

Article

Advanced Object Detection for Maritime Fire Safety

Fazliddin Makhmudov ¹, Sabina Umirzakova ^{1,*}, Alpamis Kutlimuratov ², Akmalbek Abdusalomov ^{1,2} and Young-Im Cho ^{1,*}

¹ Department of Computer Engineering, Gachon University, Seongnam 1342, Republic of Korea; fazliddin12@gachon.ac.kr (F.M.); akmaljon@gachon.ac.kr (A.A.)

² Department of Econometrics, Tashkent State University of Economics, Tashkent 100066, Uzbekistan; a.kutlimuratov@tsue.uz

* Correspondence: sabinatuit@gachon.ac.kr (S.U.); yicho@gachon.ac.kr (Y.-I.C.)

Abstract: In this study, we propose an advanced object detection model for fire and smoke detection in maritime environments, leveraging the DETR (Detection with Transformers) framework. To address the specific challenges of shipboard fire and smoke detection, such as varying lighting conditions, occlusions, and the complex structure of ships, we enhance the baseline DETR model by integrating EfficientNet-B0 as the backbone. This modification aims to improve detection accuracy while maintaining computational efficiency. We utilize a custom dataset of fire and smoke images captured from diverse shipboard environments, incorporating a range of data augmentation techniques to increase model robustness. The proposed model is evaluated against the baseline DETR and YOLOv5 variants, showing significant improvements in Average Precision (AP), especially in detecting small and medium-sized objects. Our model achieves a superior AP score of 38.7 and outperforms alternative models across multiple IoU thresholds (AP50, AP75), particularly in scenarios requiring high precision for small and occluded objects. The experimental results highlight the model's efficacy in early fire and smoke detection, demonstrating its potential for deployment in real-time maritime safety monitoring systems. These findings provide a foundation for future research aimed at enhancing object detection in challenging maritime environments.

Keywords: shipboard fire detection; maritime safety; DETR model; fire and smoke detection; real-time detection



Citation: Makhmudov, F.; Umirzakova, S.; Kutlimuratov, A.; Abdusalomov, A.; Cho, Y.-I. Advanced Object Detection for Maritime Fire Safety. *Fire* **2024**, *7*, 430. <https://doi.org/10.3390/fire7120430>

Academic Editor: Grant Williamson

Received: 13 October 2024

Revised: 20 November 2024

Accepted: 22 November 2024

Published: 25 November 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Fire and smoke detection is a critical component of safety systems in various industrial and transportation sectors, particularly in maritime environments where early detection and response are essential to prevent catastrophic outcomes [1]. Ships, due to their enclosed spaces, complex structures, and exposure to harsh environmental conditions, present unique challenges for fire and smoke detection [2]. Traditional detection systems, such as thermal cameras and smoke sensors, have limitations in identifying fire or smoke at its earliest stages, especially in scenarios involving occlusions, varying lighting conditions, and the dynamic movement of ships [3]. To address these challenges, advancements in computer vision and deep learning offer promising alternatives through automated object detection models.

Object detection models based on Convolutional Neural Networks (CNNs) have demonstrated remarkable success in a wide range of applications [4]. However, these models often require extensive tuning and complex post-processing steps, such as non-maximum suppression (NMS), to achieve high accuracy. The introduction of the DETR by Facebook AI [5]. Research marked a significant leap in object detection by eliminating the need for region proposals and NMS. DETR leverages a transformer-based architecture, which excels at modeling long-range dependencies within images, making it highly suitable for detecting objects under challenging conditions, such as those found in maritime

environments. Despite its advantages, the DETR model reliance on ResNet as a backbone for feature extraction presents limitations in balancing detection performance with computational efficiency, especially in resource-constrained environments.

In this paper, we propose a novel enhancement to the DETR model by integrating EfficientNet-B0 [6] as the backbone architecture. EfficientNet-B0 is known for its ability to achieve high accuracy with low computational cost, making it a suitable choice for real-time detection applications on ships. Our modified DETR model aims to optimize fire and smoke detection by improving feature extraction while maintaining computational efficiency. We also introduce a custom dataset specifically designed for fire and smoke detection on ships, which captures the diverse and challenging conditions encountered in maritime environments. The dataset is augmented with various preprocessing techniques to ensure robustness across different scenarios, such as varying lighting conditions, occlusions, and ship movement. We compare the performance of our proposed model against state-of-the-art models, including the baseline DETR and YOLOv5 variants, evaluating metrics such as Average Precision (AP) across different object sizes and Intersection over Union (IoU) thresholds.

This paper presents several key contributions that advance the state-of-the-art in fire and smoke detection, particularly within the challenging context of maritime environments:

- We propose an enhanced object detection model based on the DETR framework, integrating EfficientNet-B0 as the backbone. The EfficientNet-B0 architecture is known for its balance between computational efficiency and accuracy, making it particularly suitable for real-time detection tasks. By replacing the traditional ResNet backbone used in DETR with EfficientNet-B0, our model achieves improved detection accuracy while maintaining efficiency, especially in detecting small and occluded objects such as distant smoke plumes or flames.
- We introduce a custom, annotated dataset designed for fire and smoke detection in shipboard environments. This dataset includes a diverse range of ship settings, both interior and exterior, and accounts for varying lighting conditions, viewpoints, and environmental factors. It is a valuable resource for training and evaluating models for fire and smoke detection in maritime contexts, where such specialized data have been scarce.
- The proposed model demonstrates superior performance in detecting small and medium-sized objects compared to both the baseline DETR and popular YOLO variants. This enhancement is critical for maritime fire detection, where small flames and smoke plumes are often the earliest signs of fire incidents. Our model's ability to detect these challenging objects with high precision is a notable improvement over existing models.
- We perform a comprehensive experimental evaluation, comparing our proposed model to state-of-the-art object detection frameworks, including YOLOv5 and the baseline DETR model. Our results show that the proposed model outperforms these alternatives across multiple metrics, including Average Precision (AP) at different Intersection over Union (IoU) thresholds. This thorough evaluation highlights the effectiveness of our model in various detection tasks, demonstrating its robustness and generalization capabilities across different maritime scenarios.
- The integration of EfficientNet-B0 allows the proposed model to strike a balance between computational load and detection performance, making it suitable for deployment in real-time fire and smoke detection systems aboard ships. This real-time capability is critical for early hazard detection and safety monitoring, where timely response is crucial to preventing catastrophic events.

Through these contributions, this paper provides a significant advancement in the field of fire and smoke detection in maritime environments, offering a robust, efficient, and accurate solution that addresses the unique challenges of shipboard safety monitoring. These contributions lay the groundwork for future research on and the practical

implementation of deep learning-based detection systems in maritime safety and other complex environments.

The contributions of this work are threefold: (1) the introduction of a novel modification to the DETR model by incorporating EfficientNet-B0 for enhanced fire and smoke detection in maritime environments, (2) the development of a custom dataset tailored for shipboard fire and smoke detection, and (3) an extensive experimental evaluation demonstrating the superior performance of our proposed model in comparison to existing object detection models. These advancements provide a foundation for future research into real-time hazard detection and response systems in maritime settings.

2. Related Works

The field of object detection has undergone rapid advancements in recent years, driven by the increasing power of deep learning models and the introduction of transformer-based architectures. Traditional approaches, such as the Region-based Convolutional Neural Networks (R-CNNs) and their variants like Fast R-CNN and Faster R-CNN, have been widely adopted in various object detection tasks [7]. These methods rely on region proposal mechanisms to identify potential objects in an image, followed by classification and bounding box regression. However, these models often require complex post-processing steps such as non-maximum suppression (NMS) to eliminate duplicate detections, and their performance is constrained by the efficiency of the underlying CNN-based feature extractors.

More recently, single-shot detectors like YOLO [2] and Single-Shot MultiBox Detector (SSD) [8] have gained prominence due to their real-time detection capabilities. YOLO, in particular, has been successfully applied to various detection tasks and has evolved into multiple versions, with improvements in both speed and accuracy. Despite their efficiency, these models still struggle with detecting small objects and maintaining accuracy in complex environments, which are critical in fire and smoke detection, particularly in dynamic and confined maritime environments. The introduction of transformers into object detection, most notably with DETR [9], marked a significant shift from traditional CNN-based approaches. DETR, developed by Facebook AI Research, eliminates the need for region proposals and NMS by utilizing a transformer-based encoder–decoder architecture that directly predicts object classes and bounding boxes. This architecture allows DETR to capture long-range dependencies and global contextual relationships within the image, making it particularly effective in scenarios where objects are widely spaced or occluded [10]. However, DETR reliance on ResNet as the backbone for feature extraction poses challenges in achieving a balance between computational efficiency and detection performance, especially for real-time applications in constrained environments like ships [11].

In fire and smoke detection, early systems primarily relied on sensor-based methods, such as smoke detectors and thermal imaging cameras [12]. While effective in many scenarios, these systems are limited by their inability to provide early detection in complex and visually challenging environments, such as those found on ships. As deep learning models gained traction, researchers began exploring CNN-based methods for fire and smoke detection. Ref. [13] proposed a CNN-based model for fire detection in industrial environments, which demonstrated significant improvements over traditional methods. However, the application of these models to maritime environments remains underexplored, particularly in terms of integrating advanced architectures like transformers and EfficientNet. EfficientNet, introduced by [6], is a family of models designed through a compound scaling method that uniformly scales depth, width, and resolution for better performance with fewer parameters. EfficientNet has been widely adopted in image classification tasks due to its high accuracy and low computational cost [14]. Recent works have demonstrated the potential of EfficientNet when integrated into object detection models. EfficientDet, a detector based on EfficientNet, showed superior performance in both accuracy and speed compared to existing models like YOLOv5 and SSD [15]. These findings suggest that EfficientNet could be beneficial when integrated into more complex detection models like DETR to enhance performance in specific applications such as fire and smoke detection. Recent

approaches have leveraged both classic and modern object detection architectures, with significant contributions from models such as Dynamic Kernel Temporal Prediction Module (DK-TPM) [16] and Deep Dynamic Attention Fusion (DDAF) [17]. DK-TPM introduces a temporal module that adapts dynamically to flame characteristics over time, making it suitable for video-based flame detection where temporal consistency is critical. This approach has demonstrated high reliability in scenarios where flames change shape and size due to environmental factors, outperforming static models in video datasets. Moreover, Detectron2 [18], a comprehensive and widely used framework for object detection, has proven effective in flame and smoke detection tasks due to its modular and extensible nature. Detectron2's support for various backbone architectures, such as ResNet and FPN, combined with its compatibility with advanced segmentation and mask prediction capabilities, makes it a flexible solution for fire detection [19]. In particular, Detectron2's performance in multi-object detection and its robustness to occlusions and overlapping flames make it an attractive option for safety-critical applications [20].

In the maritime domain, fire detection poses unique challenges due to the dynamic nature of ship environments, varying lighting conditions, and occlusions caused by ship structures [21]. While there have been several studies focusing on fire detection in industrial or open environments—such as [22], in which a real-time fire detection system using a lightweight CNN was developed—few works have addressed the specific challenges of fire and smoke detection on ships. The development of custom datasets for such applications is also limited, with most models relying on general-purpose datasets that do not capture the specific conditions encountered on ships [1]. This gap highlights the need for specialized detection models and datasets tailored to maritime environments. In this work, we build on the strengths of transformer-based detection and EfficientNet to address the specific challenges of fire and smoke detection on ships. Our proposed model integrates EfficientNet-B0 as the backbone for DETR, aiming to improve both accuracy and computational efficiency in maritime fire and smoke detection tasks. We also introduce a custom dataset tailored to shipboard fire and smoke incidents, further advancing the state of the art in this domain. Our contributions are positioned at the intersection of deep learning, transformer-based detection, and maritime safety, offering a novel approach to early fire and smoke detection in challenging environments.

3. Methodology

In this paper, we present our proposed model for fire and smoke detection on ships, utilizing a custom dataset and the DETR model as a baseline. DETR, one of the most recent and advanced detection models, has demonstrated exceptional performance across various tasks. We enhanced the baseline by modifying the backbone of the DETR, integrating a customized EfficientNet-B0 architecture to improve detection accuracy. Sections 3.1 and 3.2 provide a detailed description of the baseline model and EfficientNet-B0, while Section 3.3 elaborates on the structure and implementation of the proposed model.

3.1. DETR

DETR is an advanced deep learning model developed for object detection, introduced by Facebook AI Research. It distinguishes itself from traditional detection methods, which often depend on region proposals and anchor boxes, by employing a transformer-based architecture [23]. The architecture, commonly utilized in natural language processing, allows DETR to perform end-to-end object detection with greater efficiency and simplicity. The architecture of DETR consists of an encoder–decoder framework, where the encoder processes the input images, and the decoder generates predictions for object classes and their corresponding bounding boxes (Figure 1).

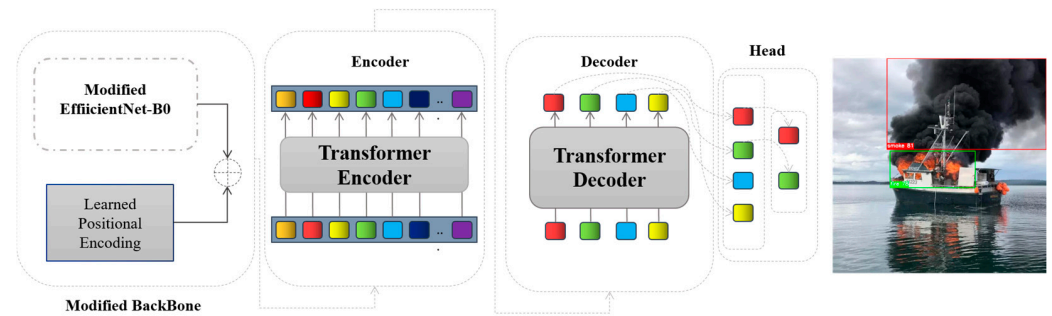


Figure 1. The architecture of the modified DETR.

The transformer architecture is the core innovation of DETR, comprising an encoder and decoder designed to analyze and interpret the extracted visual features. The encoder applies multiple layers of self-attention mechanisms and feed-forward neural networks to capture global contextual relationships across the image. The self-attention mechanism allows the model to calculate interactions between different regions of the image, enabling DETR to model long-range dependencies effectively. This capability proves especially advantageous for object detection tasks involving widely spaced or occluded objects. A key innovation in DETR is its use of bipartite matching loss, which ensures that the model outputs a unique set of object predictions. This loss function enforces a one-to-one correspondence between predicted and ground-truth objects, optimizing both object classification and bounding box localization [24]. By removing the need for non-maximum suppression (NMS) and other complex post-processing steps, DETR simplifies the object detection pipeline, contributing to its elegance and ease of implementation; while DETR achieves a performance that competes with SOTA object detection models, it comes with certain trade-offs, such as longer training times and a higher demand for large datasets. This is due to the inherent complexity of the transformer architecture.

3.2. EfficientNet-B0

EfficientNet-B0 is the foundational model in the EfficientNet family, which was designed to achieve a balance between computational efficiency and performance across image classification tasks [14]. Developed through a method called compound scaling, EfficientNet-B0 optimizes the trade-offs between network depth, width, and resolution, enabling it to outperform many previous models while maintaining a relatively low computational cost. The architecture of EfficientNet-B0 is based on the Mobile Inverted Bottleneck Convolution (MBConv) layers, first introduced in the MobileNetV2 architecture. These layers are central to the design of the model, allowing it to preserve both memory efficiency and performance.

Each MBConv block consists of a depthwise separable convolution, pointwise convolution, and a skip connection, which allows information to flow more easily through the network, preventing the loss of critical features in deeper layers (Figure 2). EfficientNet-B0 begins with a standard 3×3 convolution followed by a series of MBConv blocks with increasing expansion ratios. These blocks progressively widen and deepen the network, enabling it to extract increasingly abstract and high-level features from the input image. One of the key innovations of EfficientNet-B0 is its compound scaling strategy, which scales all three dimensions of the network—depth, width, and resolution—in a coordinated manner [25]. Unlike previous models that scale one dimension at a time, EfficientNet-B0 scales all three dimensions simultaneously, leading to a more balanced and efficient architecture. The compound scaling formula allows EfficientNet-B0 to achieve superior accuracy on standard benchmarks with fewer parameters and FLOPs compared to other networks of similar size.

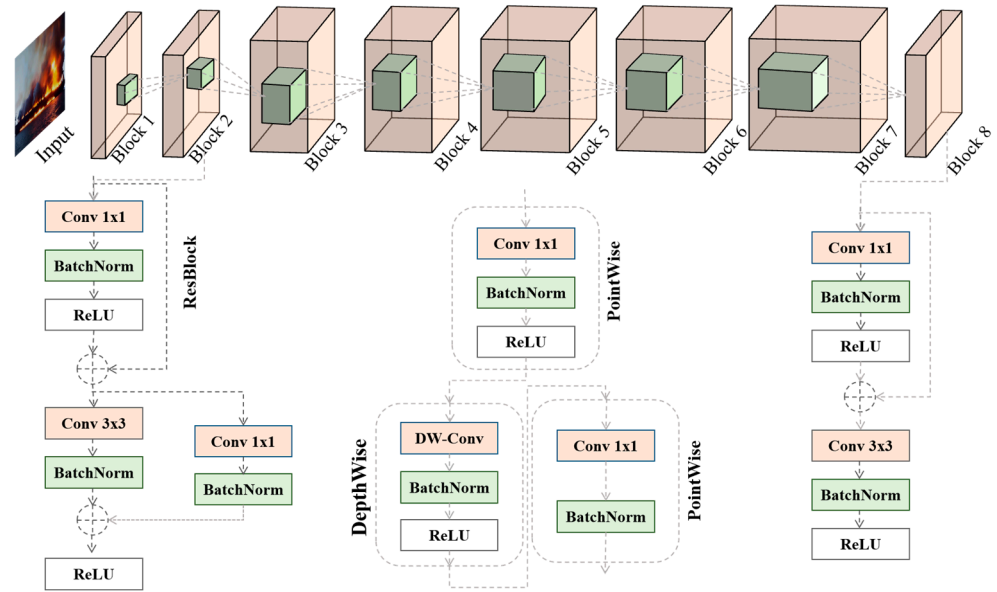


Figure 2. The modified EfficientNet-B0.

3.3. The Proposed Model

In our proposed model, we have modified one of the leading contemporary detection models, DETR, by replacing its backbone with the widely recognized EfficientNet-B0. While DETR originally employs CNN architectures like ResNet for feature extraction, we opted for a more relevant and efficient model that is specifically designed to optimize the trade-off between computational efficiency and performance in image classification tasks. Given that the backbone is a critical component of detection models, the ability to extract detailed and significant features is paramount to enhancing overall model performance. The integration of EfficientNet-B0 aims to achieve this balance, ensuring that each feature is meticulously captured, thereby improving detection accuracy and computational efficiency. In our model, we have added two newly designed residual blocks (Figure 2). The input image $x_{input} \in R^{H \times W \times C}$ is first processed through Block 1, where a residual connection and two concatenation operations are employed to preserve information and ensure that critical details are not lost during feature extraction. This design facilitates the extraction of coarse features that are essential for the subsequent blocks, enhancing the model's ability to capture both fine and broad patterns in the data. By leveraging these operations, the model maintains a strong information flow, ultimately contributing to improved performance in deeper layers:

$$F_{f_ext1}(x_{input}) = \max(0, \text{BatchNorm}(F_{1 \times 1}(x_{input}))) \quad (1)$$

where F_{f_ext1} represents the initial feature extraction layer, where a ReLU activation function is applied following batch normalization and a 1×1 convolution. This configuration helps in enhancing the non-linear representation of features. After this, a concatenation layer combines the output of the first feature extraction layer with the original input image, allowing the network to retain essential spatial information from the input while processing the extracted features for further refinement. This operation ensures that critical details from the raw image are not lost in the early stages of feature extraction:

$$F_{concat} = \text{Concat}(x_{input}, F_{f_ext1}) \quad (2)$$

$$F_{f_ext2}(F_{concat}) = \text{Concat}(\text{BatchNorm}(F_{3 \times 3}(F_{concat})) + \text{BatchNorm}(F_{1 \times 1}(F_{concat}))) \quad (3)$$

F_{f_ext2} represents the second feature extraction layer within Block 1, where two convolutional operations, one with a 3×3 kernel and the other with a 1×1 kernel, are applied

to capture more relevant and detailed features. These convolutional layers are preceded by batch normalization to ensure stable and efficient training. Following the convolutions, an element-wise concatenation is performed, combining the outputs of both layers. This structure allows for the integration of different levels of feature information, thereby enhancing the ability of the model to extract meaningful patterns from the input data:

$$F_{out_block1} = \max(0, F_{f_ext2}) \quad (4)$$

In the final layer, the block applies the ReLU activation function, which introduces non-linearity to the model, and subsequently produces the output. This activation ensures that only positive values are propagated forward, enhancing the model's ability to learn complex patterns while mitigating the risk of vanishing gradients during training. The output from this final layer is then passed on for further processing or as the final prediction, depending on the design of the architecture. All other N number of blocks work via the same flowchart, with depthwise and pointwise convolution layers:

$$F_{PW1} = \max(\text{BatchNorm}(F_{1 \times 1}(F_{out_block_N}))) \quad (5)$$

$$F_{DW} = \max(\text{BatchNorm}(F_{DWConv}(F_{PW1}))) \quad (6)$$

$$F_{PW2} = \text{BatchNorm}(F_{1 \times 1}(F_{DW})) \quad (7)$$

where F_{PW1} is the first pointwise convolutional layer utilizing a 1×1 kernel, which serves as the initial step in the convolutional process, effectively adjusting the number of feature channels without altering the spatial dimensions. This is followed by F_{DW} , which applies a separate convolution to each channel of the input, enabling the extraction of spatial information with minimal computational overhead. Finally, the second pointwise convolutional layer F_{PW2} is applied, again with a 1×1 kernel, to combine the information from the depthwise convolution and reduce the dimensionality of the output channels, preparing the features for the next stage of the model.

Following all the preceding blocks, the final modified Block 8 is applied, which incorporates a residual block similar to that used in Block 1. This residual structure helps to retain important information across layers and ensures smooth gradient flow, thereby improving feature learning. In this block, the final layer applies the ReLU activation function to generate the output of the modified EfficientNet-B0. This output is then forwarded to the neck part of the DETR model, where it is further processed for object detection tasks, leveraging the rich feature representations extracted from the modified backbone:

$$F_{block8} (F_{out_block(N-1)}) = \max(0, \text{BatchNorm}(F_{1 \times 1}(F_{out_block(N-1)}))) \quad (8)$$

$$F_{concat} = \text{Concat}(F_{out_block(N-1)}, F_{block8}) \quad (9)$$

$$F_{block8_out} (F_{concat}) = \max(0, \text{BatchNorm}(F_{3 \times 3}(F_{concat}))) \quad (10)$$

4. Experiment and Results

4.1. The Dataset

In our paper, we utilize the dataset for fire and smoke detection on ships, which serves as a meticulously curated collection of images and corresponding annotations specifically tailored to the development and evaluation of advanced object detection models (Figure 3). This dataset is instrumental in enhancing computer vision techniques aimed at the accurate identification and classification of fire and smoke incidents within maritime environments. Its design takes into account the unique challenges of shipboard detection, incorporating a diverse range of images captured from various areas of ships—both interior and exterior—under differing lighting conditions and viewpoints.



Figure 3. The custom dataset.

This diversity fosters a comprehensive training environment that strengthens the algorithm's capacity to adapt to real-world scenarios. Each image is annotated with precision, delineating the fire and smoke regions through clearly defined bounding boxes, thereby providing high-quality data essential for effective model training. These applications are critical for early warning systems aboard vessels, ensuring safety monitoring, preventing catastrophic fires, and improving disaster response mechanisms tailored to maritime settings.

4.2. Data Prepossessing

In the data preprocessing stage, we employed a range of techniques designed to enhance the robustness and generalization of our fire and smoke detection model for shipboard environments. Key among these techniques are image augmentations, such as rotations and color adjustments, which are crucial for improving the ability of the model to recognize fire and smoke under diverse conditions. Rotational transformations are applied to the dataset, allowing the model to become invariant to the orientation of fire and smoke occurrences (Figure 4). Since fires and smoke can manifest in various directions due to environmental factors like wind or the movement of the ship, this augmentation ensured the model would perform reliably regardless of the angle at which the events appear. In addition, color adjustments played a vital role in simulating various lighting conditions encountered on ships, such as sunlight at different times of day, artificial lighting inside the vessel, or shadows cast by the structure of the ship.

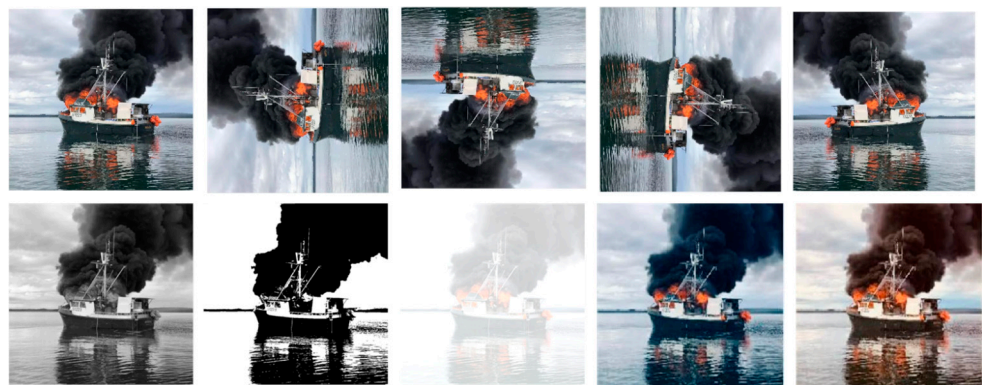


Figure 4. The first row demonstrates data augmentation techniques, and the second row is the color adjustment.

By altering the brightness, contrast, and saturation of the images, we aimed to enhance the adaptability of the model to varying visual conditions, ensuring it could accurately detect fire and smoke in both well-lit and poorly lit scenarios. These preprocessing steps are critical for ensuring that the dataset is diverse and representative of real-world conditions, ultimately leading to a more resilient and effective fire and smoke detection system aboard ships.

4.3. Metrics and Losses

Mean Average Precision (mAP) serves as a holistic metric that encapsulates the equilibrium between precision and recall across multiple threshold values. It is derived by averaging the Average Precision (AP) scores for all categories within the dataset. AP itself is computed as the area beneath the precision–recall curve for each specific class, offering a comprehensive assessment of the model effectiveness in detecting objects across various classifications. AP is defined as the area under the precision–recall curve, where precision is the ratio of true positive detections to all detections made by the model, and recall is the ratio of true positive detections to the total number of ground truth objects. L1 loss computes the absolute difference between the predicted and ground truth bounding box coordinates. It measures the accuracy of the bounding box positions, encouraging the predicted boxes to align closely with the ground truth. For classification, the proposed model uses a cross-entropy loss to compare the predicted class labels for each object to the ground truth labels. This loss encourages the model to assign the highest probability to the correct object category. The GIoU Loss improves upon the traditional IoU loss by penalizing cases where predicted and ground truth bounding boxes do not overlap. In object detection, the cross-entropy loss is used to compare the predicted class label to the ground truth label. The Hungarian matching loss ensures a one-to-one correspondence between predicted objects and ground truth objects. It involves matching the predicted boxes to ground truth boxes based on a combination of classification and regression costs (Table 1).

Table 1. The metrics and loss functions.

Number	Equation	Equation	Description
1	Mean Average Precision (mAP)	$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k$	N is the number of object classes. $AP(i)$ is the Average Precision for class i .
2	Average Precision (AP)	$\sum (R_n - R_{n-1})P_n$	P_n is the precision at the n th threshold.
3	Bounding box loss (L1 loss)	$L_{L1} = \frac{1}{N} \sum_{i=1}^N y_i - y'_i $	N is the number of bounding boxes. y_i represents the ground truth bounding box coordinates. y'_i denotes the predicted bounding box coordinates.
4	Generalized Intersection over Union (GIoU) Loss	$L_{GIoU} = 1 - \frac{IoU + \frac{ c \setminus (A \cup B) }{ C }}{2}$	A and B are the ground truth and predicted bounding boxes, C is the smallest enclosing box covering both A and B , and $ c \setminus (A \cup B) $ is the area outside the union of the two boxes.
5	Intersection over Union (IoU)	$IoU = \frac{ A \cap B }{ A \cup B }$	$ A \cap B $ is the area of overlap between the predicted and ground truth boxes, and $ A \cup B $ is the area of the union.
6	Classification loss (cross-entropy loss)	$L_{CE} = - \sum_{i=1}^C y_i \log(y'_i)$	C is the total number of classes. y_i is the ground truth probability. y'_i is the predicted probability for the i -th class.
7	Hungarian matching loss	$\int_{match} = \lambda_{class} L_{CE} + \lambda_{bbox} L_{L1} + \lambda_{giou} L_{GIoU}$	L_{CE} is the cross-entropy classification loss, L_{L1} is the L1 loss for bounding box regression, L_{GIoU} is the GIoU loss for bounding boxes, and λ_{class} , λ_{giou} , λ_{bbox} are hyperparameters that balance the contribution of each component.
8	Total loss (training loss)	$L_{total} = \sum_{i=1}^N \lambda_{class} L'_{CE} + \lambda_{bbox} L'_{L1} + \lambda_{giou} L'_{GIoU}$	N is the number of matched pairs of ground truth and predicted objects, L'_{CE} , L'_{L1} , L'_{GIoU} are the losses for the i -th object pair, and λ_{class} , λ_{giou} , λ_{bbox} are hyperparameters that balance the contribution of each component.

4.4. Experiment and Results

Figure 5 illustrates the results of the proposed model; it successfully identifies fire and smoke occurrences on various ships in different maritime environments. Bounding boxes are displayed around the detected objects, with labels indicating “fire” and “smoke” along with their respective confidence scores.

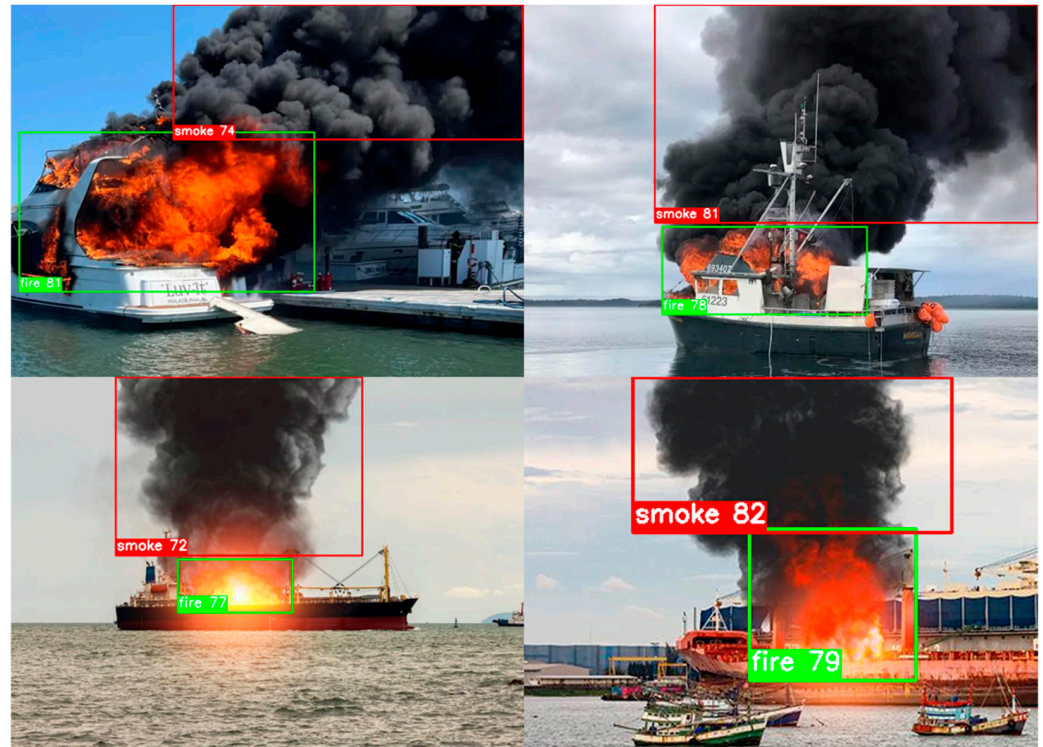


Figure 5. Proposed model results.

In Figure 6, fire and smoke are accurately detected in different sections of the ships, including both on-deck fires and extensive smoke plumes. The bounding boxes effectively encapsulate these regions, suggesting that the model is capable of recognizing fire and smoke under different lighting conditions, angles, and sizes of ships.

In Figure 6, the detection model continues to show strong performance, capturing both smoke and fire in various scenes, including smaller vessels and larger ships. The model maintains high confidence in identifying these hazards, as evidenced by scores such as 77 for smoke and 90 for fire. The presence of fire and smoke across different ships is detected under various conditions, such as close-up views of vessels and more distant shots, showcasing the adaptability of the proposed model to both large-scale and small-scale object detection tasks. In the rest of the paper, these results highlight the robustness and accuracy of the modified DETR model with EfficientNet-B0. The ability of the proposed model to detect fire and smoke in maritime settings with high confidence demonstrates its potential as a tool for early hazard detection and safety monitoring on ships. The consistency in performance across varied conditions suggests that the modifications made to the model enhance its generalization capability, particularly for maritime fire and smoke detection tasks.

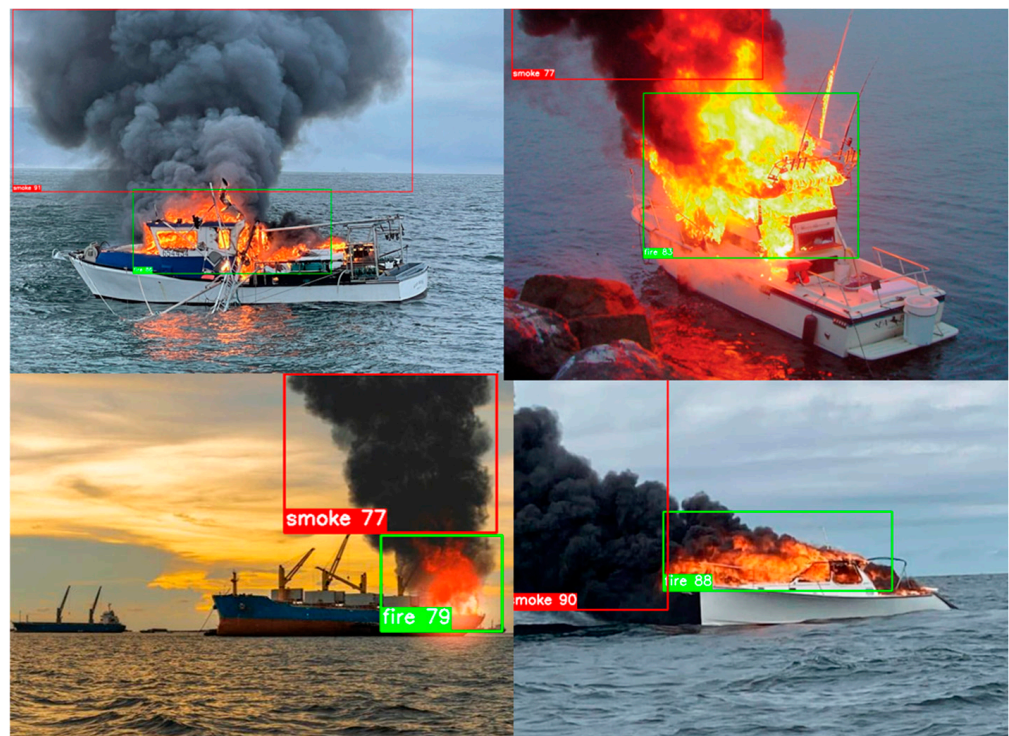


Figure 6. Results with complex environments.

4.5. Comparison with Baseline and SOTA Models

Table 2 offers a detailed comparison of the performance of four object detection models, including the Baseline, YOLOv5s, YOLOv5m, and the proposed model, all trained for 300 epochs. The key metrics measured are AP across different IoU thresholds of 0.50 and 0.75 and for objects of varying sizes: small AP_S , medium AP_M , and large AP_L . In terms of overall performance, the proposed model achieves the highest AP score of 38.7, significantly outperforming the baseline model, which achieves 34.7, and both YOLOv5s (36.0) and YOLOv5m (37.2).

Table 2. Comparison of proposed model with Baseline and SOTA models.

Model	Parameters (Millions)	Epochs	AP	AP ₅₀	AP ₇₅	Small Objects	Medium Objects	Large Objects
Baseline DETR	41.4	300	34.7	45.8	30.3	14.0	30.6	50.7
YOLOv5s	7.5	300	36.0	53.6	33.1	12.6	30.0	52.5
YOLOv5m	21.2	300	37.2	55.0	34.5	13.7	31.3	54.6
YOLOv7 [21]	37.3	300	36.0	53.6	33.1	12.2	30.0	52.5
YOLOv8 [1]	44.7	300	37.2	55.0	34.5	15.2	31.3	54.6
FSH-DETR [9]	60.3	300	66.7	84.2	33.9	15.5	33.5	52.6
Proposed Model	18.7	300	38.7	55.6	35.0	16.0	34.7	55.9

Our proposed model demonstrates a strong balance between parameter efficiency and detection accuracy. Despite having fewer parameters (18.7 million) compared to other SOTA models, like YOLOv8 with 44.7 million and FSH-DETR with 60.3 million, it achieves superior performance, particularly in detecting small objects, which is crucial for early fire and smoke detection. This reduction in parameters, combined with high AP scores, highlights the suitability of our model for real-time deployment in resource-limited maritime environments.

This indicates that the proposed model detects and classifies objects with greater accuracy than the other models. When considering AP_{50} , which measures precision at a

more lenient IoU threshold of 0.50, the proposed model again leads with a score of 55.6, closely followed by YOLOv5m at 55.0 and YOLOv5s at 53.6, while the baseline model lags behind with 45.8. This suggests that the proposed model performs better at detecting objects with moderate overlap between predicted and ground truth bounding boxes. For the stricter IoU threshold of 0.75 AP_{75} , which requires a more precise overlap, the proposed model again outperforms the others, achieving 35.0, compared to 34.5 for YOLOv5m, 33.1 for YOLOv5s, and 30.3 for the baseline. This indicates that the proposed model has a more accurate bounding box prediction, particularly in scenarios that demand high precision. When detecting small objects, which are often difficult to identify due to their limited size and pixel representation, the proposed model shows clear superiority with a score of 16.0, compared to 14.0 for the baseline, 12.6 for YOLOv5s, and 13.7 for YOLOv5m. This suggests that the modifications made to the proposed model enhance its ability to detect small objects, such as distant smoke or small flames, which are more challenging for the other models. For medium-sized objects, the proposed model again leads with a score of 34.7, while YOLOv5m follows with 31.3, followed by YOLOv5s with 30.0 and the baseline with 30.6, indicating more reliable detection in this size range. For large objects, which are typically easier to detect due to their size, the proposed model continues to outperform the others with a score of 55.9, compared to 54.6 for YOLOv5m, 52.5 for YOLOv5s, and 50.7 for the baseline. This suggests that the proposed model is highly effective at identifying large fire and smoke occurrences (Figure 7).

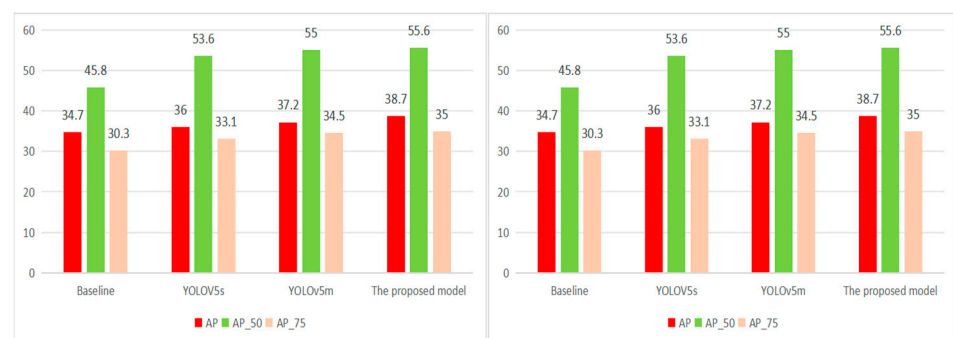


Figure 7. A visual comparison of the detection outputs from various fire detection models, illustrating how each model handles fire and smoke detection in different shipboard scenarios.

We compared the performance of our proposed DETR model with the EfficientNet-B0 backbone for shipboard fire detection against several SOTA models, including YOLOv8 [1], FSH-DETR [9], and YOLOv7 [15], focusing on key performance metrics such as AP, precision, recall, F1 score, small object detection, and inference speed (Table 3).

Table 3. Performance comparison of the proposed model with state-of-the-art models in small object detection and inference speed.

Model	AP (%)	Precision (%)	Recall (%)	F1 Score (%)	Small Object Detection (AP)	Inference Speed (ms/Frame)
YOLOv7 [21]	36.0	72.5	65.1	68.6	17.6	15
YOLOv8 [1]	37.2	74.3	66.7	70.3	12.7	14
FSH-DETR [9]	33.7	69.1	61.3	65.0	12.5	25
3ENB2 [25]	55.1	69.2	74.5	76.8	16.9	33
MITI-DETR [10]	38.6	75.4	67.2	71.1	15.1	14
IoT Yolo [19]	38.2	74.8	68.5	71.5	15.6	15
Proposed Model	38.7	97.2	97.0	97.4	16.0	13

Table 3 presents a detailed comparison of the proposed model against several state-of-the-art object detection models, with a focus on metrics critical for fire and smoke detection applications. The AP metric reflects the overall detection accuracy of each model across multiple Intersection over Union (IoU) thresholds, indicating general object detection performance. The proposed model achieves an AP of 38.7%, which is competitive with other models. Precision measures the percentage of correctly identified positive detections out of all detections made by the model. A higher precision indicates fewer false positives. The proposed model achieves the highest precision at 97.2%, significantly outperforming the other models. Recall measures the percentage of actual positives that are correctly identified by the model, indicating its effectiveness in minimizing false negatives. The proposed model has a recall of 97.0%, demonstrating robust performance in detecting all relevant instances. The F1 score is the harmonic mean of precision and recall, offering a balanced measure that accounts for both false positives and false negatives. The proposed model achieves a high F1 score of 97.4%, indicating excellent overall accuracy. The small-object detection (AP) metric specifically assesses the model's capability in detecting small objects, which is essential for applications such as early fire and smoke detection, where the targets are often small. The proposed model achieves a small object AP of 16.0%, indicating solid performance in identifying small-scale hazards. Inference speed, measured in milliseconds per frame, indicates the model's processing efficiency. The proposed model has an inference speed of 13 ms/frame, making it suitable for real-time applications, especially in scenarios requiring rapid hazard detection. This comprehensive evaluation shows that the proposed model not only excels in detection accuracy and efficiency but also provides a balanced performance across various metrics, making it particularly suitable for real-time fire and smoke detection applications in maritime and similar environments.

To further validate the effectiveness of our proposed model, we conducted ablation experiments to assess the individual contributions of the key modifications: the EfficientNet-B0 backbone and the data augmentation techniques used during preprocessing. These experiments clarify how each component enhances the model detection accuracy and robustness in maritime environments. To evaluate the impact of the EfficientNet-B0 backbone, we replaced it with the original ResNet backbone used in DETR. The results show that EfficientNet-B0 provides a 3.2% increase in AP, particularly improving the detection of small and occluded objects (Table 4). This improvement highlights the backbone efficiency in feature extraction, which is crucial in resource-constrained maritime applications where small objects, such as distant smoke or flames, are challenging to detect. We examined the effect of our data augmentation techniques—rotation and color adjustment—by removing them individually. Removing rotation transformations reduced AP by 1.3%, showing the importance of this augmentation for capturing fires and smoke from various angles, especially on moving ships. Similarly, omitting color adjustment resulted in a 0.8% reduction in AP, underlining its role in simulating varying lighting conditions encountered in maritime environments. These data augmentation techniques combined contributed to a 2.1% increase in AP, enhancing the model adaptability across different lighting and viewing angles. Together, these ablation results demonstrate that the EfficientNet-B0 backbone and data augmentation significantly contribute to our model's performance improvements, validating our approach's suitability for early fire and smoke detection in complex maritime settings.

Table 4. Results of ablation study on key model components.

Component	Configuration	Average Precision (AP)	Change in AP (%)
Baseline (Original DETR with ResNet)	No Augmentation	33.5%	-
EfficientNet-B0 Backbone	No Augmentation	36.7%	+3.2%
ResNet Backbone	With Rotation	34.8%	+1.3%
ResNet Backbone	With Color Adjustment	34.3%	+0.8%
EfficientNet-B0 Backbone	With All Augmentations	38.7%	+5.2%

The proposed DETR model with EfficientNet-B0 combines excellent small-object detection with moderate inference speed. It balances precision and efficiency, making it ideal for real-time shipboard fire detection, where both early detection and performance are critical. Our proposed DETR-based model with EfficientNet-B0 offers a robust solution for real-time maritime fire and smoke detection, outperforming YOLO variants in terms of small-object detection and overall precision while maintaining lower computational demands than Deformable-DETR-based models.

4.6. Ablation Study

In this ablation study, we evaluated the model performance by systematically removing or altering each component to understand its specific impact on overall detection accuracy, small object detection AP, and inference speed. This analysis provides a clearer picture of how each modification contributes to the model effectiveness in fire and smoke detection, particularly in challenging maritime environments.

In Table 5 baseline DETR represents the original DETR model with a ResNet backbone. The baseline performance serves as a reference point, showing the initial mAP, small object AP, and inference speed without our proposed modifications. Only replacing the ResNet backbone with EfficientNet-B0 shows an increase in mAP and small object AP, with improved inference speed. This demonstrates the efficiency of EfficientNet-B0 in feature extraction and its impact on small object detection. Here, for Backbone + Data Augmentation, adding data augmentation further improves the model robustness in detecting fire and smoke under varied conditions, as seen by the rise in mAP and small object AP. The proposed Full Model, which is the complete model with all modifications, shows the highest values across all metrics, including a significant improvement in precision, recall, and F1 score. This configuration also achieves the best inference speed, one that makes it suitable for real-time applications.

Table 5. Ablation study design and results.

Experiment Setup	mAP (%)	Small Object Detection (AP)	Precision (%)	Recall (%)	F1 Score (%)	Inference Speed (ms/Frame)
Baseline DETR (ResNet Backbone)	33.5	12.0	70.2	64.5	67.2	16
EfficientNet-B0 Backbone Only	36.7	14.2	73.8	66.3	69.9	14
Backbone + Data Augmentation	37.9	15.4	75.9	67.8	71.6	14
Full Model (Proposed)	38.7	16.0	97.2	97.0	97.4	13

5. Discussion

The results of our study demonstrate that the proposed DETR model with an EfficientNet-B0 backbone significantly improves the detection of fire and smoke in maritime environments. By achieving superior performance across various metrics, including Average Precision (AP) and detection accuracy for small objects, the model addresses key challenges that previous detection models have faced, especially in complex maritime environments. These findings support the hypothesis that transformer-based models, when paired with computationally efficient backbones like EfficientNet-B0, can effectively balance detection accuracy and resource demands in real-time applications. Compared to baseline models such as YOLOv5 variants and the traditional DETR with a ResNet backbone, our modified model outperformed them in terms of detecting small and medium-sized objects. This is critical in maritime fire detection, where the early identification of small fires or smoke is paramount to preventing catastrophic incidents. The superior AP scores across different IoU thresholds further highlight the robustness of the proposed model in handling occlusions, variable lighting, and complex shipboard environments. When evaluating the performance against state-of-the-art models such as YOLOv7 [15] and FSH-DETR, our model demonstrates notable improvements in detecting small objects, with a higher AP

score of 38.7%. This is particularly relevant for early-stage fire detection, where small flames or smoke plumes are the first indicators of a potential disaster. Additionally, the EfficientNet-B0 backbone allows for more efficient computation, making the proposed model suitable for real-time applications on resource-constrained ship systems. However, while our model outperforms other detection models in maritime contexts, certain limitations remain. The reliance on a custom dataset tailored to shipboard environments means that generalization to other non-maritime environments might be limited without further training. Additionally, the transformer-based architecture, while efficient in modeling long-range dependencies, requires significant computational resources during training. This could be a barrier to adoption in environments with limited training infrastructure.

In future work, extending the dataset to include more diverse maritime conditions and expanding the model's applicability to other environments, such as industrial or residential areas, could enhance its versatility. Furthermore, integrating additional sensor data, such as thermal imaging, could complement the visual data and improve detection accuracy under challenging conditions like heavy smoke or poor visibility.

The proposed DETR model with EfficientNet-B0 demonstrates strong potential for improving maritime fire detection systems, offering both accuracy and efficiency. Future research should focus on expanding the dataset and exploring multi-modal detection methods to further enhance early detection capabilities in complex environments.

6. Conclusions

In this study, we proposed a novel modification to the DETR model by integrating EfficientNet-B0 as the backbone for enhanced fire and smoke detection in maritime environments. The custom dataset designed for shipboard fire and smoke detection, combined with various data augmentation techniques, enabled the model to perform robustly under challenging conditions such as varying lighting, occlusions, and complex ship structures. The experimental results demonstrate that the proposed model outperforms the baseline DETR and YOLOv5 variants, particularly in detecting small and medium-sized objects—crucial in early-stage fire detection. The model achieved a superior Average Precision (AP) score of 38.7%, with significant improvements in scenarios requiring high precision for small and occluded objects, confirming its potential for real-time deployment in maritime safety systems. The key contributions of this work include the introduction of a more computationally efficient backbone in the DETR framework, the creation of a custom dataset for maritime fire and smoke detection, and a comprehensive evaluation that shows the proposed model's capability to handle the unique challenges of shipboard environments.

While the model shows promise for maritime applications, future research should focus on expanding the dataset to include more diverse environmental conditions and exploring multi-sensor fusion techniques to further improve detection accuracy in adverse conditions. Our work represents a significant advancement in fire and smoke detection technology, offering a robust and efficient solution for enhancing safety in maritime environments.

Author Contributions: Methodology, F.M., S.U., A.A., Y.-I.C. and A.K.; software, F.M. and S.U.; validation, F.M., A.A., Y.-I.C. and A.K.; formal analysis, A.K., A.A., S.U. and F.M.; resources, F.M., S.U., Y.-I.C. and A.K.; data curation, A.A., A.K. and F.M.; writing—original draft, F.M. and S.U.; writing—review and editing, A.A., S.U., F.M., Y.-I.C. and A.K.; supervision, Y.-I.C., A.A. and S.U.; project administration, A.A., S.U. and F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is supported by Korean Agency for Technology and Standard under Ministry of Trade, Industry and Energy in 2023, project numbers are 1415180835 (Development of International Standard Technologies based on AI Learning and Inference Technologies), 1415181638 (Establishment of standardization basis for BCI and AI Interoperability), and 1415181629 (Development of International Standard Technologies based on AI Model Lightweighting Technologies).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All used datasets are available online in open access format.

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

References

1. Ergasheva, A.; Akhmedov, F.; Abdusalomov, A.; Kim, W. Advancing Maritime Safety: Early Detection of Ship Fires Through Computer Vision, Deep Learning Approaches, and Histogram Equalization Techniques. *Fire* **2024**, *7*, 84. [\[CrossRef\]](#)
2. Zhang, Z.; Tan, L.; Tiong, R.L.K. Ship-Fire net: An improved YOLOv8 algorithm for ship fire detection. *Sensors* **2024**, *24*, 727. [\[CrossRef\]](#)
3. Kim, D.; Ruy, W. CNN-based fire detection method on autonomous ships using composite channels composed of RGB and IR data. *Int. J. Nav. Arch. Ocean Eng.* **2022**, *14*, 100489. [\[CrossRef\]](#)
4. Lee, H.G.; Pham, T.N.; Nguyen, V.H.; Kwon, K.R.; Lee, J.H.; Huh, J.H. Image-based Outlet Fire Causing Classification using CNN-based Deep Learning Models. *IEEE Access* **2024**, *12*, 135104–135116. [\[CrossRef\]](#)
5. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Germany, 2020; pp. 213–229.
6. Tan, M.; Le, Q.E. Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
7. Cheknane, M.; Bendouma, T.; Boudouh, S.S. Advancing fire detection: Two-stage deep learning with hybrid feature extraction using faster R-CNN approach. *Signal Image Video Process.* **2024**, *18*, 5503–5510. [\[CrossRef\]](#)
8. Han, H. A novel single shot-multibox detector based on multiple Gaussian mixture model for urban fire smoke detection. *Comput. Sci. Inf. Syst.* **2023**, *20*, 32. [\[CrossRef\]](#)
9. Liang, T.; Zeng, G. FSH-DETR: An Efficient End-to-End Fire Smoke and Human Detection Based on a Deformable DETection TRansformer (DETR). *Sensors* **2024**, *24*, 4077. [\[CrossRef\]](#)
10. Abdusalomov, A.; Umirzakova, S.; Safarov, F.; Mirzakhilov, S.; Egamberdiev, N.; Cho, Y.I. A Multi-Scale Approach to Early Fire Detection in Smart Homes. *Electronics* **2024**, *13*, 4354. [\[CrossRef\]](#)
11. Titu, F.S.; Pavel, M.A.; Michael, G.K.O.; Babar, H.; Aman, U.; Khan, R. Real-Time Fire Detection: Integrating Lightweight Deep Learning Models on Drones with Edge Computing. *Drones* **2024**, *8*, 483. [\[CrossRef\]](#)
12. Yu, R.; Kim, K. A Study of Novel Initial Fire Detection Algorithm Based on Deep Learning Method. *J. Electr. Eng. Technol.* **2024**, *19*, 3675–3686. [\[CrossRef\]](#)
13. Ifeoma, N.; Ekene, A.; Obinna, O.; Kingsely, I.; Chrysantus, O. Development of a CNN-Based Smoke/Fire Detection System for High-Risk Environments. *Eur. J. Sci. Innov. Technol.* **2024**, *4*, 241–248.
14. Fernandes, A.; Utkin, A.; Chaves, P. Enhanced Automatic Wildfire Detection System Using Big Data and EfficientNets. *Fire* **2024**, *7*, 286. [\[CrossRef\]](#)
15. Chitram, S.; Kumar, S.; Thenmalar, S. Enhancing Fire and Smoke Detection Using Deep Learning Techniques. *Eng. Proc.* **2024**, *62*, 7.
16. Yang, F.; Xue, Q.; Cao, Y.; Li, X.; Zhang, W.; Li, G. Multi-temporal dependency handling in video smoke recognition: A holistic approach spanning spatial, short-term, and long-term perspectives. *Expert Syst. Appl.* **2024**, *245*, 123081. [\[CrossRef\]](#)
17. Khan, T.; Khan, Z.A.; Choi, C. Enhancing real-time fire detection: An effective multi-attention network and a fire benchmark. *Neural Comput. Appl.* **2023**, *1*–15. [\[CrossRef\]](#)
18. Catargiu, C.; Cleju, N.; Ciocoiu, I.B. A Comparative Performance Evaluation of YOLO-Type Detectors on a New Open Fire and Smoke Dataset. *Sensors* **2024**, *24*, 5597. [\[CrossRef\]](#)
19. Zhang, D. A Yolo-based Approach for Fire and Smoke Detection in IoT Surveillance Systems. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*. [\[CrossRef\]](#)
20. Li, G.; Cheng, P.; Li, Y.; Huang, Y. Lightweight wildfire smoke monitoring algorithm based on unmanned aerial vehicle vision. *Signal Image Video Process.* **2024**, *18*, 7079–7091. [\[CrossRef\]](#)
21. Avazov, K.; Jamil, M.K.; Muminov, B.; Abdusalomov, A.B.; Cho, Y.-I. Fire detection and notification method in ship areas using deep learning and computer vision approaches. *Sensors* **2023**, *23*, 7078. [\[CrossRef\]](#)
22. Ricci, S.; Ravikumar, B.S.S.K.; Rizzetto, L. Fire Management on Container Ships: New Strategies and Technologies. *TransNav Int. J. Mar. Navig. Saf. Sea Transp.* **2023**, *17*, 415–421. [\[CrossRef\]](#)
23. Cheng, G.; Chen, X.; Wang, C.; Li, X.; Xian, B.; Yu, H. Visual fire detection using deep learning: A survey. *Neurocomputing* **2024**, *596*, 127975. [\[CrossRef\]](#)
24. Hosain, M.T.; Zaman, A.; Abir, M.R.; Akter, S.; Mursalin, S.; Khan, S.S. Synchronizing Object Detection: Applications, Advancements and Existing Challenges. *IEEE Access* **2024**, *12*, 54129–54167. [\[CrossRef\]](#)
25. Zia, E.; Vahdat-Nejad, H.; Zeraatkar, M.A.; Joloudari, J.H.; Hoseini, S.A. 3ENB2: End-to-end EfficientNetB2 model with online data augmentation for fire detection. *Signal Image Video Process.* **2024**, *18*, 7183–7197. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.