Artificial Intelligence Applications And Innovations Vol. 2 No. 1 (2025)

Received 27 August 2025
Revised 25 October 2025
Accepted 19 December 2025
Published online 1 January 2025

A Novel Hybrid CNN-Mamba Framework with DySample-Enhanced YOLOv11 for Automated Pediatric Wrist Fracture Detection

Mahdi. Zarrin[1], Jafar. Tanha[1*], Haniyeh. Nikkhah[1]

[1] Faculty of Electrical and Computer Engineering, University of Tabriz, Iran

*** Corresponding author email address**: tanha@tabrizu.ac.ir

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Zarrin, M., Tanha, J., & Nikkhah, H. (2025). A Novel Hybrid CNN-Mamba Framework with DySample-Enhanced YOLOv11 for Automated Pediatric Wrist Fracture Detection. *Artificial Intelligence Applications and Innovations, 2*(1), 11-30.
https://doi.org/10.61838/jaiai.2.1.2

A B S T R A C T

Wrist fractures, particularly distal radius and ulna fractures, are among the most common injuries in pediatric populations. Early and accurate detection of these injuries is critical for preventing long-term complications, yet interpreting pediatric wrist radiographs remains a challenging task due to the subtle nature of some abnormalities. In response to this challenge, we propose a novel hybrid framework for automated medical image detection, combining the strengths of convolutional neural networks (CNNs) and Mamba-based encoders to capture both local and global feature dependencies. To address the challenges in fusing features from these two distinct architectures, we introduce the Feature Aggregation Attention Module (FAAM), which dynamically combines the feature maps for more robust representation. Additionally, we enhance the YOLOv11 framework by replacing conventional upsampling in the neck with the Dysample technique, which improves feature propagation and refinement. We evaluate our method on the GRAZPEDWRI-DX dataset, a comprehensive collection of pediatric wrist trauma X-rays, demonstrating significant improvements in fracture detection. Our approach achieves an mAP@0.5 of 69.12% and an mAP@0.95 of 48.4%, showcasing its effectiveness in both general and challenging detection scenarios.

Keywords: *Wrist Fractures Detection, Object Localization, CNN-Mamba Framework, Hybrid Deep Learning, Medical Imaging, Feature Fusion*

1. Introduction

Wrist abnormalities are among the most frequent injuries encountered in pediatric populations, with wrist fractures (particularly distal radius and ulna fractures) being the most common. These injuries often occur during periods of rapid growth and increased physical activity, especially during adolescence [1, 2]. Early detection and accurate diagnosis are essential to prevent long-term complications such as malunion, growth plate disturbances, or chronic pain. Radiography, especially X-ray imaging, is the primary diagnostic tool due to its accessibility, speed, and cost-effectiveness [3]. However, the interpretation of pediatric wrist radiographs can be challenging, even for trained clinicians. Studies report that diagnostic errors in emergency radiograph readings can reach up to 26% [4, 5], exacerbated by the ongoing global shortage of radiologists. These issues highlight the pressing need for automated, accurate, and scalable solutions to assist in interpreting pediatric trauma images. In this context, deep learning (especially object detection models) has emerged as a powerful tool, showing promising results in identifying abnormalities from radiographs [6, 7].

Object detection is a core task in computer vision that identifies and locates objects within an image. It performs two functions simultaneously: classification, which labels

an object (e.g., a fracture), and localization, which draws a bounding box to pinpoint its exact position [8]. This technology allows models to assist in tasks like identifying and marking abnormalities on medical images. Traditional object detection techniques like the sliding window method were limited by their inefficiency and lack of contextual understanding. The evolution of object detection has since been marked by the development of region-based approaches and, more significantly, the rise of single-stage detectors like the YOLO (You Only Look Once) family. YOLO models perform object localization and classification in a single forward pass, enabling real-time detection with high accuracy. This makes them particularly well-suited for applications such as emergency triage or automated radiology systems [9]. Over time, the YOLO architecture has evolved through various iterations (YOLOv3 [10] through YOLOv12 [11]) each introducing improvements in speed, accuracy, and flexibility. The version employed in this paper, YOLOv11 [12], integrates state-of-the-art design principles such as anchor-free detection, decoupled head structures, and enhanced feature pyramids, further optimizing the trade-off between inference speed and detection accuracy. Despite their advancements, YOLO models have typically relied on convolutional neural networks (CNNs) as their backbone [13], which brings inherent limitations, especially when interpreting complex medical images like pediatric wrist X-rays.

While CNNs are excellent at capturing local features like edges and textures, their small receptive fields restrict their ability to model long-range dependencies, the relationships between distant parts of an image [14]. A subtle diagnosis in an X-ray might depend on cues from multiple, spatially separated areas, and a purely local view is often insufficient to capture these crucial associations. For example, a minor bone fracture might be subtle but is linked to significant soft-tissue swelling or dislocation a few centimeters away, a connection a traditional CNN might miss. To overcome this, several architectural solutions have emerged, each with its own trade-offs. For instance, Dilated CNNs [15], can expand their receptive fields but may suffer from a "gridding" effect, while Vision Transformers (ViTs) [14], effectively model global relationships but come with high computational cost and memory requirements. Similarly, Non-local Networks [16], also capture long-range dependencies but are limited by their significant computational expense. Global Convolutional Networks (GCNs) [17], use larger kernels for broader context but do not explicitly model relationships between non-contiguous pixels.

Given the shortcomings of traditional architectures, state-space models [18] have recently emerged as a compelling alternative, offering a fundamentally different approach to sequence modeling. Among these, Mamba [19] stands out as a novel and efficient sequence modeling framework. Unlike transformers, Mamba uses linear time state-space transitions, making it much more efficient in handling long sequences and high-resolution data. It captures both short- and long-term dependencies while maintaining low computational complexity and strong generalization capabilities [20]. These properties make Mamba particularly well-suited for the medical imaging domain, where global context, spatial continuity, and subtle feature integration are essential. In the context of radiographic analysis, Mamba can effectively capture interrelated features spread across the image (such as small fractures accompanied by soft tissue swelling or periosteal reactions) without the overhead of self-attention mechanisms [21].

In this study, we introduce a novel and comprehensive framework for medical image detection, designed to overcome the limitations of single-architecture models. Our proposed solution is built upon a dual-encoder architecture that combines the strengths of two distinct state-of-the-art models: a CNN-based encoder, which leverages its inherent inductive biases for superior extraction of fine-grained local features, and a Mamba-based encoder, which efficiently models both short-term (local) and long-term (global) dependencies through its linear-time state-space transitions. A critical challenge in this multi-encoder design is the optimal fusion of features from these fundamentally different architectures. To address this, we introduce the Feature Aggregation Attention Module (FAAM), a key contribution of this research. The FAAM is strategically applied to the last three layers of both encoders, intelligently and dynamically weighing and combining their feature maps to create a unified and enriched representation. This fused feature set is then passed to a modified YOLOv11 framework for the final object detection task. We further enhance performance by replacing conventional upsampling in the neck with a novel and more effective approach called Dysample [22], which ensures superior feature propagation and refinement. This integrated framework is meticulously designed to provide a

more robust and comprehensive understanding of complex radiographs, thereby enabling more accurate and detailed automated dictations. We apply this approach to the GRAZPEDWRI-DX [23] dataset, one of the most detailed pediatric wrist trauma X-ray datasets available, encompassing over 20,000 images annotated across multiple categories. Our contributions are as follows:

- First, we design and implement a novel hybrid framework that utilizes a CNN-based encoder for robust local feature extraction and a Mamba-based encoder, which efficiently captures both short-term (local) and long-term (global) dependencies, as dual encoders. This dual-encoder output is then integrated with the detection capabilities of YOLOv11's neck and head.
- Second, we introduce a novel Feature Aggregation Attention Module (FAAM) to address the critical challenge of efficiently and optimally combining the distinct feature sets extracted from the CNN and Mamba encoders.
- Third, we enhance the YOLOv11 framework by replacing its conventional upsampling mechanism in the neck with a new and more effective approach called "Dysample", ensuring superior feature propagation and refinement for improved detection performance.
- Lastly, we conduct extensive experiments on the GRAZPEDWRI-DX dataset to validate the effectiveness of our approach in detecting fractures and other wrist abnormalities.

The rest of the paper is organized as follows. In Section 2, we provide a comprehensive review of related work, discussing studies that have utilized the YOLO framework with traditional backbones, as well as those exploring the combination of Mamba and YOLO for medical image detection. We also review key findings from prior work on the dataset used in this research. Section 3 details our proposed methodology, beginning with an explanation of the dual-encoder branches, followed by the design of our Feature Aggregation Attention Module (FAAM), and concluding with a description of the modifications made to the YOLOv11 neck and head. Section 4 outlines our experimental setup. Section 5 presents the comprehensive results of our approach, followed by the conclusion in Section 6.

2. Related work

Fracture detection is a crucial and advancing area in medical image analysis, with computer vision techniques playing a pivotal role in improving diagnostic accuracy. This section provides a comprehensive and systematic review of the literature most relevant to our research. The studies are organized into three distinct categories to provide a clear understanding of the progression and current state of the art in this domain. First, we examine studies that utilize the YOLO framework with its traditional backbones for medical images detection. Next, we review recent and emerging work that explores the combination of Mamba and YOLO architectures, a key focus of our study. Finally, we summarize prior research and key findings specifically from the GRAZPEDWRI-DX dataset that we employ in this work.

2.1. YOLO with traditional Backbones for Medical Image Detection:

A recent study introduced YOLO-Med, an efficient end-to-end multi-task network for simultaneous object detection and semantic segmentation in biomedical images. The model incorporates a cross-scale task-interaction module alongside task-specific decoders, enabling improved feature fusion. Evaluated on the Kvasir-seg and a private biomedical dataset, YOLO-Med achieved a strong balance between accuracy and inference speed [24]. Tao Zhou et al. [25] presents a novel deep learning model called Mandible-YOLO, designed to improve fracture detection by addressing the limitations of existing methods, which often struggle with the unique and varied characteristics of fractures. The core of this research lies in its three main contributions: the Multi-scale Residual Feature Enhancement Module (MRFEM), which uses a multi-branch fusion strategy and different convolution kernel sizes to enhance the model's ability to perceive fractures of varying scales; the Spatial-Channel Feature Hybrid Module (SCFHM), which improves fracture recognition by using a dual-branch attention mechanism to extract both spatial and channel information; and the Global–Local Feature Hybrid Module (GLTHM), which integrates a Transformer module with a self-attention mechanism to enhance the model's capacity for capturing long-distance features, thereby improving localization. In essence, the proposed model is a highly specialized architecture engineered to handle the specific complexities of fracture detection by systematically improving feature extraction, attention, and long-range dependency capture. Deepak Puthanpura et al. [25] present a comprehensive, hybrid deep learning framework called TFL-net, specifically designed to improve tibia fracture diagnosis and localization due to its

challenging nature. The core of this research lies in its three key methodologies: it first employs a VGG16-based deep feature extraction combined with a Support Vector Machine (SVM) classifier for accurate fracture classification. Second, a custom-trained YOLOv8 network is central to the framework, enabling real-time and precise localization of fracture regions with high mean average precision (0.964 mAP@0.5) and efficiency. Lastly, an Extreme Gradient Boosting (XGBoost) module is integrated for individualized healing time prediction based on patient-specific data, with the entire system accessible via a graphical user interface to streamline the clinical workflow.

2.2. *Hybrid Mamba and YOLO Architectures for Medical Image Detection:*

A recent study proposed the BrYOLO-Mamba model, which integrates state-space models (SSMs) into a Br-ORSS Block and introduces improvements to the input and neck modules of YOLO-Mamba for bronchoscopic image analysis. The model was evaluated on three bronchoscopic datasets and achieved notable improvements, with a 2.95% increase in accuracy and a 3.41% improvement in F1-score compared to YOLO-Mamba, while also demonstrating high sensitivity and reduced computational cost [26]. Khalili et al. [27] developed The Mamba-YOLOvX model to improve detection and localization of thoracic abnormalities in chest X-rays. The model integrates global and local lesion information through combined spatial and channel attention mechanisms and selective scanning blocks, while a projection-based augmentation strategy leveraging rib segmentation and keypoint alignment was proposed to enhance anatomical consistency across datasets. Evaluated on VinDr-CXR, ChestX-ray8, and CXR-AL14, the model achieved average precision scores of 0.366, 0.153, and 0.615, respectively, demonstrating significant gains in precision, recall, and mean average precision, particularly for small lesions, and confirming strong cross-dataset generalizability. Another recent study proposed a single-stage Mamba-YOLACT instance segmentation model for head–neck lymph nodes in MRI, introducing a Cross-field and Cross-direction Feature Enhancement (CCFE) module and a MambaNet-based prediction head to better capture global dependencies and lesion features. On a head–neck lymph node MRI dataset, the model achieved APdet 69.8%, APseg 70.9%, ARdet 55.3%, ARseg 56.4%, mAPdet 39.4%, and mAPseg 41.0%, demonstrating accurate segmentation performance for clinical decision support [28].

2.3. *Review of the GRAZPEDWRI-DX Dataset:*

Ahmed et al. [29] evaluated single-stage detection models YOLOv5, YOLOv6, YOLOv7, and YOLOv8 for pediatric wrist abnormality detection on the GRAZPEDWRI-DX dataset, demonstrating that they outperform the widely used two-stage Faster R-CNN. Among them, YOLOv8m achieved the highest fracture sensitivity (0.92) and mAP (0.95), while YOLOv6m yielded the highest overall sensitivity (0.83) and YOLOv8x recorded the best overall mAP (0.77), underscoring the potential of single-stage models in pediatric wrist imaging. A recent study was the first to apply the YOLOv9 algorithm for fracture detection, using the GRAZPEDWRI-DX dataset with extended training via data augmentation. The proposed model achieved a 3.7% improvement over the state-of-the-art, increasing mAP50–95 from 42.16% to 43.73%, demonstrating its potential as a computer-assisted diagnostic tool for radiologists and surgeons [30]. Another recent study proposed the YOLOv8+GC model, which integrates a lightweight Global Context (GC) block into YOLOv8 to enhance global feature modeling for fracture detection. On the GRAZPEDWRI-DX dataset, the model improved mAP50 from 63.58% to 66.32%, achieving state-of-the-art performance compared to the baseline YOLOv8 [31]. Ju, R.Y et al. [32] applied the YOLOv8 algorithm with data augmentation to the GRAZPEDWRI-DX pediatric wrist trauma dataset, achieving state-of-the-art performance with an mAP50 of 0.638, surpassing the improved YOLOv7 (0.634) and baseline YOLOv8 (0.636). The authors also developed a clinical application, Fracture Detection Using YOLOv8 App, to support surgeons in fracture diagnosis and surgical planning.

3. Methodology

Our proposed model consists of three main components: a dual-branch encoder, Feature Aggregation Attention Module (FAAM), and a modified YOLOv11-style neck with DySample, as shown in Fig. 1. The architecture is specifically designed to improve wrist fracture detection accuracy while maintaining computational efficiency.

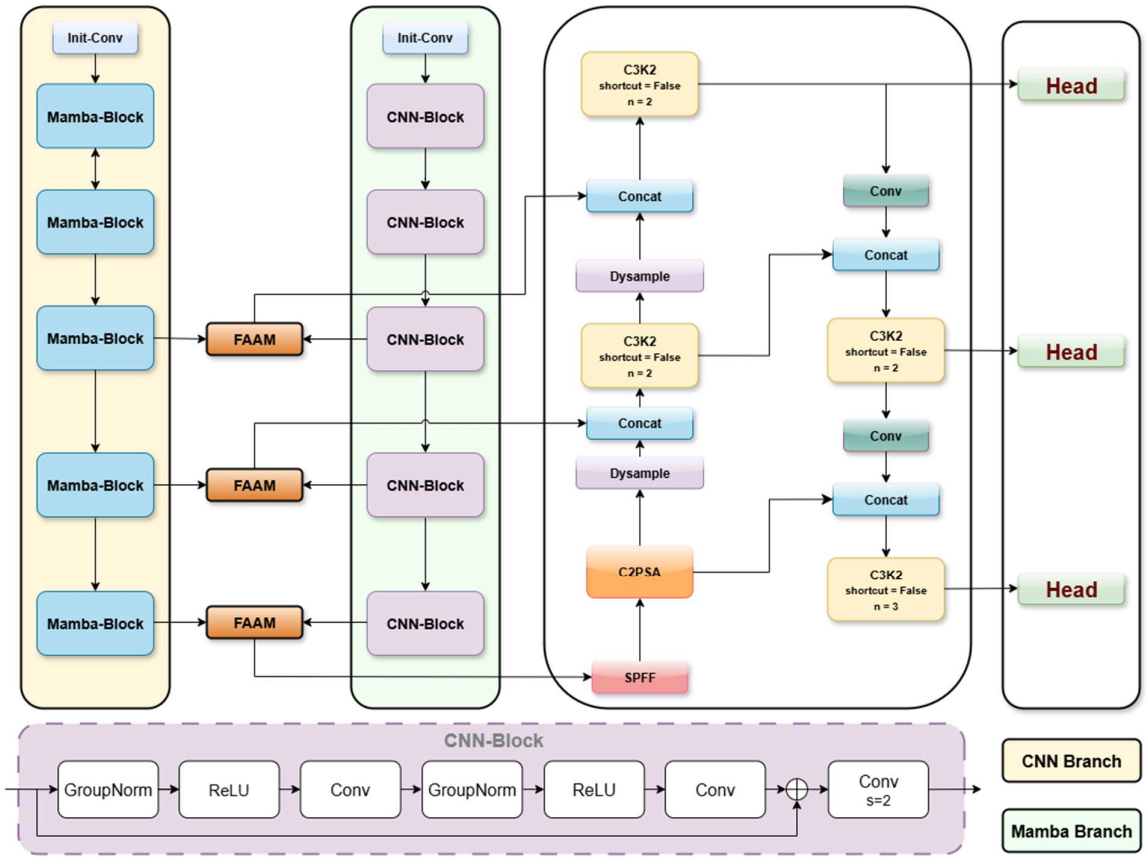

Figure 1. The architecture of our proposed dual-encoder framework, which integrates CNN-based and Mamba-based encoders with a modified YOLOv11 neck and head.

3.1. *Dual-Branch Encoder Design*

Wrist abnormality detection models often rely on a single feature extraction path within an encoder–decoder framework. However, this single-path approach struggles to simultaneously capture fine-grained local details and broader contextual information, a challenge inherent to the nature of different deep learning architectures. Convolutional neural networks (CNNs) are highly effective at modeling local spatial relationships through hierarchical receptive fields, making them well-suited for extracting precise textures, edges, and object boundaries in high-resolution images. However, their localized nature limits their ability to efficiently capture long-range dependencies, requiring either very deep architectures or additional context modules, which can increase computational complexity. Conversely, the Mamba architecture efficiently models both local and long-range dependencies via its Selective Scan mechanism, but it lacks the strong local inductive bias of CNNs for fine-detail preservation. To address these limitations, we propose a Dual-Branch Encoder Architecture that synergistically integrates CNN and Mamba branches, as shown in Fig. 1, leveraging hierarchical feature fusion to enhance segmentation accuracy and robustness across multiple scales. The encoder comprises two parallel feature extraction pathways, CNN and Mamba branches, whose outputs are fused at the last 3 layers. At layer l, the CNN and Mamba branches produce feature maps $F_A^l \in R^{C_l \times H_l \times W_l}$ and $F_B^l \in R^{C_l \times H_l \times W_l}$ respectively, where $C_l \in \{32,64,128,256,512\}$ correspond to downsampling scales of $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$.

3.1.1. *Mamba Branch*

This branch contains 5 stages with Mamba blocks followed by convolutional layers with stride 2 for progressive downsampling, as shown in Fig. 1, creating a hierarchical feature representation that captures both local

and global contextual information. In the following sections, we explain the state space model and 2D-selective-scan mechanism.

A. State space model

Mamba is built upon the State Space Model (SSM) framework. Unlike Transformer-based methods, which have quadratic time complexity, Mamba achieves linear-time complexity while demonstrating strong representational capacity for long-sequence modeling, offering significant efficiency in large-scale data processing [33].

The SSM is traditionally employed to represent linear time-invariant (LTI) systems. It establishes a mapping from an input sequence to an output sequence through a latent hidden state, thereby capturing temporal dynamics and encoding dependencies between inputs and outputs. Formally, the continuous-time SSM can be expressed using a system of linear ordinary differential equations (ODEs):

$$h(t) \in R^N, \quad x(t) \in R^L, \quad y(t) \in R^L \tag{1}$$

$$\frac{d}{dt}h(t) = Ah(t) + Bx(t) \tag{2}$$

$$y(t) = Ch(t) + Dx(t) \tag{3}$$

where $h(t)$ is the hidden state vector at time t, $A \in \mathbb{R}^{N \times N}$ is the state transition matrix, $B \in \mathbb{R}^{N \times L}$ is the input matrix, $C \in \mathbb{R}^{L \times N}$ is the output matrix, and $D \in \mathbb{R}^{L \times L}$ is the feed-forward matrix. The matrix A governs the temporal evolution of the hidden state, while B, C and D define the interactions between the input signal $x(t)$, state $h(t)$, and output response $y(t)$.

For deep learning applications, it is common to discretize these continuous-time dynamics to align with the sampling frequency of real-world signals [10]. This ensures that the model captures temporal dependencies at discrete intervals. Applying the zero-order hold principle, the discretized state-space equations are:

$$A_d = e^{\Delta A}, \quad B_d = (e^{\Delta A} - I)A^{-1}B, \quad C_d = C \tag{4}$$

$$h_k = A_d h_{k-1} + B_d x_k \tag{5}$$

$$y_k = C_d h_k + D x_k \tag{6}$$

where Δ denotes the discretization step size and I is the identity matrix. These discretization rules provide a tractable discrete-time representation of the continuous SSM, enabling seamless integration of Mamba into modern deep learning frameworks.

B. 2D-selective-scan mechanism

A fundamental limitation of directly applying Mamba to vision tasks lies in the incompatibility between two-dimensional (2D) visual data and one-dimensional (1D) sequential inputs. While 2D spatial information is crucial in visual understanding, it plays a relatively minor role in 1D sequence modeling. As a result, a direct mapping from 2D to 1D often yields restricted receptive fields, which fail to capture correlations with unobserved spatial regions [19].

To overcome this challenge, the Visual State Space Model (VSSM) introduces the 2D Selective-Scan (SS2D) mechanism [19], which forms the core of the framework. As illustrated in Fig. 2, SS2D first expands image patches into four distinct scanning directions, producing four independent sequences. This quad-directional scanning allows each element of the feature map to aggregate contextual information from all other positions across multiple orientations, thereby constructing a global receptive field while maintaining linear computational complexity.

Each generated sequence is then passed through the Selective-Scan State Space Model (S6), which models dependencies along the scanning direction. Finally, the sequences are merged to reconstruct the 2D feature map. Formally, given an input feature map z, the SS2D process is defined as:

$$z_i^{\text{expand}} = \text{expand}(z), \quad i \in \{1,2,3,4\} \tag{7}$$

$$z_i^{\text{out}} = S6\left(z_i^{\text{expand}}\right) \tag{8}$$

$$z^{\text{merge}} = \text{merge}(z_1^{\text{out}}, z_2^{\text{out}}, z_3^{\text{out}}, z_4^{\text{out}}) \tag{9}$$

Here, $expand(\cdot)$ and $merge(\cdot)$ denote the scan expansion and merging operations, respectively. The S6 block serves as the core VSSM operator, enabling each element in the 1D sequence to interact with previously scanned elements via a compact hidden state, thereby capturing both local and global dependencies efficiently.

3.1.2. CNN Branch

The second branch consists of five sequential CNN blocks that extract traditional convolutional features through a hierarchical structure, as shown in Fig. 1. Each CNN block follows a standardized architecture comprising GroupNorm layers for normalization, ReLU activation functions for non-linearity, and convolutional layers for feature extraction. The internal structure of each block includes two GroupNorm-ReLU-Conv sequences followed by a residual connection and an additional convolutional layer with stride 2 for downsampling. This design enables progressive spatial dimension reduction while increasing

feature depth, allowing the network to capture multi-scale anatomical patterns essential for fracture detection. The consistent block structure ensures stable gradient flow throughout the deep network while maintaining computational efficiency.

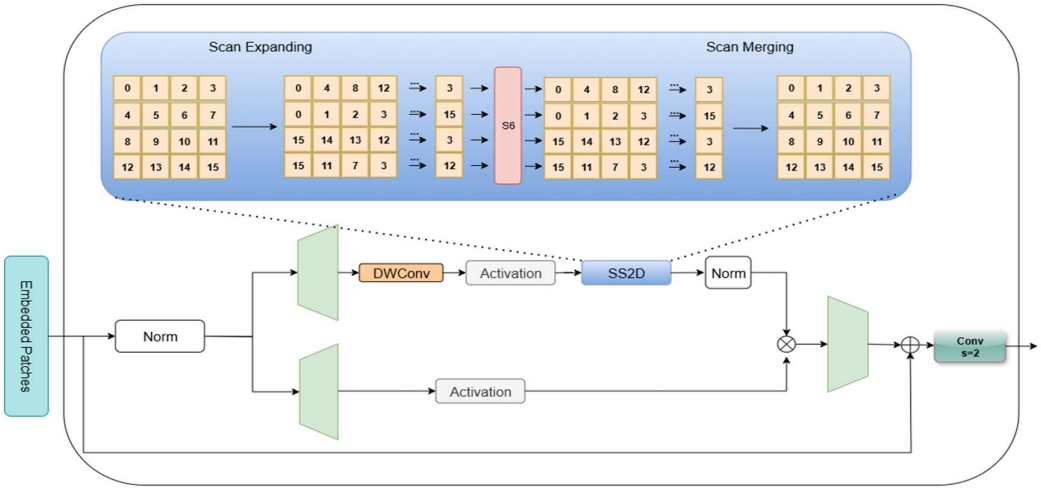

Figure 2. The 2D Selective-Scan (SS2D) mechanism.

3.2. Feature Aggregation Attention Module (FAAM)

To effectively merge the complementary information extracted from the CNN and Mamba encoders, we introduce FAAM, as illustrated in Fig. 3. The primary objective of FAAM is to enable dynamic, bidirectional interaction between the two feature streams, allowing each encoder to guide and refine the other. Instead of simply concatenating the outputs, FAAM establishes a selective communication pathway that emphasizes informative spatial regions and discriminative channels, leading to a more expressive fused representation.

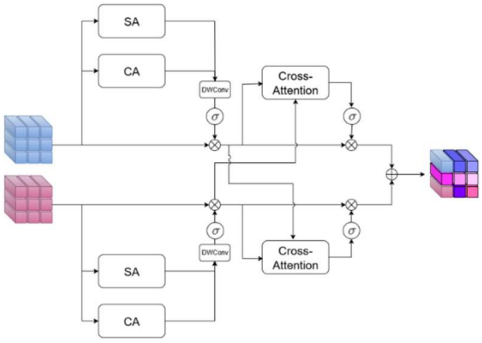

Figure 3. The architecture of the Feature Aggregation Attention Module (FAAM).

Each input branch (denoted as F_1 for the CNN encoder and F_2 for the Mamba encoder) is first enhanced through Spatial Attention (SA) and Channel Attention (CA) blocks. The SA mechanism identifies where critical visual cues (e.g., fine fracture lines or tissue boundaries) occur within the feature map, while the CA mechanism emphasizes which feature channels are most informative for detection. The outputs of these attention maps are combined and refined using a Depthwise Convolution (DWConv) followed by a sigmoid activation function $\sigma(\cdot)$, producing attention-weighted features $\widehat{F_1}$ and $\widehat{F_2}$. These operations are mathematically expressed as:

$$\widehat{F_1} = \sigma\left(\mathrm{DWConv}\left(\mathrm{SA}(F_1) + \mathrm{CA}(F_1)\right)\right) \odot F_1 \quad (10)$$

$$\widehat{F_2} = \sigma\left(\mathrm{DWConv}\left(\mathrm{SA}(F_2) + \mathrm{CA}(F_2)\right)\right) \odot F_2 \quad (11)$$

Where $\odot$ denotes element-wise multiplication. These steps ensure that each branch focuses on its most relevant spatial and channel characteristics before interacting with the other.

To promote cross-branch communication, FAAM employs a bidirectional Cross-Attention mechanism. In this stage, the refined features from one branch act as queries, while the other provides the keys and values, enabling each encoder to selectively absorb complementary information from its counterpart. This process allows the CNN branch, for instance, to utilize the Mamba branch's global context

to enhance its localized fracture features, and vice versa. The cross-attention refinement can be represented as:

$$\widetilde{F_1} = \sigma\left(\text{CrossAttn}\left(\widehat{F_1}, \widehat{F_2}\right)\right) \odot \widehat{F_1} \quad (12)$$

$$\widetilde{F_2} = \sigma\left(\text{CrossAttn}\left(\widehat{F_2}, \widehat{F_1}\right)\right) \odot \widehat{F_2} \quad (13)$$

After cross-attention, the two refined outputs are concatenated along the channel dimension to form the final fused feature representation:

$$F_{\text{fused}} = \text{Concat}\left(\widetilde{F_1}, \widetilde{F_2}\right) \quad (14)$$

As shown in Fig. 3, FAAM plays a pivotal role in unifying the outputs of the CNN and Mamba encoders. By facilitating a bidirectional exchange of information, FAAM allows the model to simultaneously capture fine-grained local structures and long-range contextual dependencies within pediatric wrist radiographs. This synergy ensures that subtle yet clinically significant patterns (such as faint fracture lines or soft tissue irregularities) are preserved and emphasized during feature fusion. Furthermore, FAAM achieves this enhanced representational capability through an efficient design that leverages lightweight depthwise convolutions and attention gating, adding minimal computational cost. The ablation results presented in Table 5 underscore its importance: incorporating FAAM consistently improved both mAP@0.5 and mAP@0.95, demonstrating superior detection accuracy and localization precision. Collectively, these findings confirm that FAAM is a critical component of the proposed framework, enabling a balanced, context-aware, and computationally efficient understanding of complex medical images.

3.3. *Enhanced Neck Architecture*

Our model utilizes an enhanced YOLOv11n-style neck. Key improvements, detailed below, include the integration of the DySample module for optimized feature propagation and refinement, and the incorporation of C3K2, C2PSA, and SPFF modules (Figures 4 and 5).

3.3.1. *YOLOv11-Style Design*

In this study, we enhance the neck architecture by incorporating Dysample to improve the overall performance of a YOLOv11-inspired design. Dysample refines the feature extraction process by optimizing the sampling strategy, which aids in more effective feature processing across different scales. This enhancement works in conjunction with the existing C3K2, C2PSA, and SPFF modules, which are designed for efficient feature processing, position-sensitive attention, and spatial pyramid feature fusion, respectively. The specific implementation of these modules within the neck architecture is visually represented in Fig. 4, providing a clear depiction of the system's design. A more detailed explanation of Dysample, including its impact and functionality within the network, is provided in the following section.

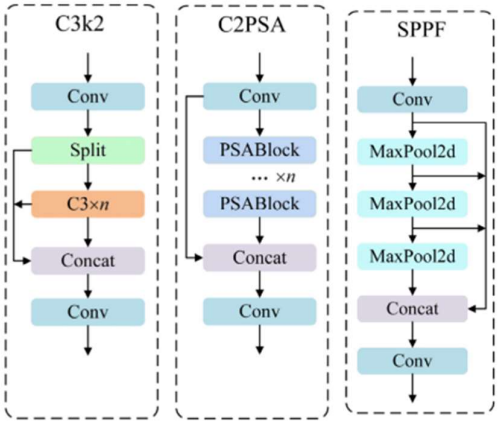

Figure 4. Structural designs of the C3K2, C2PSA, and SPFF modules within the enhanced neck architecture.

3.3.2. *DySample*

DySample is a lightweight, learnable, sampling-based upsampling module that replaces conventional upsampling operations in the neck of the detector[3]. The primary purpose of DySample is to improve the quality and informativeness of upsampled feature maps while keeping computational cost and parameter counts low. In standard detection pipelines, simple upsampling methods or heavy alternatives either produce blurred or artifact-prone outputs or introduce a large number of parameters and extra computation. DySample addresses this trade-off by predicting a small, per-location sampling offset field and using that field to resample the input features at adaptive, informative positions. In practice this means the network can learn where to pull information from the lower-resolution feature map to reconstruct higher-resolution details that are most useful for downstream localization and classification.

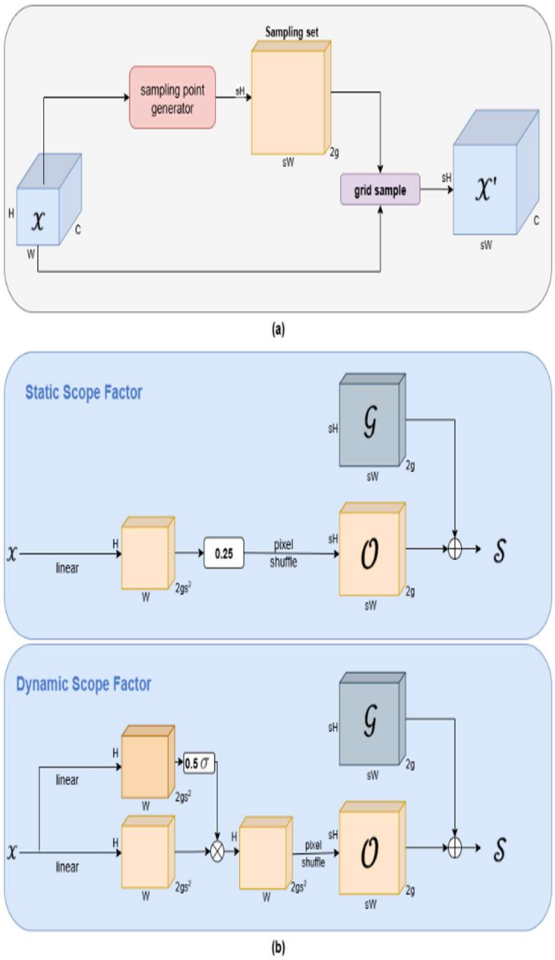

Figure 5. Sampling-based dynamic upsampling and module designs in DySample. (a) Sampling based dynamic upsampling. (b) Sampling point generator in DySample [22].

Functionally, DySample operates in two coordinated steps. First, a light projection (implemented as a small linear or 1×1 convolutional layer) maps the input feature map to a compact offset tensor; this tensor encodes coordinate offsets (relative to a regular upsampling grid) for each target sampling position. The offset tensor is reshaped and combined with the default sampling grid to form a learned sampling set. Second, a differentiable resampling operation (grid sampling with bilinear interpolation) uses the learned sampling set to produce the upsampled feature map. Because the sampling locations are predicted conditioned on the local features, the module effectively performs content-aware upsampling: it concentrates capacity on image regions and feature channels that matter for detection, rather than uniformly interpolating everywhere. Fig. 5 visualizes this sampling-based dynamic upsampling and the sampling point generator, clarifying how offsets are produced and applied to the grid.

DySample confers several concrete advantages when integrated into the neck of our YOLO-style detector. First, it reduces common upsampling artifacts (checkerboard effects from transposed convolutions, and oversmoothing from plain bilinear upsampling) because sampling locations are optimized during training and bilinear resampling produces smooth, spatially consistent outputs. Second, it improves preservation of high-frequency, localized detail (a key requirement for detecting subtle pediatric fractures) by adaptively steering sampling toward edges and anatomical boundaries rather than blindly interpolating. Third, DySample is computationally efficient: offset generation requires only a small projection and the grid-sampling operation is hardware-friendly and fully differentiable, so the module can be trained end-to-end with negligible increase in inference latency relative to heavyweight dynamic convs. This efficiency is reflected in our ablation results, where introducing Dysample yields clear improvements in mAP metrics while adding only a modest number of parameters (see Table 5).

4. Experiments

In this part, we give a detailed explanation of how our experiments were set up to test and support our method. We describe the datasets used, focusing on their key features and why they are important for our research. We also explain how the method was implemented, including parameter settings and any optimization techniques applied. Lastly, we discuss the evaluation methods used to measure how well the approach performed, ensuring the results are thoroughly analyzed.

4.1. Datasets

This research uses the GRAZPEDWRI-DX database, an open-access database of pediatric wrist trauma radiographs acquired by the Medical University of Graz. Comprising 20,327 de-identified images from 6,091 patients (mean age 10.9 years, ranging from 0.2 to 19 years; 2,688 females, 3,402 males, and 1 unknown), the dataset spans a decade of imaging from 2008 to 2018. Each radiograph, presented in 16-bit grayscale PNG format, includes posteroanterior and

lateral projections, providing comprehensive views of the wrist anatomy.

Annotations were meticulously performed between 2018 and 2020 by a team of expert radiologists and medical students, with validation by three seasoned pediatric radiologists. The dataset encompasses 74,459 image tags and 67,771 labeled objects, detailing various pathologies such as fractures, periosteal reactions, and bone lesions. These annotations employ bounding boxes, polygons, and lines to accurately delineate areas of interest.

The GRAZPEDWRI-DX dataset was chosen for this study for several important reasons. First, it provides detailed annotations, which is different from many other pediatric datasets that only give simple labels. This makes it possible to analyze wrist problems in more detail. Second, the dataset includes images from children of various ages, which helps study how wrist anatomy changes as kids grow. Finally, the dataset is publicly available under a Creative Commons license, meaning anyone can access it. This openness promotes transparency and allows other researchers to use the data for their own studies, making it a valuable tool for research in pediatric radiology and machine learning. To illustrate the dataset's richness and labeling quality, Fig. 6 presents example X-ray images where each labeled object type is shown individually. These include different types of findings such as fractures, periosteal reactions, metal implants, and soft tissue abnormalities. Each object is visually highlighted using bounding boxes or polygons, offering a clear view of how the dataset captures and categorizes wrist pathologies in children.

4.1.1. Analysis of Objects in the Dataset

The GRAZPEDWRI-DX dataset includes annotations for nine distinct objects related to pediatric wrist radiographs. These include common abnormalities such as fractures and periosteal reactions, as well as rarer findings like bonelesions and foreign bodies. The object "text", which is used to indicate side markers, appears in nearly all images.

Table 1 below summarizes the total number of instances for each object, their occurrence ratio across the dataset, and the frequency distribution of how many times each object type appears per image (i.e., zero, one, two, or more instances).

Table 1. Object Distribution in the GRAZPEDWRI-DX Dataset.

Object	Instances	Ratio (%)	Zero	One	Two	More
Text	20,274	99.74	–	–	–	–
Fracture	13,550	66.60	6,777	9,212	4,137	201
Periostealreaction	2,235	11.00	18,092	1,273	885	77
Metal	708	3.48	19,620	347	219	141
Pronatorsign	566	2.78	19,761	456	71	39
Softtissue	439	2.16	19,888	221	82	136
Boneanomaly	192	0.94	20,135	42	24	126
Bonelesion	42	0.21	20,285	11	8	23
Foreignbody	8	0.04	20,319	0	0	8

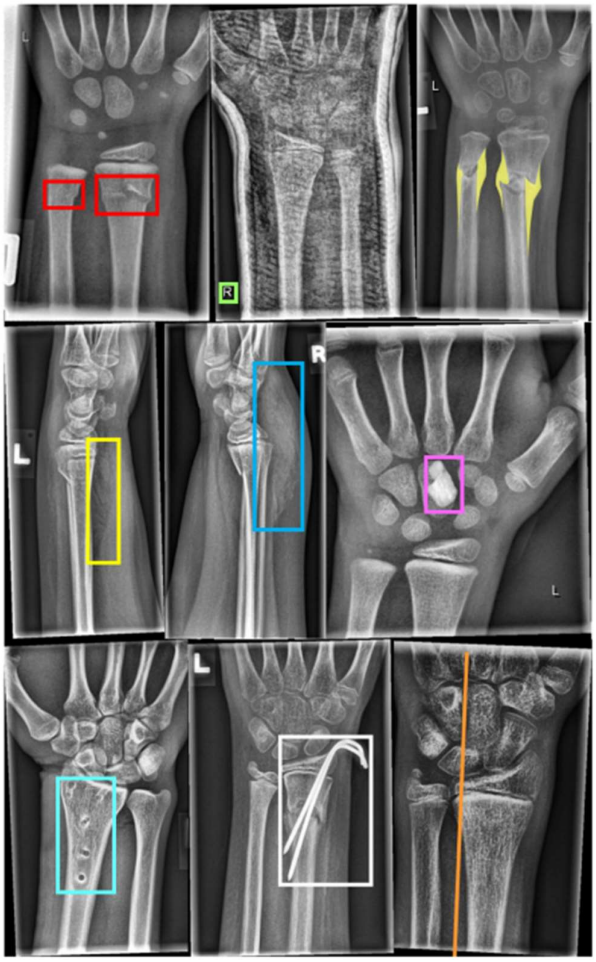

Figure 6. Examples of different objects labelled by the human experts. In the first row from left to right: Fracture (bounding box), text (bounding box), periosteal reaction (polygon). In the second row from left to right: Pronator quadratus sign (bounding box), soft tissue swelling (bounding box), foreign body (bounding box). In the last row from left to right: Bone anomaly (bounding box), metal (bounding box), and axis (line). The middle radiograph in the first row shows a cast. The right images in the first and third row were tagged with osteopenia [23].

4.1.2. Dataset Split

To ensure a robust and unbiased evaluation of our model's performance, the entire dataset of 20,327 images was systematically partitioned into three distinct subsets: a training set, a validation set, and a testing set. This approach is crucial for preventing data leakage and ensuring that the model's final performance is measured on unseen data, providing a realistic estimate of its generalization capability. The split was performed randomly to maintain a representative distribution of the data across all three sets. The dataset was divided as follows: a training set of 15,245 images (approximately 75% of the total dataset) was used to train the object detection models; a validation set of 4,066 images (approximately 20%) was used to fine-tune the model's hyperparameters and monitor its performance during training; and a test set of 1,016 images (approximately 5%) was held out and used only once, after the final model was trained, to ensure an objective and reliable assessment of the model's true performance on new, unseen data.

4.2. Implementation details

All the experiments in this paper are implemented based on the PyTorch framework. Considering the substantial computational cost during the training process for deep learning models, all models in this work were implemented and trained on an RTX 3090Ti GPU. In preprocessing data, all the input images were uniformly resized to the fixed resolution of 640 × 640 so that all input data would be compatible with the network's input dimension. A batch size of 8 was selected for training, balancing computational efficiency with memory constraints to achieve optimal performance of the model. To enhance the network's effectiveness, SiLU activation was applied after batch normalization layers. This combination not only stabilized training by normalizing intermediate feature distributions but also introduced non-linearity, enabling the network to learn more complex patterns. The Stochastic Gradient Descent (SGD) optimizer was chosen for optimizing the network's parameters. Although it does not have adaptive learning rates like Adam, SGD is a highly effective optimizer for object detection tasks due to its simplicity and efficiency in training large-scale models. A "poly" learning rate policy was used to control the dynamic process of learning rates during training. The learning rate of each iteration was estimated by $\left(1-\frac{iter}{max_{iter}}\right)^{power}$, where the initial learning rate was set to 0.01, and the power parameter was fixed at 0.9. This approach allows the learning rate to gradually decrease as the training progresses, preventing overshooting of the optimal solution and facilitating fine-tuning of weights in the later stages of training. For better regularization of the model and to prevent overfitting, a weight decay value of 0.0001 was used, which penalizes large weight updates for smoother learning. Table 2 outlines the hyperparameters that were tuned during the training process of the model. Hyperparameter optimization is a critical step in model development, as the selection of appropriate hyperparameter values can significantly impact the model's performance and generalization capabilities.

Table 2. Hyperparameters used for training the model.

Hyperparameters	Value
Batch size	8
Epochs	100
Learning rate	0.001 and 0.01
Optimizer	SGD

4.3. Evaluation metrics

To assess the performance of the object detection model, we utilized three widely recognized evaluation metrics: Precision, Sensitivity (Recall), and mean Average Precision (mAP). These metrics provide complementary insights into the model's ability to correctly identify and localize objects within the images, which is critical in clinical and diagnostic settings such as pediatric radiograph analysis.

- Precision:

Precision measures the proportion of correctly identified objects among all objects predicted by the model. It reflects how accurate the model is in its positive predictions. A higher precision value indicates fewer false positives, meaning the model is more reliable when it claims an object is present.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

Where:

- TP (True Positives): Correctly detected objects.
- FP (False Positives): Incorrectly detected objects.

- Sensitivity (Recall):

Sensitivity, also referred to as Recall, assesses the model's ability to detect all relevant objects. It represents the proportion of actual objects that were successfully identified. A high sensitivity value indicates that the model misses fewer true objects in the data.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{16}$$

Where:

- TP (True Positives): Correctly detected objects.
- FN (False Negatives): Missed objects.

- Mean Average Precision (mAP):

To evaluate object detection performance, Intersection over Union (IoU) is commonly used to measure the accuracy of predicted object locations. IoU is defined as the ratio of the area of overlap between the predicted and ground truth bounding boxes to the area of their union. For a given image, let A be the set of predicted bounding boxes and B be the set of ground truth bounding boxes.

IoU can be computed as:

$$\text{IoU}(A, B) = \frac{A \cap B}{A \cup B} \quad (17)$$

Where $A, B \in [0, 1]$

Typically, if IoU > 0.5, the detection is considered a true positive (TP); otherwise, it is treated as a false positive (FP).

Using TP and FP values computed via IoU, the Average Precision (AP) for each object class c is defined as:

$$\text{AP(c)} = \frac{TP(c)}{TP(c) + FP(c)} \quad (18)$$

Once AP has been calculated for each object class, the Mean Average Precision (mAP) is obtained by averaging AP across all object classes C:

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^{C} AP(c) \quad (19)$$

mAP is the primary metric used to quantify the performance of object detection algorithms. Specifically, mAP 0.5 refers to the mean Average Precision calculated using a threshold of IoU > 0.5. This metric balances precision and recall, offering a comprehensive evaluation that accounts for both accurate detection and minimization of false positives.

While mAP@0.5 is a standard metric, some applications, particularly those requiring very precise localization, use a stricter threshold.

mAP@0.95 (Mean Average Precision at an IoU threshold of 0.95) is a more stringent metric that evaluates a model's performance under highly precise localization conditions. For a detection to be considered a true positive (TP), its IoU with the ground truth must be greater than 0.95. Because of this high threshold, mAP@0.95 is a much more challenging metric to achieve. A model with a high mAP@0.95 score not only correctly identifies objects but also places their bounding boxes with near-perfect accuracy, leaving little to no room for error. This makes it particularly valuable for tasks where object boundaries are critical, such as in medical imaging or autonomous driving systems.

5. Results

This section presents a comprehensive evaluation of our proposed model's performance on the GRAZPEDWRI-DX dataset. The primary goal of this analysis is to demonstrate the model's high effectiveness in accurately detecting and localizing pediatric wrist pathologies, as confirmed through a series of rigorous comparative experiments. This section is divided into two subsections. The first subsection presents the quantitative results obtained from our experiments, while the second subsection compares and discusses the qualitative results.

5.1. Quantitative results

The quantitative results provide numerical evidence of the model's precision, sensitivity, and overall detection capability. This section presents a detailed analysis of our model's performance through a series of comparative tables, including a comparison of different versions of YOLO, a comparison with state-of-the-art methods, and a detailed review of our proposed method. Additionally, we have also shown the results obtained on each class.

5.1.1. Comparison of YOLO Versions

Table 3 illustrates a comprehensive quantitative evaluation of our proposed model "Ours" against various versions of YOLO architectures on the GRAZPEDWRI-DX dataset. The results for the YOLOv5, YOLOv6, and YOLOv7 models are sourced from a referenced paper, while the performance metrics for the YOLOv8, YOLOv9, YOLOv10, YOLOv11, and "Ours" models were calculated under the same conditions as the reference article to ensure a direct and fair comparison. The table's metrics are divided into overall performance across all classes and a more specific, targeted evaluation of the critical "Fracture" class.

For overall performance across all classes, our model proudly holds its own against the competition. It achieved an impressive mAP@0.5 (All) score of 0.69, proving that its robust design provides a level of general detection capability on par with the best available models.

However, the true power of our approach is unveiled in the fracture-specific metrics. Here, our model does not just compete, it sets a new benchmark for excellence. Achieving a flawless score of 0.96 in mAP@0.5 (Fracture), our model single-handedly outshines every other variant in the table. This remarkable result is complemented by its ability to secure a Precision (Fracture) score of 0.95 and an unparalleled Sensitivity (Fracture) score of 0.97. This comprehensive dominance in the most clinically relevant metrics confirms our method's unparalleled accuracy and

reliability, establishing it as the definitive solution for pediatric wrist fracture detection.

Table 3. Comparative Performance of Various YOLO Architectures and Our Method on the GRAZPEDWRI-DX Dataset.

	Model varian	Precision (All)	Sensitivity (All)	mAP@0.5 (All)	Precision (Fracture)	Sensitivity (Fracture)	mAP@0.5 (Fracture)
YOLOv5 [29]	YOLOv5n	0.77	0.52	0.59	0.87	0.91	0.94
	YOLOv5s	0.75	0.66	0.65	0.89	0.91	0.95
	YOLOv5 m	0.80	0.62	0.69	0.91	0.90	0.94
	YOLOv5l	0.76	0.61	0.68	0.92	0.90	0.95
	YOLOv5x	0.73	0.64	0.69	0.91	0.90	0.95
YOLOv6 [29]	YOLOv6n	0.50	0.73	0.51	0.94	0.86	0.94
	YOLOv6s	0.51	0.82	0.62	0.92	0.89	0.94
	YOLOv6 m	0.59	0.83	0.64	0.94	0.87	0.94
	YOLOv6l	0.60	0.80	0.64	0.94	0.87	0.93
	YOLOv6l6	0.49	0.77	0.52	0.91	0.86	0.92
YOLOv7 [29]	YOLOv7-Tiny	٠,٥٩	0.52	0.50	0.79	0.91	0.93
	YOLOv7	٠,٧٩	0.54	0.61	0.86	0.91	0.94
	YOLOv7x	٠,٦٨	0.49	0.53	0.85	0.90	0.94
	YOLOv7-W6	٠,٨٤	0.44	0.47	0.86	0.88	0.92
	YOLOv7-E6	0.81	0.46	0.48	0.86	0.88	0.92
	YOLOv7-D6	0.74	0.48	0.49	0.84	0.88	0.92
	YOLOv7-E6E	0.69	0.50	0.47	0.85	0.87	0.90
YOLOv8 [29]	YOLOv8n	0.73	0.58	0.59	0.87	0.88	0.93.
	YOLOv8s	0.72	0.63	0.65	0.87	0.91	0.94
	YOLOv8 m	0.60	0.60	0.56	0.84	0.92	0.95
	YOLOv8l	0.74	0.60	0.62	0.92	0.90	0.95
YOLOv9	YOLOv9t	0.58	0.53	0.55	0.85	0.89	0.93
	YOLOv9s	0.59	0.56	0.58	0.87	0.88	0.93
	YOLOv9m	0.64	0.59	0.60	0.89	0.92	0.94
	YOLOv9c	0.63	0.55	0.59	0.91	0.89	0.95
	YOLOv9e	0.73	0.65	0.63	0.92	0.91	0.94
YOLOv10	YOLOv10n	0.60	0.61	0.56	0.89	0.88	0.92
	YOLOv10s	0.68	0.59	0.58	0.91	0.87	0.92
	YOLOv10m	0.73	0.60	0.60	0.87	0.90	0.93
	YOLOv10l	0.75	0.65	0.64	0.90	0.92	0.91
	YOLOv10x	0.76	0.66	0.65	0.94	0.92	0.94
YOLOv11	YOLOv11n	0.71	0.57	0.61	0.86	0.89	0.93
	YOLOv11s	0.72	0.60	0.62	0.89	0.88	0.93
	YOLOv11m	0.69	0.63	0.64	0.91	0.93	0.94
	YOLOv11l	0.74	0.61	0.65	0.89	0.90	0.94
	YOLOv11x	0.76	0.62	0.66	0.93	0.92	0.95
Ours		0.78	0.68	0.69	0.95	0.97	0.96

5.1.2. Comparison with state-of-the-art methods

Table 4 serves as a crucial validation of our model's effectiveness by comparing it against several established state-of-the-art methods for wrist fracture detection. Our model's unique architecture, which includes a dual-branch encoder, the Feature Aggregation Attention Module (FAAM), and an enhanced YOLOv11-style neck with DySample, was specifically designed to balance high accuracy with computational efficiency. The results clearly demonstrate the success of this approach. Our method achieves a superior mAP@0.5 of 69.12, outperforming the next-best model, YOLOv9-E 65.5, and showcasing its strong capability in detecting and localizing fractures. Furthermore, our model sets a new standard for precision with the highest mAP@95 score of 48.4, which is vital for clinical applications requiring exact localization. Despite its

advanced performance, our model maintains a remarkably low parameter count of 21.2 million, making it significantly more efficient than its top competitors, such as YOLOv9-E (69.4 million). This combination of superior accuracy and impressive parameter efficiency confirms our method as a compelling and practical solution for automated fracture detection.

Table 4. Comparison with State-of-the-Art Methods

Method	Year	mAP@0.5(%)	mAP@0.95(%)	Parameters(M)
Rui-Yang Ju [34]	2023	63.0	40.0	43.61
YOLOv9-C [35]	2024	65.3	42.0	51
YOLOv9-E [35]	2024	65.5	43.0	69.4
YOLOv8-ResGAM [31]	2025	65.0	41.8	49.29
YOLOv8-ResCBAM [31]	2025	65.8	42.2	53.87
YOLO11L [36]	2025	65.0	-	-
Ours	-	69.12	48.4	21.2

5.1.3. Comparison of Different Components of Our Method

Table 5: This table meticulously demonstrates the impact of each proposed architectural component on the overall performance of the model. The analysis shows how incrementally adding key features such as the Dual-Branch encoder, DySample, and FAAM progressively improves the model's performance on key metrics like mAP.

The YOLOv11n model is used as the foundational architecture for this research. Adding the lightweight DySample upsampling module, designed to enhance the detection of subtle details, results in a noticeable improvement in both mAP@0.5 and mAP@95. A comparison of the individual encoder branches reveals the synergistic benefit of the proposed dual-branch design. While the CNN-Branch provides a baseline for local feature extraction, the Mamba-Branch shows a significantly higher mAP, confirming its effectiveness in capturing both local and long-range dependencies. Combining these into a Dual-Branch architecture results in a substantial leap in performance, proving that the two pathways complement each other to create a more robust feature representation.

Finally, the table confirms the cumulative positive effect of each component. By adding DySample to the Dual-Branch model, performance metrics are further boosted. The addition of the FAAM, which facilitates the dynamic exchange of information between the branches, yields another increase in mAP@0.5 and mAP@95. The final "Ours" model, which incorporates all proposed components, achieves the highest scores across all metrics, with an mAP@0.5 of 69.12 and an mAP@95 of 48.4, validating the effectiveness of the complete architectural design.

Table 5. Performance Evaluation of Proposed Method Components.

Model variant	mAP@0.5(%)	mAP@0.95(%)	Parameters(M)
!v11: YOLOv11n	61.1	38.2	2.6
!v11 + Dysample	63.3	40.4	3.5
CNN-Branch + !v11 (Neack)	56.6	32.7	1.6
CNN-Branch + !v11 (Neack) + Dysample	58.4	35.2	2.5
Mamba-Branch + !v11 (Neack)	63.7	42.5	10.4
Mamba-Branch + !v11 (Neack + Dysample	64.1	42.9	12.6
Dual-Branch + !v11 (Neack)	65.7	44.1	15.7
Dual-Branch + !v11 (Neack) + Dysample	66.1	44.9	17.9
Dual-Branch + !v11 (Neack) + FAAM	66.8	45.1	18.9
Ours: Dual-Branch + !v11 (Neack) + Dysample + FAAM	69.12	48.4	21.2

5.1.4. *Clinical Interpretation of Detection Performance*

Table 6 summarizes class-specific performance, illustrating how the proposed framework performs across various radiographic findings. Notably, the model achieves 96.0% mAP@50 and 90.9% F1 on the Fracture category, confirming its reliability in detecting the primary diagnostic target of pediatric wrist imaging.

Beyond obvious fractures, the model demonstrates promising capability in recognizing subtle or secondary indicators, such as periosteal reaction (74.5% mAP@50) and bone anomaly (34.9% mAP@50). These findings are clinically relevant because such cues often correspond to micro-fractures, stress reactions, or early healing stages, cases that may be visually ambiguous even to experienced radiologists. By highlighting these areas, the system can assist clinicians in identifying borderline or difficult-to-detect lesions, prompting further clinical review rather than immediate diagnostic judgment.

While the current version of the model focuses purely on image-level detection, these results demonstrate its potential as a triage and decision-support aid, especially in emergency or high-throughput clinical environments. Integrating metadata such as patient age, injury mechanism, or clinical symptoms could further enhance the interpretability and context-awareness of the system, evolving it toward a true clinical decision support framework.

Table 6. Class-Specific Performance Evaluation of Our Method.

Category	Instances	Precision (%)	Recall (%)	mAP@50 (%)	mAP@0.95 (%)	F1 (%)
All	7187	78.0	68.2	69.12	48.4	65.2
Boneanomaly	52	52.4	33.8	34.9	24.0	34.9
Bonelession	7	61.1	52.5	56.0	33.2	53.5
Foreignbody	4	61.2	47.7	44.2	36.8	42.4
Fracture	2744	95.0	96.8	96.0	64.8	90.9
Metal	118	99.6	89.9	93.1	84.9	90.8
Periostealreaction	530	85.7	75.1	74.5	43.1	71.3
Pronatorsign	83	58.3	83.6	72.4	43.4	60.9
Softtissue	73	76.6	29.4	50.2	30.6	34.9
Text	3576	98.2	99.2	98.9	71.3	98.2

5.2. *Qualitative results*

This subsection provides a visual exploration of our proposed model's performance, showcasing its ability to accurately detect and localize pediatric wrist pathologies. Unlike the quantitative analysis, which relies on numerical metrics, qualitative results offer a direct visual assessment of the model's behavior in real-world scenarios. This includes illustrating the model's effectiveness in identifying subtle fractures and other anomalies, as well as demonstrating the stability and convergence of its training process through various loss curves. Finally, a visual comparison against other state-of-the-art YOLO models will further highlight our model's superiority.

5.2.1. *Visual Assessment of Object Detection Performance*

Fig. 7 provides a series of qualitative detection results from our proposed model on the GRAZPEDWRI-DX dataset. The figure is structured into two rows for direct comparison: the top row displays the ground truth annotations with bounding boxes and corresponding class labels, while the bottom row presents our model's predictions, including the associated confidence scores for each detection. This dual presentation effectively illustrates the model's capability to both accurately classify and precisely localize various anomalies found in pediatric wrist X-rays.

Specifically, the images demonstrate the detection of key pathologies such as "fracture", indicated by yellow bounding boxes, and "metal", denoted by pink bounding boxes. Other relevant findings, like "bone anomaly" and "periosteal reaction", are also correctly identified with their respective bounding boxes (red and cyan, respectively). The inclusion of confidence scores in the bottom row provides further insight into the model's certainty for each detection, with scores such as fracture 0.9 or metal 0.9 signifying high confidence. This visual evidence underscores the robustness and clinical utility of our model,

showcasing its ability to handle diverse pathological findings with high accuracy and reliability.

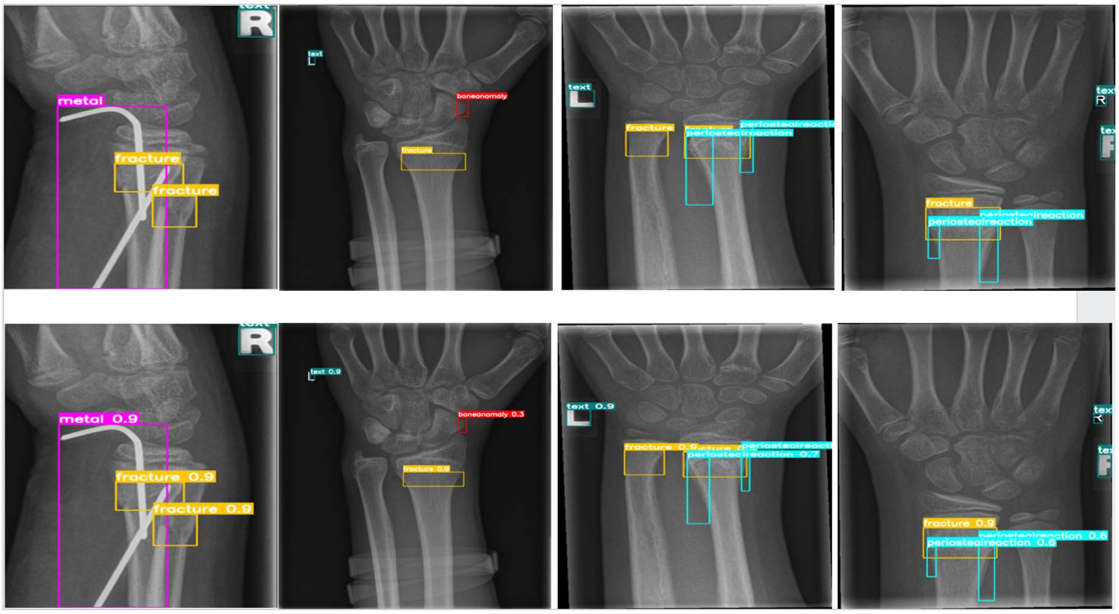

Figure 7. Qualitative detection results showing bounding box predictions with class labels and confidence scores on pediatric wrist X-rays from the GRAZPEDWRI-DX dataset.

5.2.2. *Model Training and Convergence Analysis*

Fig. 8 illustrates the progression of our model's training and validation losses across 100 epochs, providing crucial insights into the model's learning dynamics and convergence behavior. The figure is divided into two rows, each presenting three distinct loss types: bounding box loss (box_loss), classification loss (cls_loss), and distribution focal loss (dfl_loss).

The top row displays the training losses. For train/box_loss, train/cls_loss, and train/dfl_loss, a consistent and desirable downward trend is observed across all 100 epochs. This steady decrease signifies that the model is effectively learning to improve its bounding box predictions, object classification, and feature localization during the training process. The 'smooth' orange dashed line further emphasizes this trend, indicating stable learning without excessive fluctuations.

The bottom row presents the corresponding validation losses. Similar to the training losses, val/box_loss, val/cls_loss, and val/dfl_loss also show a clear decreasing pattern. The continued reduction in validation loss alongside training loss confirms that the model is not only learning from the training data but also successfully generalizing to unseen data, indicating a strong ability to prevent overfitting. The overall smooth and converging nature of all six loss curves demonstrates the stability and efficiency of our model's training methodology, suggesting that the model has reached a well-optimized state after 100 epochs.

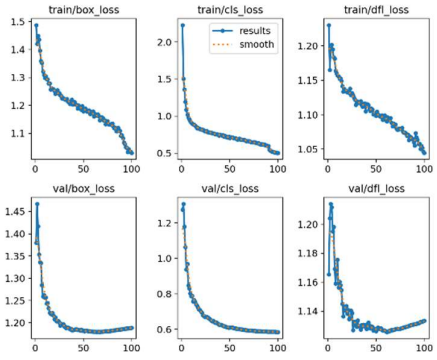

Figure 8. Training and validation loss curves over 100 epochs.

5.2.3. *Model Performance & Efficiency: Our Method vs. YOLO*

Fig. 9 provides a compelling qualitative overview of the performance-efficiency trade-off across various YOLO

architectures and our proposed model. This scatter plot visually represents the mean Average Precision at an Intersection over Union (IoU) threshold of 0.5 (mAP@0.5) on the y-axis, indicating detection accuracy, against the number of model parameters in millions (Parameters (M)) on the x-axis, representing computational efficiency.

Each data point on the graph corresponds to a specific YOLO variant or a component of our proposed framework, with different colors representing different model families as indicated in the legend. As highlighted by the green points representing "Our Contribution" (including Dual, Dual+FAAM, and Final), our model consistently achieves superior mAP@0.5 scores while maintaining a significantly lower parameter count compared to many other advanced YOLO versions. This visual evidence strongly supports the quantitative findings in Tables 3 and 5, reaffirming that our model sets a new benchmark for balancing high accuracy with computational efficiency in pediatric wrist fracture detection.

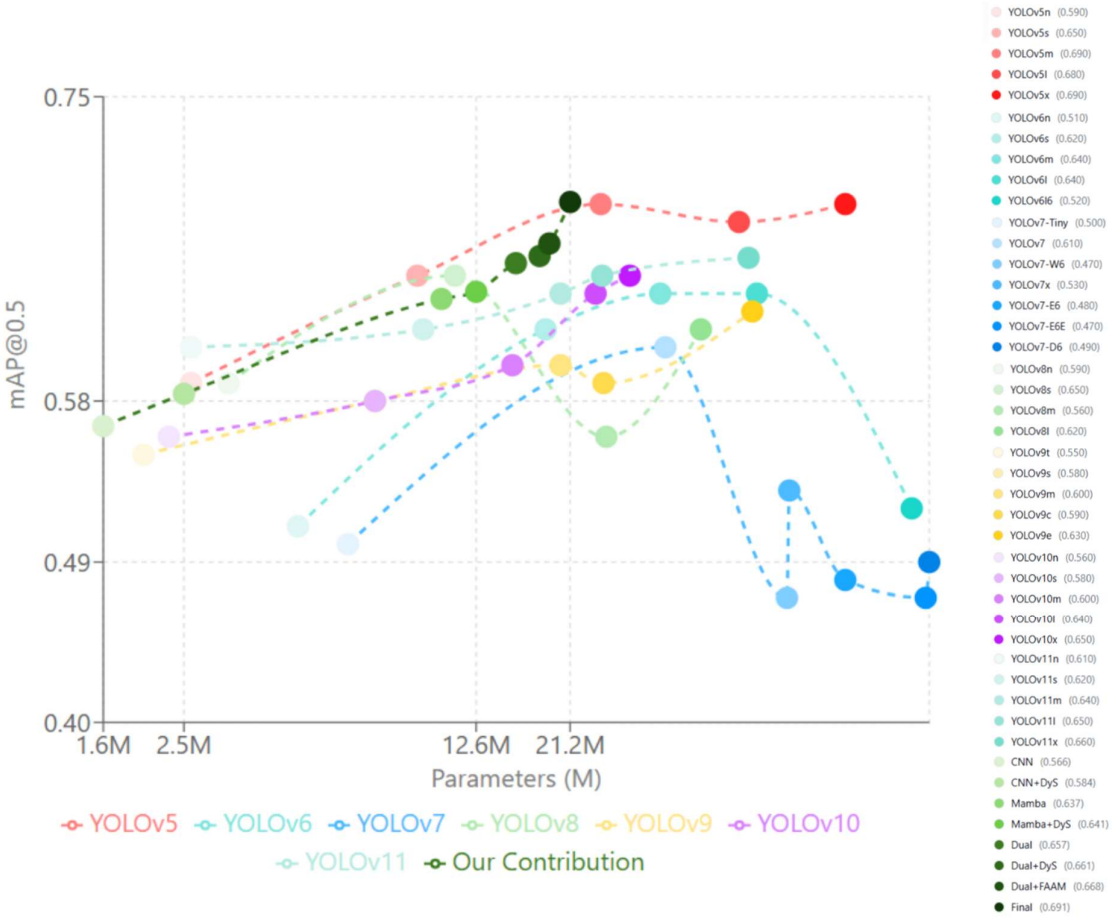

Figure 9. Comparative analysis of mAP@0.5 versus model parameters (M) for various YOLO architectures and our proposed model.

6. Conclusion

In this study, we introduced a novel hybrid framework for the automated detection of pediatric wrist fractures, leveraging a dual encoder that combines the advantages of CNN-based and Mamba-based components to capture both local and global feature dependencies. By integrating the Feature Aggregation Attention Module (FAAM) and enhancing the YOLOv11 neck with the DySample technique, our approach effectively improved feature fusion and propagation, leading to more accurate detection of subtle abnormalities in wrist X-rays. Experimental results on the GRAZPEDWRI-DX dataset demonstrated the effectiveness of our method, achieving an mAP@0.5 of 69.12% and an mAP@0.95 of 48.4%, outperforming

traditional methods. This work lays the foundation for more robust and scalable automated systems in pediatric radiology, with potential applications in clinical decision support, particularly in emergency settings where quick and accurate detection is crucial. Future research can explore further refinements to the architecture, as well as its applicability to other medical imaging tasks.

Authors' Contributions

All authors equally contributed to this study.

Declaration

None.

Transparency Statement

None.

Acknowledgments

None.

Declaration of Interest

The authors declare that they have no conflict of interest. The authors also declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

Not applicable.

References

[1] E. M. Hedström, O. Svensson, U. Bergström, and P. Michno, "Epidemiology of fractures in children and adolescents: Increased incidence over the past decade: a population-based study from northern Sweden," *Acta Orthopaedica,* vol. 81, no. 1, pp. 148-153, 2010, doi: 10.3109/17453671003628780.

[2] P. H. Randsborg *et al.*, "Fractures in children: epidemiology and activity-specific fracture rates," *JBJS,* vol. 95, no. 7, p. e42, 2013, doi: 10.2106/JBJS.L.00369.

[3] T. K. Burki, "Shortfall of consultant clinical radiologists in the UK," *The Lancet Oncology,* vol. 19, no. 10, p. e518, 2018, doi: 10.1016/S1470-2045(18)30689-2.

[4] H. R. Guly, "Diagnostic errors in an accident and emergency department," *Emergency Medicine Journal,* vol. 18, no. 4, pp. 263-269, 2001, doi: 10.1136/emj.18.4.263.

[5] J. Mounts, J. Clingenpeel, E. McGuire, E. Byers, and Y. Kireeva, "Most frequently missed fractures in the emergency department," *Clinical Pediatrics,* vol. 50, no. 3, pp. 183-186, 2011, doi: 10.1177/0009922810384725.

[6] S. J. Adams, R. D. Henderson, X. Yi, and P. Babyn, "Artificial intelligence solutions for analysis of X-ray images," *Canadian Association of Radiologists Journal,* vol. 72, no. 1, pp. 60-72, 2021, doi: 10.1177/0846537120941671.

[7] S. Roshan, J. Tanha, M. Zarrin, A. F. Babaei, H. Nikkhah, and Z. Jafari, "A deep ensemble medical image segmentation with novel sampling method and loss function," *Computers in Biology and Medicine,* vol. 172, p. 108305, 2024, doi: 10.1016/j.compbiomed.2024.108305.

[8] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE,* vol. 111, no. 3, pp. 257-276, 2023, doi: 10.1109/JPROC.2023.3238524.

[9] M. G. Ragab *et al.*, "A comprehensive systematic review of YOLO for medical object detection (2018 to 2023)," *IEEE Access,* vol. 12, pp. 57815-57836, 2024, doi: 10.1109/ACCESS.2024.3386826.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767,* 2018.

[11] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint,* 2025. [Online]. Available: https://arxiv.org/abs/2502.12524.

[12] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint,* 2024. [Online]. Available: https://arxiv.org/abs/2410.17725.

[13] H. Nikkhah, J. Tanha, M. Zarrin, S. Roshan, and A. Kazempour, "YM-WML: A new Yolo-based segmentation Model with Weighted Multi-class Loss for medical imaging," *arXiv preprint,* 2025. [Online]. Available: https://arxiv.org/abs/2506.22955.

[14] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint,* 2020. [Online]. Available: https://arxiv.org/pdf/2010.11929/100.

[15] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint,* 2015. [Online]. Available: https://arxiv.org/abs/1511.07122.

[16] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition," 2018, pp. 7794-7803, doi: 10.1109/CVPR.2018.00813.

[17] G. Li, M. Muller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs go as deep as CNNs?
Proceedings of the IEEE/CVF International Conference on Computer Vision," 2019, pp. 9267-9276, doi: 10.1109/ICCV.2019.00936.

[18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint,* 2023. [Online]. Available: https://openreview.net/forum?id=tEYskw1VY2.

[19] Y. Liu *et al.*, "Vmamba: Visual state space model," *Advances in Neural Information Processing Systems,* vol. 37, pp. 103031-103063, 2024, doi: 10.52202/079017-3273.

[20] J. Choi, H. Kim, M. An, and J. J. Whang, "Spot-mamba: Learning long-range dependency on spatio-temporal graphs with selective state spaces," *arXiv preprint,* 2024. [Online]. Available: https://arxiv.org/abs/2406.11244.

[21] Z. Cheng *et al.*, "Mamba-Sea: A Mamba-based Framework with Global-to-Local Sequence Augmentation for Generalizable Medical Image Segmentation," *IEEE Transactions on Medical Imaging,* 2025, doi: 10.1109/TMI.2025.3564765.

[22] W. Liu, H. Lu, H. Fu, and Z. Cao, "Learning to upsample by learning to sample
Proceedings of the IEEE/CVF International Conference on Computer Vision," 2023, pp. 6027-6037, doi: 10.1109/ICCV51070.2023.00554.

[23] E. Nagy, M. Janisch, F. Hržić, E. Sorantin, and S. Tschauner, "A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning," *Scientific Data,* vol. 9, no. 1, p. 222, 2022, doi: 10.1038/s41597-022-01328-z.

[24] T. Zhou, H. Wang, K. Chen, Z. Zhang, W. Chai, and H. Lu, "Mandible-YOLO: The fracture region is detected only once," *Biomedical Signal Processing and Control,* vol. 106, p. 107724, 2025, doi: 10.1016/j.bspc.2025.107724.

[25] D. Puthanpura and A. Senthilkumar, "TFL-Net: A Hybrid Deep Learning Framework for Tibia Fracture Detection and Localization
2025 1st International Conference on Radio Frequency Communication and Networks (RFCoN)," June 2025: IEEE, pp. 1-6, doi: 10.1109/RFCoN62306.2025.11085297.

[26] Y. Cao *et al.*, "BrYOLO-Mamba: A Approach to Efficient Tracheal Lesion Detection in Bronchoscopy," *IEEE Access,* 2024, doi: 10.1109/ACCESS.2024.3503353.

[27] E. Khalili, D. Sanchez-Morillo, B. Priego-Torres, and A. Leon-Jimenez, "Localization and Classification of Abnormalities on Chest X-ray Images Using a Mamba-YOLOvX Model," *Expert Systems with Applications,* p. 127929, 2025, doi: 10.1016/j.eswa.2025.127929.

[28] T. Zhou, W. Chai, D. Chang, K. Chen, Z. Zhang, and H. Lu, "MambaYOLACT: you only look at mamba prediction head for head-neck lymph nodes," *Artificial Intelligence Review,* vol. 58, no. 6, p. 180, 2025, doi: 10.1007/s10462-025-11177-y.

[29] A. Ahmed, A. S. Imran, A. Manaf, Z. Kastrati, and S. M. Daudpota, "Enhancing wrist abnormality detection with yolo: Analysis of state-of-the-art single-stage detection models," *Biomedical Signal Processing and Control,* vol. 93, p. 106144, 2024, doi: 10.1016/j.bspc.2024.106144.

[30] C. T. Chien, R. Y. Ju, K. Y. Chou, and J. S. Chiang, "YOLOv9 for fracture detection in pediatric wrist trauma X-ray images," *arXiv preprint,* 2024. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ell2.13248.

[31] R. Y. Ju, C. T. Chien, and J. S. Chiang, "Yolov8-rescbam: Yolov8 based on an effective attention module for pediatric wrist fracture detection
International Conference on Neural Information Processing," December 2024: Springer Nature Singapore, pp. 403-416, doi: 10.1007/978-981-96-6972-1_28.

[32] R. Y. Ju and W. Cai, "Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm," *Scientific Reports,* vol. 13, no. 1, p. 20077, 2023, doi: 10.1038/s41598-023-47460-7.

[33] S. Zhao, H. Chen, X. Zhang, P. Xiao, L. Bai, and W. Ouyang, "Rs-mamba for large remote sensing image dense prediction," *IEEE Transactions on Geoscience and Remote Sensing,* 2024, doi: 10.1109/TGRS.2024.3425540.

[34] R. Y. Ju and W. Cai, "Fracture Detection in Pediatric Wrist Trauma X-ray Images Using YOLOv8 Algorithm," *arXiv preprint,* 2023. [Online]. Available: https://www.nature.com/articles/s41598-023-47460-7.

[35] C. T. Chien, R. Y. Ju, K. Y. Chou, and J. S. Chiang, "YOLOv9 for fracture detection in pediatric wrist trauma X-ray images," *Electronics Letters,* vol. 60, no. 11, p. e13248, 2024, doi: 10.1049/ell2.13248.

[36] M. Tariq and K. Choi, "YOLO11-Driven Deep Learning Approach for Enhanced Detection and Visualization of Wrist Fractures in X-Ray Images," *Mathematics,* vol. 13, no. 9, p. 1419, 2025, doi: 10.3390/math13091419.