qualitative investigation into the rehabilitation experience of patients following wrist fracture. Hand Therapy 15(3), 54–61 (201)
3. Blüthgen, C., Becker, A.S., de Martini, I.V., Meier, A., Martini, K., Frauenfelder, T.: Detection and localization of distal radius fractures: Deep learning system versus radiologists. European journal of radiology 126, 108925 (2020)
4. Boochever, S.S.: His/ris/pacs integration: getting to the gold standard. Radiology management 26(3), 16–24 (2004)
5. Burki, T.K.: Shortfall of consultant clinical radiologists in the UK. The Lancet Oncology 19(10), e518 (2018)
6. Burkow, J., Holste, G., Otjen, J., Perez, F., Junewick, J., Alessio, A.: Avalanche decision schemes to improve pediatric rib fracture detection. In: Medical Imaging 2022: Computer-Aided Diagnosis. vol. 12033, pp. 611–618. SPIE (2022)
7. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze excitation networks and beyond. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp. 0–0 (2019)
8. Chen, Y., Kalantidis, Y., Li, J., Yan, S., Feng, J.: Aˆ 2-nets: Double attention networks. Advances in neural information processing systems 31 (2018)
9. Choi, J.W., Cho, Y.J., Lee, S., Lee, J., Lee, S., Choi, Y.H., Cheon, J.E., Ha, J.Y.: Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography. Investigative Radiology 55(2), 101–110 (2020)
10. Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H.: Expectation-maximization attention networks for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9167–9176 (2019)
11. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 510–519 (2019)
12. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. Advances in Neural Information Processing Systems 33, 21002–21012 (2020)
13. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)

14. Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., Hanel, D., Gardner, M., Gupta, A., Hotchkiss, R., et al.: Deep neural network improves fracture detection by clinicians. Proceedings of the National Academy of Sciences 115(45), 11591–11596 (2018)

15. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8759–8768 (2018)

16. Liu, Y., Shao, Z., Hoffmann, N.: Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv preprint arXiv:2112.05561 (2021)

17. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient CNN architecture design. In: Proceedings of the European conference on computer vision (ECCV). pp. 116–131 (2018)

18. Mounts, J., Clingenpeel, J., McGuire, E., Byers, E., Kireeva, Y.: Most frequently missed fractures in the emergency department. Clinical pediatrics 50(3), 183–186 (2011)

19. Su, Z., Adam, A., Nasrudin, M.F., Ayob, M., Punganan, G.: Skeletal fracture detection with deep learning: A comprehensive review. Diagnostics 13(20), 3245 (2023)

20. Tanzi, L., Vezzetti, E., Moreno, R., Aprato, A., Audisio, A., Massè, A.: Hierarchical fracture classification of proximal femur x-ray images using a multistage deep learning approach. European journal of radiology 133, 109373 (2020)

21. Tsai, H.C., Qu, Y.Y., Lin, C.H., Lu, N.H., Liu, K.Y., Wang, J.F.: Automatic rib fracture detection and localization from frontal and oblique chest x-rays. In: 2022 10th International Conference on Orange Technology (ICOT). pp. 1–4. IEEE (2022)

22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)

23. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: 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. pp. 7464–7475 (2023)
24. Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H.: Cspnet: A new backbone that can enhance learning capability of cnn. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 390–391 (2020)

25. Wang, C.Y., Liao, H.Y.M., Yeh, I.H.: Designing network design strategies through gradient path analysis. arXiv preprint arXiv:2211.04800 (2022)

26. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: 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 (2020)

27. Warin, K., Limprasert, W., Suebnukarn, S., Inglam, S., Jantana, P., Vicharueang, S.: Assessment of deep convolutional neural network models for mandibular fracture detection in panoramic radiographs. International Journal of Oral and Maxillofacial Surgery 51(11), 1488–1494 (2022)

28. Warin, K., Limprasert, W., Suebnukarn, S., Paipongna, T., Jantana, P., Vicharueang, S.: Maxillofacial fracture detection and classification in computed tomography images using convolutional neural network-based models. Scientific Reports 13(1), 3434 (2023)

29. Wolbarst, A.B.: Looking within: how X-ray, CT, MRI, ultrasound, and other medical images are created, and how they help physicians save lives. Univ of California Press (1999)
30. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)

31. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)