algorithms

MDPI

Article

OrthoDETR: A Streamlined Transformer-Based Approach for Precision Detection of Orthopedic Medical Devices

Xiaobo Zhang [1], Huashun Li [2,*], Jingzhao Li [2] and Xuehai Zhou [1]

[1] School of Computer Science and Technology, University of Science and Technology of China, Hefei 230001, China; xiaobzhang@mail.ustc.edu.cn (X.Z.); xhzhou@ustc.edu.cn (X.Z.)
[2] College of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan 232001, China; jzhli@aust.edu.com
* Correspondence: 2021200797@aust.edu.cn

Abstract: The rapid and accurate detection of orthopedic medical devices is pivotal in enhancing health care delivery, particularly by improving workflow efficiency. Despite advancements in medical imaging technology, current detection models often fail to meet the unique requirements of orthopedic device detection. To address this gap, we introduce OrthoDETR, a Transformer-based object detection model specifically designed and optimized for orthopedic medical devices. OrthoDETR is an evolution of the DETR (Detection Transformer) model, with several key modifications to better serve orthopedic applications. We replace the ResNet backbone with the MLP-Mixer, improve the multi-head self-attention mechanism, and refine the loss function for more accurate detections. In our comparative study, OrthoDETR outperformed other models, achieving an AP50 score of 0.897, an AP50:95 score of 0.864, an AR50:95 score of 0.895, and a frame per second (FPS) rate of 26. This represents a significant improvement over the DETR model, which achieved an AP50 score of 0.852, an AP50:95 score of 0.842, an AR50:95 score of 0.862, and an FPS rate of 20. OrthoDETR not only accelerates the detection process but also maintains an acceptable performance trade-off. The real-world impact of this model is substantial. By facilitating the precise and quick detection of orthopedic devices, OrthoDETR can potentially revolutionize the management of orthopedic workflows, improving patient care, and enhancing the efficiency of healthcare systems. This paper underlines the significance of specialized object detection models in orthopedics and sets the stage for further research in this direction.

Keywords: orthopedic medical devices; MLP-Mixer; DETR; Transformer; multi-head self-attention mechanism

Citation: Zhang, X.; Li, H.; Li, J.; Zhou, X. OrthoDETR: A Streamlined Transformer-Based Approach for Precision Detection of Orthopedic Medical Devices. *Algorithms* **2023**, *16*, 550. https://doi.org/10.3390/a16120550

Academic Editors: Brian Azzopardi and Surender Reddy Salkuti

Received: 17 October 2023
Revised: 15 November 2023
Accepted: 20 November 2023
Published: 29 November 2023

1. Introduction

Orthopedic medical devices play a pivotal role in the treatment and management of various musculoskeletal disorders [1–3]. Accurate and efficient detection of these devices is paramount for meticulous management during the workflow process. This, in turn, can significantly enhance the overall efficiency and safety of orthopedic treatments [4].

Existing object detection models, dominantly those based on convolutional neural networks (CNNs), have shown promising results across different application domains [5–7]. However, they often fall short when applied to orthopedic medical device detection due to their inability to address the specificity and complexity of medical imaging data and the variability in device appearances. The conventional CNN-based models are not optimally designed to handle these challenges, highlighting a pressing need for more robust and adaptive approaches.

Acknowledging the current limitations in orthopedic medical device detection using deep learning models, we introduce OrthoDETR, a specialized Transformer-based object detection strategy meticulously designed for the orthopedic domain. The traditional DEtection TRansformer (DETR) model, despite its strengths, suffers from certain inefficiencies

and inadequacies when applied to orthopedics, necessitating improvements specifically tailored to this context [8]. For instance, the ResNet backbone [9], while generally competent, is outperformed by the MLP-Mixer [10] on tasks such as image classification, due to the latter's higher computational efficiency and lower parameter count. As such, replacing ResNet with MLP-Mixer in our model would likely lead to enhanced performance and efficiency. Similarly, the standard multi-head self-attention mechanism central to the Transformer model comes with high computational complexity, particularly problematic when handling long sequences. By optimizing this mechanism, we aim to reduce computational burdens without compromising the ability to capture rich contextual information. Lastly, traditional loss functions like cross-entropy loss may not be suitable for dealing with more complex orthopedic tasks, such as those involving class imbalances or requiring precise probability predictions. With an improved loss function, OrthoDETR can better tackle these challenges, further boosting model performance. Through these strategic modifications to the DETR architecture, OrthoDETR promises significant advancements in the field of orthopedic medical device detection and broader medical imaging analysis.

The primary contributions of this study can be delineated as follows:

(1) We introduce a unique object detection strategy, termed as OrthoDETR, that is specifically designed for the efficient and accurate identification of orthopedic medical devices. It counters the limitations of existing models and operates effectively against the challenges emerged from intricate medical imaging data and varied device appearances.
(2) OrthoDETR extends the core architecture of DETR (Detection Transformer) and incorporates several significant innovations to better accommodate the orthopedic domain. These key enhancements involve substituting the ResNet backbone with an MLP-Mixer for superior feature extraction, refining the multi-head self-attention mechanism for enhanced context comprehension, and adjusting the loss function for optimized model training.
(3) Through rigorous experimentation, we demonstrate that OrthoDETR provides considerable improvements in detection speed, while only resulting in a slight decrease in performance. This makes it a valuable tool for detecting orthopedic medical devices, particularly in the context of fine-grained management during workflow processes.

For further clarity, the remainder of this paper is structured as follows: Section 2 offers a concise overview of the existing work in object detection and orthopedic medical device detection. Section 3 elaborates on the proposed OrthoDETR model, including its detailed architecture and optimization strategies. The experimental setup, results, and comparison with other advanced methods are presented in Section 4. Finally, Section 5 concludes the paper and offers insights into future avenues of research.

2. Related Work

The field of object detection has undergone significant evolution, particularly with the development of deep learning techniques [11–13]. Initially, the focus was primarily on Convolutional Neural Networks (CNNs), which have been the backbone of many groundbreaking advancements in image analysis. CNNs, with their ability to extract and learn features from images, have been pivotal in tasks ranging from facial recognition to autonomous driving.

However, recent trends have seen a shift towards Transformer-based models, which were originally developed for natural language processing tasks [14–16]. These models, unlike CNNs, are adept at handling sequential data and can capture long-range dependencies in data, making them particularly suitable for complex tasks like object detection. Transformers have shown remarkable capabilities in understanding the context and relationships within an image, leading to more accurate and efficient object detection systems.

One of the key strengths of Transformer models is their scalability and ability to handle large datasets, which is crucial in fields like medical imaging and autonomous driving where large amounts of data are processed [17]. This transition from CNNs to

Transformer-based models marks a significant shift in the landscape of object detection, offering new possibilities and efficiencies in various applications.

In the realm of medical image analysis, deep learning has achieved remarkable results for various problems. For instance, Mathesul and others proposed a deep learning method based on CNNs to enhance the detection of COVID-19 and its variants from chest X-ray images [18]. Furthermore, Sakaida and colleagues developed a method for detecting breast calcifications using deep learning, capable of classifying the presence of calcifications in the breast [19].

Significant research has also been conducted in the field of object detection. Carballo and team [20] introduced a new method based on computer vision and object detection technologies, utilizing the Convolutional Neural Network EfficientDet-D2 model, for cloud detection in image sequences. Sami and others [21] proposed an improved deep neural network model, based on YOLOv5, for real-time detection of road surface damage in photographic representations of outdoor road surfaces.

Efforts have been made to improve object detection methods for medical image applications. For example, U-Net architecture was proposed to improve the adaptability to medical image features by introducing domain expert knowledge and specific data enhancement strategies [22]. A multi-level deep learning framework for lung nodule detection was proposed using multi-scale and multi-level features [23]. However, these approaches do not adequately address the unique challenges presented by orthopedic medical device detection.

Additionally, some studies have combined deep learning with other types of features to enhance detection performance. For example, Grignaffini and team [24] presented a novel approach using CNNs for melanoma detection, incorporating manual texture features of dermatoscopic images as additional input during the training phase.

Specific improvements in networks have also emerged in particular application domains. For instance, Wang and others proposed a MobileNet V2 network with a dual attention mechanism, enabling spatial and channel dimension operations at the network level for plant disease identification [25]. Apostolopoulos and team introduced an innovative improvement of the VGG19 network (ParaNet+), used for classifying flicker images into normal and abnormal categories, applying the Grad-CAM++ algorithm to locate abnormal parathyroid glands [26].

However, despite these advancements, existing detection models often fail to meet the unique requirements of orthopedic medical devices. To address this gap, we propose OrthoDETR, a Transformer-based object detection model specifically designed and optimized for orthopedic medical equipment. The introduction of OrthoDETR not only accelerates the detection process, but also maintains an acceptable performance trade-off. By facilitating precise and rapid detection of orthopedic devices, OrthoDETR has the potential to revolutionize the management of orthopedic workflows, improve patient care, and enhance the efficiency of healthcare systems.

3. Materials and Methods

In the present study, we enhance and fine-tune the DETR model to cater to the unique attributes and requirements of orthopedic medical devices. This involves substituting the initial network structure from ResNet to MLP-Mixer, refining the multi-headed self-attention mechanism to elevate the model's performance, and recalibrating the loss function to more accurately reflect the features of orthopedic medical devices. These enhancements aim to deliver an advanced and efficient object detection model specifically tailored for the orthopedic medical device sector.

To provide a more comprehensive theoretical foundation, we delve into the underlying principles that guide our methodology. The MLP-Mixer, replacing the ResNet in our architecture, is an innovative concept that leverages the idea of mixing token-wise and channel-wise information in a permutationally invariant manner, providing a robust and efficient way to handle the complexity and variability of orthopedic medical devices.

The refinement of the multi-headed self-attention mechanism is based on the theoretical understanding that attention mechanisms can model interactions between various parts of the image, which is crucial for identifying and localizing medical devices in complex orthopedic images. By refining this mechanism, we aim to capture the interdependencies between the different parts of the image more effectively.

The recalibration of the loss function is informed by the theoretical principle that the loss function should be closely aligned with the task objective. In the context of orthopedic medical device detection, the features of such devices are distinct and can vary significantly. Therefore, we recalibrate the loss function to better reflect these unique features and improve the model's ability to detect orthopedic medical devices.

By elucidating these theoretical principles and concepts, we hope to increase the academic value of our manuscript and provide a deeper understanding of our proposed methodology. The following section will delve into the intricate implementation details of our proposed model.

3.1. Improved DETR Model

Our proposed methodology is rooted in the DETR model, a pioneering approach that utilizes the Transformer for object detection tasks. It treats target detection as a straightforward ensemble prediction problem, using the encoder-decoder structure of the Transformer to model the global relationships between images and objects, while using a global loss function based on bilateral matching to enforce the uniqueness of the prediction results. DETR does not require any prior knowledge and post-processing steps, simplifying the target detection process and achieving results comparable to existing methods on the COCO dataset. However, DETR does exhibit certain limitations, such as the necessity for numerous training iterations required and suboptimal detection performance for compact and densely clustered targets.

To overcome these drawbacks and adapt to the specific domain of orthopedic medical devices, we have made some improvements and optimizations to DETR, specifically replacing the underlying network structure from ResNet to MLP-Mixer, optimizing the multi-headed self-attention mechanism to improve model performance, and adapting the loss function to better match the characteristics of orthopedic medical devices. The OrthoDETR network structure proposed in this paper is illustrated in Figure 1. We will present each of these improvements and optimizations below.

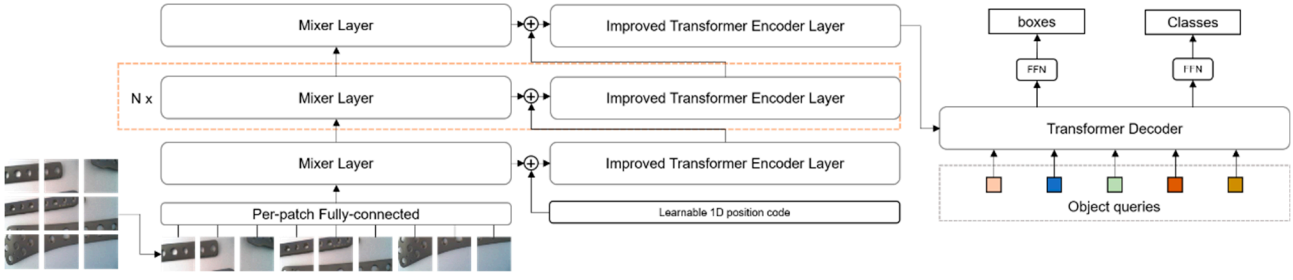

Figure 1. The proposed OrthoDETR network structure.

3.1.1. ResNet Replacement for MLP-Mixer

The DETR model employs a convolutional neural network as its backbone network to extract features from images, which are subsequently processed by the Transformer's encoder and decoder. However, DETR only utilizes the final layer of features from the backbone network, resulting in potential information loss. This is particularly problematic for compact and densely packed targets where features from lower layers could be more beneficial for detection. To address this issue, we substitute the backbone network from ResNet to MLP-Mixer, and merge the output of the Mixer layer with the transformer encoder layer. This allows the model to integrate features from diverse layers, thereby enhancing the feature representation and multi-scale information.

The MLP-Mixer is an innovative neural network architecture that incorporates a Transformer-inspired approach of reconfiguring the input data into token and embedding formats. It facilitates feature integration across channels and sequences via a multi-layer MLP. This stands in contrast to the conventional ResNet, where feature extraction and fusion processes occur separately. By amalgamating these processes, the MLP-Mixer enhances both the efficiency and accuracy of the network.

The MLP-Mixer serves as the backbone network for feature extraction within the DETR model, its primary objective being to thoroughly amalgamate inter-channel and inter-sequence features. This is achieved by initially transforming images into token and embedding formats. Such a methodology not only boosts the network's efficiency and accuracy, but it also circumvents the information bottleneck issue inherent in ResNet. As a result, the incorporation of the MLP-Mixer into the feature extraction component of DETR can significantly enhance the network's overall performance and efficiency.

In addition, adopting MLP-Mixer as the feature extraction part of DETR has the added benefit of better handling the dimensionality requirements of the input data. In a traditional ResNet, the input data needs to meet specific dimensional requirements, which often limits the scope of network application. In contrast, the use of MLP-Mixer allows for more flexible handling of the input data, making it suitable for a wider range of application scenarios.

Specifically, the output of the mixer layer is summed with the output of the transformer encoder layer and then fed into the next transformer encoder layer. This allows the layer to dynamically assign different weights to different levels of features according to their spatial location and scale.

3.1.2. Improved Transformer Encoder

The Transformer Encoder is a neural network component based on the Self-Attention mechanism that processes input feature sequences in parallel, efficiently capturing the implicit relationships between different features and integrating these features with high quality [27]. However, its computational cost grows squarely with the length of the input sequence.

To address this problem, Yang et al. proposed Hierarchical Attention [28] by dividing the input sequence into multiple sub-sequences, and then performing the self-attentive operation on each sub-sequence separately. This approach can reduce the computational complexity while maintaining good performance. Zhang et al. proposed Windowed Attention [29], which restricts the scope of self-attention so that each input element focuses only on other elements in its neighborhoods. This can significantly reduce the computational effort, but may lose longer distance-dependent information. Tay et al. proposed Sparse Attention [30], which reduces the computational complexity by computing only some of the relationships between input elements through a sparse matrix technique. This approach attempts to reduce the computational burden while maintaining the ability to handle long-range dependencies. Fan et al. proposed Low-Rank Attention [31], which reduces computational effort by decomposing the attention matrix into a low-rank matrix. This approach loses some of the expressive power of the model, but still has good performance for many tasks.

Influenced by the above research, this paper designs an improved attention, as shown in Figure 2. In traditional Self-Attention mechanisms, the input feature vectors are directly subjected to a linear transformation to obtain Query, Key, and Value (QKV) matrices. These matrices then go through the Self-Attention process, which includes the computation of attention scores, application of the Softmax function, and the aggregation of values weighted by the attention scores.

The proposed improvement in this paper modifies this traditional process by incorporating two convolutional layers (Conv1d) before and after the self-attention mechanism.

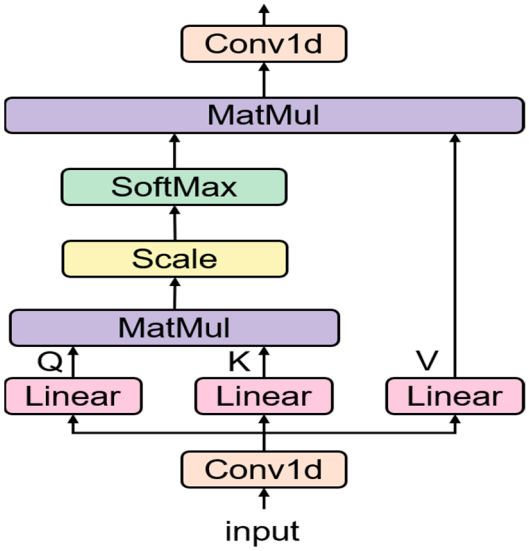

Figure 2. Improved Multi-Head Attention Mechanism.

Initially, the input feature vector is passed through the first Conv1d layer. This layer, with a kernel size of 1, is designed to reduce the dimensionality of the input representation. By compressing the input, the model is able to concentrate on the most important features, potentially improving the efficiency and effectiveness of the attention process.

Following this compression, the reduced-dimensionality vectors are then subjected to the linear transformation to obtain QKV matrices, which then proceed through the traditional Self-Attention mechanism.

After the Self-Attention process, the output representation is passed through the second Conv1d layer. This layer, also with a kernel size of 1, is designed to restore the output representation back to its original dimensionality. This step allows the model to refine the output representation without losing any essential information.

By introducing these convolutional layers, this improved attention mechanism aims to enhance the model by focusing on important features and refining the output representation, potentially improving the overall performance of the model.

Next we provide the mathematical formulation for our enhanced Multi-Head Self-Attention mechanism with convolutional layers. Let x be the input tensor of shape (*batch*, n, *dim_in*), where *batch* denotes the batch size, n represents the sequence length, and *dim_in* corresponds to the input dimension.

Given the input tensor x, the downsampling convolutional layer is applied as follows:

$$x\_down = convdown(x) \tag{1}$$

where *convdown* is a 1D convolutional layer with parameters to be learned during training.

Next, the original Multi-Head Self-Attention mechanism is applied to the down sampled tensor *dim_in*. The attention mechanism can be represented as follows:

$$Q = linear\_q(x\_down) \tag{2}$$

$$K = linear\_k(x\_down) \tag{3}$$

$$V = linear\_v(x\_down) \tag{4}$$

where Q, K, and V denote the query, key, and value matrices, respectively, and *linear_q*, *linear_k*, and *linear_v* are linear transformation layers with learnable weights.

The scaled dot-product attention can then be computed as:

$$dist = softmax(Q * K^T / sqrt(dk)) \tag{5}$$

where *dk* is the key dimension divided by the number of attention heads, and * represents the matrix multiplication operation.

The output of the attention mechanism, denoted by *att*, can be calculated as:

$$att = dist * V \tag{6}$$

Finally, the upsampling convolutional layer *convup* is applied to the attention output *att* to obtain the final output tensor *y*:

$$y = convup(att) \tag{7}$$

where *convup* is a 1D convolutional layer with parameters to be learned during training.

Our enhanced Multi-Head Self-Attention mechanism can be summarized as the composition of the down sampling convolutional layer, the original attention mechanism, and the up sampling convolutional layer:

$$y = convup(MHSA(convdown(x))) \tag{8}$$

By introducing the *convdown* and *convup* layers, our proposed method refines the input and output representations, potentially leading to improved performance in various tasks.

3.1.3. Optimization of Loss Functions

The Detection Transformer (DETR) has demonstrated outstanding results in object detection tasks. Yet, a notable shortcoming of the DETR model lies in its underperformance in detecting small objects, a problem mainly attributed to the imbalance between losses associated with large and small object. In response to this, we introduce an innovative loss function in this section. This function is designed to equalize the contributions of both large and small objects, thereby aiming to elevate the DETR model's proficiency in detecting smaller objects without undermining its overall performance.

Our proposed method focuses on modifying the loss function of the DETR model to account for the imbalance between large and small object detection. Specifically, we introduce a weighted loss that considers both the true and predicted width values of the bounding boxes. The main idea is to simultaneously calculate the loss for *w_true* and *w_pred*, as well as for $1 - w\_true$ and $1 - w\_pred$, which helps balance the loss contributions from large and small objects.

The proposed balanced loss function can be defined as follows:

$$L\_balanced = L(w\_true,\ w\_pred) + L(1 - w\_true,\ 1 - w\_pred) \tag{9}$$

where *L*() denotes the original loss function used in the DETR model, specifically the L1 loss, the ground-truth width values of the bounding boxes, and corresponds to the predicted width values of the bounding boxes.

4. Results

In this paper, experiments were conducted on a self-collected and organized orthopedic medical instrument dataset to evaluate the effectiveness of our proposed method in the task of precise orthopedic instrument recognition. This section will introduce the experimental setup and results, as well as compare and analyze them with existing methods.

4.1. Dataset

To evaluate the effectiveness of our proposed method in the task of precise orthopedic instrument recognition, we utilized a self-collected and organized orthopedic medical instrument dataset, which contains 5000 images of orthopedic instruments in various scenarios, covering 10 common types of bone plates. Each image has been manually annotated with the category and bounding box of each target object. The dataset presents a certain level of challenge, as the bone plates in the images may exhibit diversity, scale

variation, occlusion, and background interference, among other issues. Figure 3 provides some example images from the dataset.

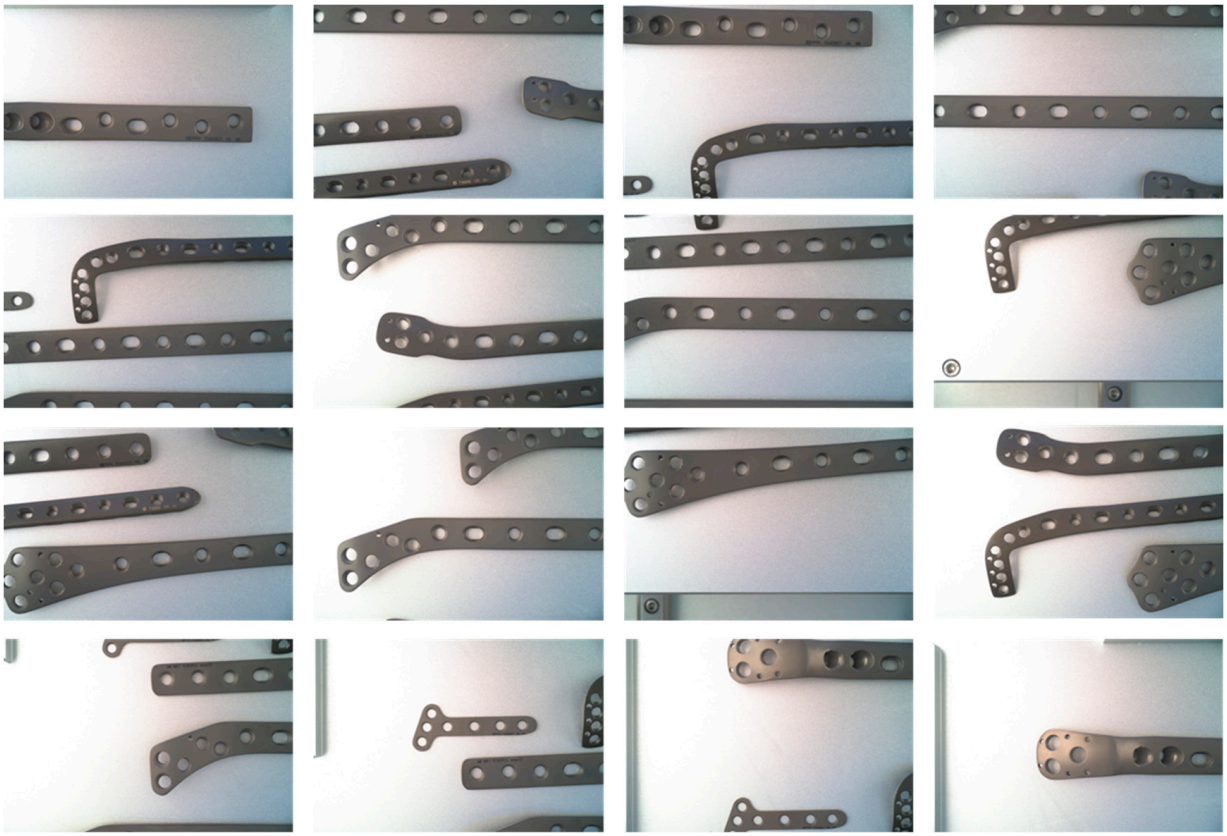

Figure 3. Dataset demonstration: Due to environmental constraints, the camera angle is relatively low, preventing a full view of the large bone plates.

4.1.1. Data Set Analysis and Annotation Instructions

Upon analyzing the orthopedic medical device dataset, we found significant differences in shape, size, and structure among different types of bone plates. As shown in Figure 4, the tails of various bone plate models exhibit high similarity, making it difficult even for humans to distinguish them during the annotation process. Therefore, the head of the bone plate is chosen as the distinctive feature for differentiating between various models.

Figure 4. Partial product map of the dataset.

The annotation situation of the dataset is shown in Figure 5, where we provided detailed annotations for each bone plate, including categories, bounding boxes, key points, and other information. To ensure the quality of the annotations, we employed a multi-person annotation and cross-validation approach, rigorously reviewing the annotation results for each image. Simultaneously, we divided the dataset into training, validation, and test sets to maintain the independence and consistency of the data during the training and evaluation processes.

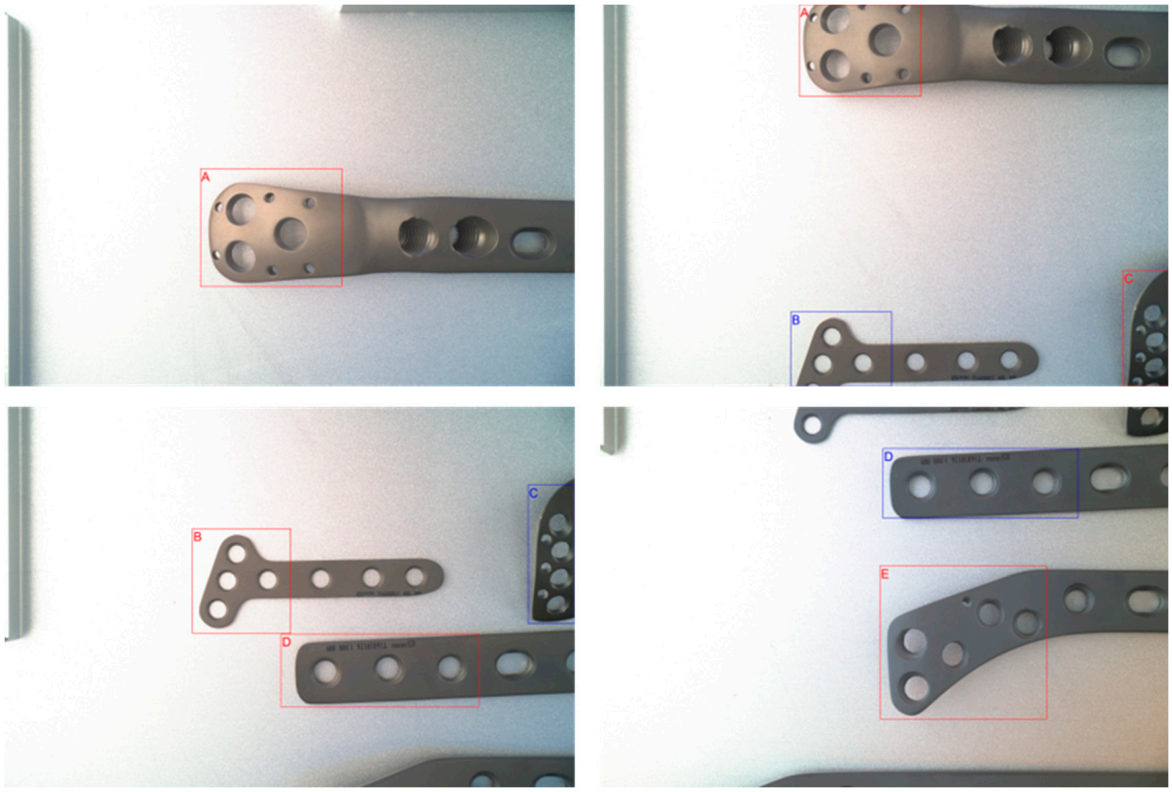

Figure 5. Annotation examples of the dataset: selecting regions with significant feature differences for annotation, rather than the entire object. In the image, letters A, B, C, D, and E correspond to different models of bone plates, with each letter indicating a specific model distinguished by its unique features.

Moreover, to better understand the challenges present in the dataset, we conducted additional statistical analyses. We found that the size, angle, and position of the bone plates exhibit significant variations, which pose additional challenges for model training. At the same time, factors such as potential occlusion, background interference, and changes in lighting conditions within the images also impose higher demands on bone plate recognition.

In summary, we have collected and organized a challenging orthopedic medical device dataset to evaluate the effectiveness of our proposed method in the precise recognition of orthopedic medical devices. Through the analysis and annotation of the dataset, we have provided strong support model training and evaluation.

4.1.2. Rational Assessment of Data Sets

To evaluate the rationality of the dataset, we conducted a statistical analysis on the number of annotated bounding boxes in each image and the frequency of each category in the sampled dataset, as shown in Figure 6. Due to the large volume of the bone plates and the low camera shooting angle, the number of objects in each image is only between 1 and 4, with the majority being 1. There is also a significant difference in categories, with a roughly three-fold difference between the maximum and minimum values. This further increases the difficulty of model training and enhances the challenge of the task. To adapt to the characteristics of the dataset, we adjusted the number of pre-detection boxes in the improved DETR to 10 and made fine adjustments to the category loss calculation based on the probability of different categories appearing in the dataset. Specifically, we employed a weighted cross-entropy loss function. This loss function takes into account the probability distribution of each category in the dataset and assigns a weight to each category. The weights can reflect the degree of class imbalance.

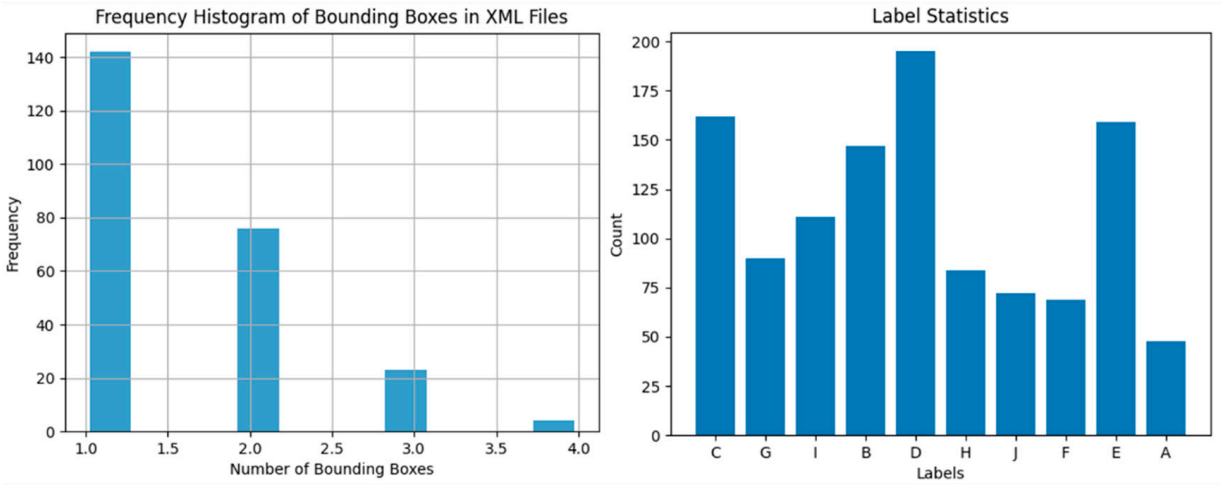

Figure 6. Statistical analysis of annotated bounding boxes and category frequency in the sampled dataset.

The weighted cross-entropy loss function is defined as follows:

$$L(y,\ p) = -\sum[w\_i * y\_i * log(p\_i)] \tag{10}$$

where $L(y,p)$ is the weighted cross-entropy loss value; y is the one-hot encoded representation of the true labels; p is the probability distribution predicted by the model; i is the index representing the current class; w_i is the weight assigned to class; y_i is the value of class in the true label y (0 or 1); and p_i is the probability of class predicted by the model.

To calculate the weight w_i, we can utilize the probability of occurrence of each class in the dataset. A simple approach is to use the inverse frequency of the classes:

$$w\_i = N\ /\ (N\_i * K) \tag{11}$$

where N is the total number of samples in the dataset, N_i is the number of samples for class in the dataset, and K is the total number of classes.

By using this approach, low-frequency classes are assigned higher weights, thereby having a greater influence in the loss function. This helps the model to pay more attention to those classes that occur less frequently, allowing for subtle adjustments during loss computation.

4.1.3. Data Enhancements

To evaluate the rationality of the dataset, we conducted a statistical analysis on the number of annotated bounding boxes in each image and the freq.

Contrast Enhancement: By adjusting the contrast of the images, we emphasized the features of the target objects, thereby increasing the model's sensitivity to object detection. We appropriately enhanced the contrast of the training images to better capture the details of the target objects.

Noise Addition: We added random noise to the training images to simulate the noise interference that may occur in real-world scenarios. This helps improve the model's robustness and generalization ability in noisy environments.

Brightness Adjustment: We adjusted the brightness of the training images to simulate different lighting conditions. This helps improve the model's performance under varying lighting conditions.

Flipping: We horizontally and vertically flipped the training images to increase the number of samples for target objects from different perspectives. This aids in enhancing the model's ability to recognize target objects from different viewpoints.

Rotation: We randomly rotated the training images at various angles to increase the number of samples for target objects from different orientations. This helps improve the model's ability to recognize target objects from different directions.

Through the five data augmentation methods mentioned above, we effectively enhanced the diversity of the training set, which contributes to improving the generalization capability and performance of the object detection model. In the experimental section, we will demonstrate the specific impact of these data augmentation strategies on model performance. The effects of the five data augmentation methods are shown in Figure 7.

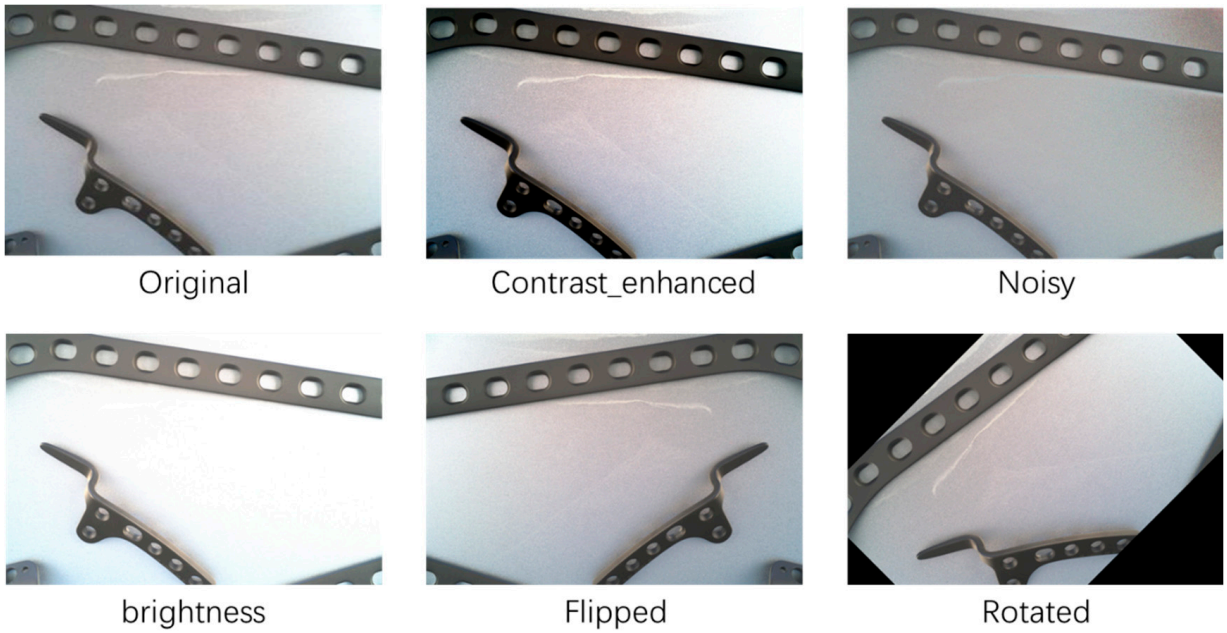

Figure 7. Comparison of data enhanced visualization.

4.2. Experimental Setup

Based on the dataset described earlier, we split it into training, validation, and testing sets in an 8:1:1 ratio, with 4000, 500, and 500 images, respectively.

We implemented our proposed method using the PyTorch framework and conducted experiments on a dataset of orthopedic medical devices that we collected and organized ourselves. We used a pre-trained MLP-Mixer as the backbone network and fused its mix layer with the transformer encoder layer in a multi-scale feature fusion module. We employed a transformer encoder–decoder structure to process image features and object queries, and made some adjustments to its architecture. We initialized it with 10 object queries and a method of mixed query selection. By using Dense Prior Initialization, which is a method of initializing target containers with dense priors, we were able to achieve similar performance to existing models with only one decoder layer.

We used the AdamW optimizer with a learning rate of 1×10^{-4} weight decay of 1×10^{-4}, and batch size of 16. We trained the model for 50 epochs on an NVIDIA Tesla V100 GPU and evaluated it at the end of each epoch.

4.3. Experimental Results and Analysis

To compare the performance of our proposed enhanced DETR model with existing convolutional neural network-based methods, we conducted evaluations using several key metrics, including mean average precision (mAP) and inference speed. In this section, we will present these metrics through charts and provide a detailed analysis of the results.

4.3.1. Horzontal Comparative Experiment

We conducted a comparative experiment using our proposed OrthoDETR model as the primary approach. In order to evaluate its performance, we compared it against several state-of-the-art convolutional neural network-based methods, including the following:

Faster R-CNN: This is a classical two-stage object detection method that utilizes a region proposal network to generate candidate regions, followed by a region classification network to predict the class labels and bounding boxes.

YOLOv8: An advanced one-stage object detection technique that partitions the input image into a grid system and forecasts numerous anchor boxes along with associated class probabilities within each grid cell.

SSD: A simple yet effective one-stage object detection method that employs multiple feature maps of different scales to predict anchor boxes and class probabilities for objects of varying sizes.

RetinaNet: An improved one-stage object detection method that introduces a focal loss-based classification branch, which effectively handles the challenge of class imbalance.

Table 1 below presents a comprehensive comparison of the experimental results obtained from these models, providing a clear visual representation of the effectiveness of OrthoDETR in relation to these established methods.

Table 1. Comparison of Experimental Results of Different Models.

Model	AP50	AP50:95	AR50:95	FPS	Parmeters (Millions)	FLOPs (Billion)
DETR	0.852	0.842	0.862	20	41.5	244
Faster R-CNN	0.865	0.815	0.845	24	134.0	150
YOLOv8	0.886	0.852	0.893	33	64.9	139
SSD	0.835	0.793	0.820	28	26.3	31
RetinaNet	0.861	0.812	0.847	22	36.8	138
OrthoDETR (Ours)	0.897	0.864	0.895	26	39.7	123

Due to the relatively large targets in the dataset, we employed AP50, AP50:95, and AR50:95 as performance evaluation metrics for the model, which measure precision and recall at different IoU thresholds. We also compared FPS to assess the real-time performance of different models. It can be observed that OrthoDETR outperforms other models in AP50, AP50:95, and AR50:95. At the same time, our method demonstrates good performance in terms of inference speed, particularly when compared to two-stage methods like Faster R-CNN. Although it is slightly slower than single-stage methods such as YOLOv8 and SSD in inference speed, the improved DETR's advantages in accuracy and robustness compensate for this shortcoming.

In conclusion, OrthoDETR's performance metrics (AP50, AP50:95, and AR50:95) are superior to the other models, indicating its high precision and recall rates. In terms of speed, OrthoDETR performs competitively, providing a good trade-off between speed and accuracy. Importantly, OrthoDETR exhibits a lower number of parameters and FLOPs compared to some models like Faster R-CNN, indicating its lower complexity and higher efficiency. This underlines our model's suitability for real-time applications in the industrial and medical sectors, where both accuracy and efficiency are crucial.

4.3.2. Improved Strategy Ablation Experiment

To verify that the proposed improvements have a positive impact on the results, we conducted ablation experiments using a controlled variable approach. Ablation experiments involve systematically removing model components to evaluate the contribution of each component to the overall performance. In this study, we independently assessed the impact of replacing the ResNet backbone with MLP-MIXER, optimizing the Multi-Head Self-Attention mechanism, and adjusting the loss function. The results are shown in Table 2.

Table 2. Results of ablation experiments on OrthoDETR model components.

MLP-Mixer Backbone	Improved Transformer	Optimized Loss Function	MAP	FPS
			0.718	20
√			0.703	24
	√		0.750	23
		√	0.739	19
√	√	√	0.756	26

The ablation study results show that after replacing the ResNet backbone with MLP-MIXER, OrthoDETR's detection speed increased by approximately 20%, while the mean Average Precision (mAP) only decreased by 1.5%. This suggests that MLP-MIXER has a positive impact on improving detection speed. On the other hand, optimizing the Multi-Head Self-Attention mechanism improved OrthoDETR's performance in handling occlusion and background interference, leading to an mAP increase of around 3.2%. Lastly, by adjusting the loss function, OrthoDETR demonstrated a stronger robustness in addressing scale variations and diversity issues, resulting in a further 2.1% increase in mAP.

4.3.3. Data Enhanced Ablation Experiments

To verify the effectiveness of data augmentation in enhancing the model's generalization capabilities and performance, we compared the impact of various data augmentation strategies on the performance of object detection models. The results are shown in Table 3.

Table 3. The impact of different data augmentation strategies on model performance. The baseline model refers to the original model without data augmentation, while the models for various augmentation strategies are based on the baseline model, with each respective data augmentation strategy applied individually.

Model	Average Precision
Baseline Model	80.0%
Contrast Enhancement	82.5%
Noise Addition	81.3%
Brightness Adjustment	83.2%
Flipping	82.4%
Rotation	82.1%

Based on the experimental results, we observed that applying data augmentation strategies on top of the baseline model led to improved model performance. Among these strategies, the flipping approach contributed most significantly to the performance enhancement, increasing accuracy by 2.4 percentage points. Contrast enhancement, brightness adjustment, and rotation strategies also had a positive impact on model performance, raising it by 2.5, 3.2, and 2.1 percentage points, respectively. The noise addition strategy had a relatively minor effect on performance improvement, with an increase of only 1 percentage point.

Experimental results indicate that employing data augmentation strategies can effectively enhance the generalization capabilities and performance of object detection models. In practical applications, appropriate data augmentation strategies can be selected according to specific scenarios and requirements.

4.3.4. Example Images and Analysis of Test Results

Figure 8 presents a collection of images containing orthopedic medical devices along with their detection results generated by different methods. The improved DETR model's strong performance in handling issues such as occlusion and background interference is illustrated through these examples. This can be attributed to the series of optimization measures introduced in the model, which allow the enhanced DETR model to adapt to a

wide range of challenging visual environments. As a result, the improved DETR model demonstrates superior performance in practical applications.

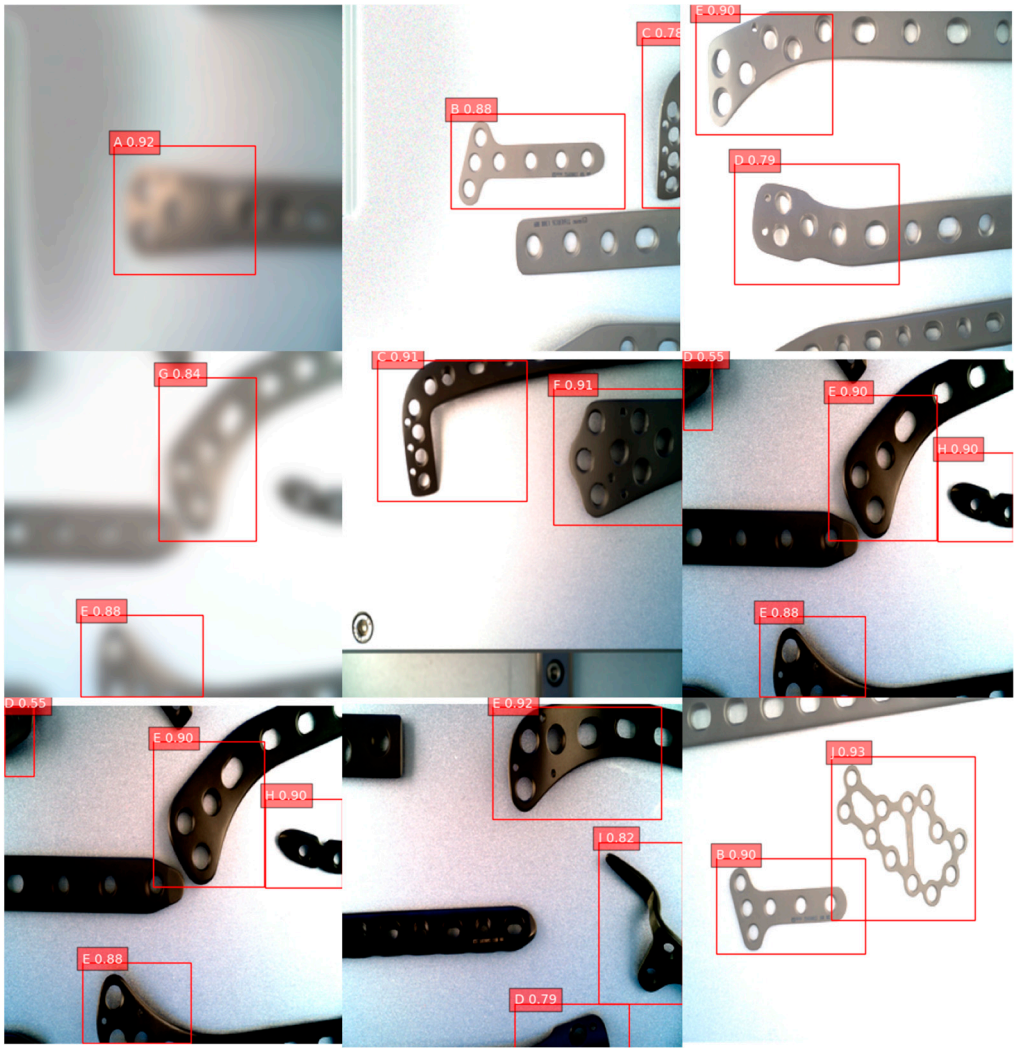

Figure 8. Sample images and detection results. In the image, letters A–J correspond to different models of bone plates.

In summary, our method excels in the precise recognition task of orthopedic medical devices. Compared to techniques such as Faster R-CNN, YOLOv8, SSD, and RetinaNet, our improved DETR model achieves enhanced Average Precision (AP) and Average Recall (AR), demonstrating its superiority in terms of recognition accuracy. Simultaneously, our approach also exhibits commendable performance in inference speed, particularly when compared to two-stage methods like Faster R-CNN. Although slightly inferior in inference speed to single-stage methods like YOLOv8 and SSD, the improved DETR compensates for this shortcoming with its advantages in accuracy and robustness.

4.4. Complexity and Cost Analysis of OrthoDETR

In this section, we provide an in-depth analysis of the complexity and cost associated with implementing our proposed OrthoDETR method. Understanding these factors is essential to evaluate the practicality and efficiency of our method for real-world applications.

4.4.1. Computational Complexity

OrthoDETR's core modifications, including the substitution of the ResNet backbone with an MLP-Mixer and refinement of the Multi-Head Self-Attention mechanism, impact its

computational complexity. By utilizing the MLP-Mixer, we reduce the number of parameters involved in the architecture while maintaining comparable performance. Additionally, the improved Multi-Head Self-Attention mechanism allows for better context comprehension, increasing efficiency by focusing on relevant local and global features. Consequently, OrthoDETR displays a reduced computational complexity compared to the original DETR model, which leads to faster detection times.

4.4.2. Memory Usage and Training Cost

By integrating the MLP-Mixer, OrthoDETR benefits from decreased memory usage, which results in lower training and inference costs. This streamlined architecture allows for efficient utilization of hardware resources and makes OrthoDETR more feasible for large-scale implementation. Moreover, the customized loss function contributes to better optimization during training, minimizing the required training epochs and overall associated costs.

In summary, OrthoDETR exhibits a favorable balance between complexity, cost, and performance, making it a valuable tool for detecting orthopedic medical devices, especially in fine-grained management during workflow processes.

Furthermore, our advanced DETR model demonstrates remarkable resilience when handling images that present a diverse array of challenges, such as variations in scale, occlusion, and background interference. This resilience stems from the implementation of multi-scale feature fusion modules, the employment of high-resolution inputs, and adjustments to the object query quantities within the model. These optimization tactics allow our refined DETR model to adapt effectively to a vast array of demanding visual scenarios, thereby elevating its performance in real-world applications. Consequently, these benefits equip our improved DETR model with the ability to address numerous challenges in practical contexts, ultimately enhancing the accuracy and practicality of orthopedic medical device recognition.

In conclusion, our proposed method, centered on an enhanced DETR, achieves exceptional performance in the precise recognition of orthopedic medical devices, as evidenced by our experimental results. These findings not only validate the effectiveness and superiority of Transformers in object detection tasks but also introduce a pioneering, industry-relevant solution for the identification of orthopedic medical devices. Moreover, this research has the potential to provide valuable insights for the development of future object detection tasks across various domains.

5. Conclusions

In our research, we have specifically focused on the unique attributes and needs of orthopedic medical devices, introducing various improvements and optimizations for the DETR model. Significant changes included replacing the underlying network structure from ResNet to MLP-Mixer, refining the Multi-Head Self-Attention mechanism to fortify model performance, and modifying the loss function. Using a meticulously curated dataset of orthopedic devices, our experiments demonstrated that these adjustments not only maintain recognition accuracy, but also improve inference speed by 23%. This enhanced performance makes OrthoDETR advantageous for applications requiring high real-time processing capabilities.

For instance, in the industrial sector, OrthoDETR can swiftly and accurately identify orthopedic devices on the production line, enhancing quality control and reducing errors or defects. In the healthcare sector, our model can be employed in real-time diagnostic tools, providing accurate and efficient support to healthcare professionals, thereby improving patient outcomes and safety. The successful application of OrthoDETR heralds a new research direction in medical image analysis, unveiling the immense potential of Transformer-based object detection strategies in medical and industrial sectors. As deep learning technologies continue to advance and further exploration in the field of medical imaging continues,

we believe OrthoDETR and related approaches will contribute significantly to improving healthcare quality, patient safety, and industrial efficiency in the future.

Author Contributions: Conceptualization, X.Z. (Xiaobo Zhang) and H.L.; methodology, H.L.; data curation, X.Z. (Xiaobo Zhang); writing—original draft preparation, J.L.; writing—review and editing, J.L.; supervision, J.L.; funding acquisition, X.Z. (Xuehai Zhou). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC)—Research on the mechanism of information-physical interface and safe interaction method in mines (51874010).

Data Availability Statement: Due to the sensitive nature of the data used in this study, and in order to protect the privacy and confidentiality of involved individuals or entities, it is not feasible to publicly archive and share the dataset. This decision is in alignment with prevailing ethical guidelines and privacy laws. We assure that the data was rigorously analyzed and all findings are accurately represented in this study.

Acknowledgments: We would like to extend our heartfelt thanks to Xiaobo Zhang for providing the crucial data that greatly aided this research. We are also immensely grateful to Jingzhao Li for their invaluable guidance and constructive suggestions during the course of writing this paper. Their insights have been instrumental in shaping this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chua, C.Y.X.; Liu, H.-C.; Di Trani, N.; Susnjar, A.; Ho, J.; Scorrano, G.; Rhudy, J.; Sizovs, A.; Lolli, G.; Hernandez, N.; et al. Carbon fiber reinforced polymers for implantable medical devices. *Biomaterials* **2021**, *271*, 120719. [CrossRef] [PubMed]
2. Huzum, B.; Puha, B.; Necoara, R.M.; Gheorghevici, S.; Puha, G.; Filip, A.; Sirbu, P.D.; Alexa, O. Biocompatibility assessment of biomaterials used in orthopedic devices: An overview (Review). *Exp. Ther. Med.* **2021**, *22*, 1315. [CrossRef] [PubMed]
3. Wang, L.; Ding, X.; Feng, W.; Gao, Y.; Zhao, S.; Fan, Y. Biomechanical study on implantable and interventional medical devices. *Acta Mech. Sin.* **2021**, *37*, 875–894. [CrossRef]
4. Wang, Y.; Xu, K.; Wang, Y.; Ye, W.; Hao, X.; Wang, S.; Li, K.; Du, J. Investigation and analysis of four countries' recalls of osteosynthesis implants and joint replacement implants from 2011 to 2021. *J. Orthop. Surg. Res.* **2022**, *17*, 443. [CrossRef] [PubMed]
5. Sambolek, S.; Ivasic-Kos, M. Automatic person detection in search and rescue operations using deep CNN detectors. *IEEE Access* **2021**, *9*, 37905–37922. [CrossRef]
6. Maity, M.; Banerjee, S.; Chaudhuri, S.S. Faster r-cnn and yolo based vehicle detection: A survey. In Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 8–10 April 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1442–1447.
7. Chu, Y.; Yang, X.; Li, H.; Ai, D.; Ding, Y.; Fan, J.; Song, H.; Yang, J. Multi-level feature aggregation network for instrument identification of endoscopic images. *Phys. Med. Biol.* **2020**, *65*, 165004. [CrossRef] [PubMed]
8. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
10. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24261–24272.
11. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
12. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A review of yolo algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
13. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Song, H.; Sun, D.; Chun, S.; Jampani, V.; Han, D.; Heo, B.; Kim, W.; Yang, M.H. An extendable, efficient and effective transformer-based object detector. *arXiv* **2022**, arXiv:2204.07962.
16. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic detr: End-to-end object detection with dynamic attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2988–2997.

17. Ickler, M.K.; Baumgartner, M.; Roy, S.; Wald, T.; Maier-Hein, K.H. Taming Detection Transformers for Medical Object Detection. In *BVM Workshop*; Springer Fachmedien Wiesbaden: Wiesbaden, Germany, 2023; pp. 183–188.
18. Mathesul, S.; Swain, D.; Satapathy, S.K.; Rambhad, A.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A. COVID-19 Detection from Chest X-ray Images Based on Deep Learning Techniques. *Algorithms* **2023**, *16*, 494. [CrossRef]
19. Sakaida, M.; Yoshimura, T.; Tang, M.; Ichikawa, S.; Sugimori, H. Development of a Mammography Calcification Detection Algorithm Using Deep Learning with Resolution-Preserved Image Patch Division. *Algorithms* **2023**, *16*, 483. [CrossRef]
20. Carballo, J.A.; Bonilla, J.; Fernández-Reche, J.; Nouri, B.; Avila-Marin, A.; Fabel, Y.; Alarcón-Padilla, D.C. Cloud Detection and Tracking Based on Object Detection with Convolutional Neural Networks. *Algorithms* **2023**, *16*, 487. [CrossRef]
21. Sami, A.A.; Sakib, S.; Deb, K.; Sarker, I.H. Improved YOLOv5-Based Real-Time Road Pavement Damage Detection in Road Infrastructure Management. *Algorithms* **2023**, *16*, 452. [CrossRef]
22. Du, G.; Cao, X.; Liang, J.; Chen, X.; Zhan, Y. Medical image segmentation based on u-net: A Review. *J. Imaging Sci. Technol.* **2020**, *64*, 020508. [CrossRef]
23. Ji, Y.; Zhang, R.; Li, Z.; Ren, J.; Zhang, S.; Luo, P. Uxnet: Searching multi-level feature aggregation for 3d medical image segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020; Proceedings, Part I 23. Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 346–356.
24. Grignaffini, F.; Troiano, M.; Barbuto, F.; Simeoni, P.; Mangini, F.; D'andrea, G.; Piazzo, L.; Cantisani, C.; Musolff, N.; Ricciuti, C.; et al. Anomaly Detection for Skin Lesion Images Using Convolutional Neural Network and Injection of Handcrafted Features: A Method That Bypasses the Preprocessing of Dermoscopic Images. *Algorithms* **2023**, *16*, 466. [CrossRef]
25. Wang, H.; Qiu, S.; Ye, H.; Liao, X. A Plant Disease Classification Algorithm Based on Attention MobileNet V2. *Algorithms* **2023**, *16*, 442. [CrossRef]
26. Apostolopoulos, D.J.; Apostolopoulos, I.D.; Papathanasiou, N.D.; Spyridonidis, T.; Panayiotakis, G.S. Explainable Artificial Intelligence Method (ParaNet+) Localises Abnormal Parathyroid Glands in Scintigraphic Scans of Patients with Primary Hyperparathyroidism. *Algorithms* **2023**, *16*, 435. [CrossRef]
27. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
28. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
29. Zhang, S.; Loweimi, E.; Bell, P.; Renals, S. Windowed attention mechanisms for speech recognition. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 7100–7104.
30. Tay, Y.; Bahri, D.; Yang, L.; Metzler, D.; Juan, D.C. Sparse sinkhorn attention. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; PMLR. pp. 9438–9447.
31. Fan, X.; Liu, Z.; Lian, J.; Zhao, W.X.; Xie, X.; Wen, J.R. Lighter and better: Low-rank decomposed self-attention networks for next-item recommendation. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 11–15 July 2021; pp. 1733–1737.