information

MDPI

Article

Fast Object Detection Leveraging Global Feature Fusion in Boundary-Aware Convolutional Networks

Weiming Fan [1], **Jiahui Yu** [2] **and Zhaojie Ju** [3,*]

[1] School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China
[2] Department of Biomedical Engineering, Zhejiang University, Hangzhou 310058, China
[3] School of Computing, University of Portsmouth, Portsmouth PO1 3HE, UK
* Correspondence: zhaojie.ju@port.ac.uk

Abstract: Endoscopy, a pervasive instrument for the diagnosis and treatment of hollow anatomical structures, conventionally necessitates the arduous manual scrutiny of seasoned medical experts. Nevertheless, the recent strides in deep learning technologies proffer novel avenues for research, endowing it with the potential for amplified robustness and precision, accompanied by the pledge of cost abatement in detection procedures, while simultaneously providing substantial assistance to clinical practitioners. Within this investigation, we usher in an innovative technique for the identification of anomalies in endoscopic imagery, christened as Context-enhanced Feature Fusion with Boundary-aware Convolution (GFFBAC). We employ the Context-enhanced Feature Fusion (CEFF) methodology, underpinned by Convolutional Neural Networks (CNNs), to establish equilibrium amidst the tiers of the feature pyramids. These intricately harnessed features are subsequently amalgamated into the Boundary-aware Convolution (BAC) module to reinforce both the faculties of localization and classification. A thorough exploration conducted across three disparate datasets elucidates that the proposition not only surpasses its contemporaries in object detection performance but also yields detection boxes of heightened precision.

Keywords: object detection; polyps; endoscopy; deep learning; computer vision

Citation: Fan, W.; Yu, J.; Ju, Z. Fast Object Detection Leveraging Global Feature Fusion in Boundary-Aware Convolutional Networks. *Information* **2024**, *15*, 53. https://doi.org/10.3390/info15010053

Academic Editors: Ilias Theodorakopoulos and Nikolaos Mitianoudis

Received: 16 November 2023
Revised: 19 December 2023
Accepted: 28 December 2023
Published: 17 January 2024

1. Introduction

Deep learning has surfaced as a transformative influence in the realm of medical image analysis, profoundly impacting the domain of endoscopy, which encompasses specialized imaging modalities for diagnostics and therapeutic interventions [1]. This symbiotic relationship has yielded noteworthy applications. Convolutional neural networks (CNNs) demonstrate exceptional proficiency in real-time polyp detection during colonoscopy, enhancing the prospects of early intervention and the potential prevention of colorectal cancer [2]. When it comes to diagnosing Barrett's esophagus, a precursor to esophageal adenocarcinoma, deep learning surpasses human assessments by adeptly scrutinizing endoscopic images, thereby facilitating early detection [3]. Deep learning models additionally gauge the quality of endoscopic procedures through real-time video analysis, ensuring adherence to established standards, consequently amplifying the dependability and efficacy of the practice. These applications underscore the capacity of deep learning to elevate precision, efficiency, and reliability within the realm of endoscopy, thereby propelling forward patient care. The competence to scrutinize copious volumes of image and video data reinforces the pivotal role that deep learning assumes in the domain of endoscopy. This amalgamation has inaugurated an era of innovation, thereby elevating healthcare practices and promising a brighter future for medical diagnostics and treatments [4,5].

In recent years, the domain of object detection has witnessed remarkable progress, primarily attributable to the emergence of deep learning methodologies. Object detection, the task of precisely identifying and localizing objects within images or video frames, holds paramount importance in various applications, encompassing surveillance, autonomous driving, and medical imaging [6,7]. The YOLO (You Only Look Once) series of models has played a pivotal role in addressing critical challenges within this domain [8,9]. One primary challenge tackled by the YOLO series pertains to the detection of small objects. Traditional methods faced limitations in accurately discerning diminutive objects due to spatial constraints. YOLOv3 addressed this issue by introducing multi-scale detection capabilities, thereby enhancing the model's proficiency in identifying small objects [10]. YOLO models strike an equilibrium between processing speed and detection accuracy, rendering them highly suitable for real-time applications. Another significant challenge that the YOLO series resolves is multi-class object detection. YOLO empowers the detection of objects belonging to diverse categories in a single inference pass, resulting in heightened efficiency [11]. This is particularly valuable in applications such as autonomous vehicles navigating intricate environments, where the identification and categorization of multiple objects within a scene are imperative. Notably, YOLO excels in real-time object detection, even in dynamic and rapidly evolving scenarios [12]. This capability has rendered YOLO indispensable in applications like video surveillance, traffic management, and augmented reality. YOLO accomplishes this through the optimization of network architecture and inference processes. In summary, the YOLO series has transformed object detection by effectively addressing challenges related to small object detection, providing real-time capabilities, and facilitating multi-class detection. These technological advancements have significantly expanded the realm of object detection applications across diverse domains, thereby carrying substantial implications for safety, security, and operational efficiency. As the field of object detection continues its evolution, YOLO remains at the forefront of this transformative landscape.

Over the course of the last decade, computer vision has demonstrated remarkable advancements; nonetheless, it grapples with substantial challenges when confronted with real-world applications. Particularly noteworthy among these challenges is the predicament of intra-class variation, where objects belonging to the same class manifest significant disparities attributable to factors such as occlusion, varying illuminative conditions, diverse poses, and alterations in viewpoint [13]. Furthermore, objects may incur non-rigid deformations or undergo rotations, scalings, and blurriness, thereby complicating their extraction and precise recognition [14]. In certain instances, objects may find themselves ensconced in inconspicuous surroundings, further heightening the complexity of recognition. Another conspicuous challenge in the realm of computer vision pertains to the extensive multitude of object categories necessitating classification. Effectively addressing this challenge mandates unfettered access to voluminous, high-quality annotated data for the purpose of training object detectors. However, the dearth of such data poses a substantial impediment, rendering the development of robust models a more intricate task. Additionally, the matter of training object detectors with limited exemplars remains a fertile area of ongoing research. Efficiency represents a paramount concern within the discipline, considering the substantial computational resources demanded by modern models to attain precise object detection. This challenge is accentuated by the growing ubiquity of mobile and edge devices, underscoring the exigency of developing streamlined object detection methodologies to further advance the field of computer vision. The successful surmounting of these challenges constitutes a fundamental imperative in order to fully harness the potential of computer vision across a diverse spectrum of real-world applications.

In this undertaking, our objective is to explore a novel methodology that can effectively address precision and minimize expenditure. To commence, in the interest of shortening the information path, we employ low-level fine-grained localization signals to fortify the feature pyramid, thus creating an enhancement from shallow to deep layers. In practice, shallow attributes have been utilized in the systems denoted by [15–22]. Yet, there has been a scarcity of research concerning the propagation of shallow attributes to enhance the entire feature hierarchy for instance recognition. Subsequently, to rectify the issue of imbalanced feature hierarchy, our approach recognizes that deep high-level features from the backbone network carry more substantial semantic content, whereas the shallow low-level attributes are primarily content-descriptive. In the domain of object detection, low-level and high-level information are mutually complementary. Our research elucidates the necessity of achieving balanced feature integration from each resolution. However, different aggregation sequences can result in integrated features focusing more on adjacent resolutions and less on others. In the information flow, each fusion operation dilutes semantic information from non-adjacent levels. We leverage balanced semantic features integrated at the same depth to enhance multi-level attributes. Lastly, we introduce the concept of boundary-aware convolutions, where each side of the barrel is individually positioned based on the surrounding context. Furthermore, to maintain precise local bounding boxes during non-maximum suppression, we recommend adjusting classification scores based on barrel confidence, further enhancing overall performance. We have demonstrated cutting-edge performance across multiple datasets. Employing ResNet as the foundational network, our model has outperformed several advanced object detectors in single-scale testing for object detection tasks, underscoring the effectiveness of our approach.

In this segment, we furnish an exposition of our principal contributions:

1. We introduce a boundary-aware convolution technique, denominated as BAC, meticulously crafted for the efficient detection of objects within the realm of endoscopy.
2. We proffer a stratagem to elevate the attributes residing in the shallow layers, thus engendering equilibrium in the domain of features. We optimize multi-tier features by judiciously harmonizing the influence of superficial and profound informational strata.
3. We execute comprehensive assessments of the envisaged framework across three distinct datasets. These evaluations exhibit unwaveringly noteworthy enhancements in comparison to the most advanced detectors, encompassing both singular-stage and dual-stage detectors.

2. Related Work

2.1. YOLO

YOLO made its debut on the 8th of June in 2015, ushering in a distinctive approach by framing the detection task as a regression conundrum [23]. It achieved the simultaneous output of both positional and class-related information through the conduit of a solitary neural network. When juxtaposed with the Fast R-CNN, YOLO notably curtailed the incidence of background errors in its predictions. YOLO captured substantial attention owing to its astonishing swiftness and the augmentation of localization precision. In recent years, dedicated researchers have tirelessly toiled on the amelioration of the YOLO framework. In the year 2020, Glenn Jocher, the luminary CEO of Ultralytics, unveiled YOLO (v5) on GitHub, endowing it with a plethora of invaluable attributes. These include test-time augmentation, model ensembling, hyperparameter evolution, and the faculty to export models in an array of formats like ONNX, CoreML, and TFLite. For the enhancement of accuracy in video detection, the deployment of automatic anchoring techniques was instrumental. This innovation bestowed dedicated anchor boxes to each component of the network. As for YOLO (v6), it witnessed enhancements across the backbone, neck, and head of the model, ingeniously addressing the pragmatic concerns germane to industrial video detection applications [24]. YOLO (V6) prides itself on its twofold improvement in inference speed compared to V5, all the while achieving a higher mean average precision (mAP).

The realm of natural image processing has commenced to harvest the rewards of triumphant object detection. Simon et al. ushered in a model employing the Complexer-YOLO architecture for real-time 3D object detection [25]. They harnessed spatial LiDAR data and incorporated a 2D scene understanding to attain competitive performance on the KITTI benchmark. Real-time models in the domains of Automated Driving Systems (ADS) and Driver Assistance Systems (DAS) have hitherto grappled with issues related to diminished accuracy and suboptimal performance. Han et al. proffered an innovative real-time object detection model, O-YOLO-v2, seamlessly integrated within a deep learning framework [26]. This model introduced a fresh architecture that amplifies the network's feature extraction capabilities by embedding convolutional layers at diverse junctures. Simultaneously, it adeptly addressed the quandaries stemming from augmented network depth, notably the predicaments of gradient vanishing or exploding, through the assimilation of residual modules. This judicious approach yielded successful experimental outcomes on the KITTI dataset. In a bid to surmount the challenge of detecting faces at varying scales, Chen et al. introduced a face detector christened YOLO-face, predicated on the bedrock of the YOLOv3 framework [27]. This stratagem aspired to elevate face detection performance by leveraging anchor boxes more tailored for facial detection and a regression loss function of greater precision. This enhanced detector conspicuously elevated accuracy while upholding alacrity in the realm of detection.

The unique attributes of endoscopy images, when juxtaposed with their natural counterparts, bestow upon them marked disparities, courtesy of their specialized essence. These disparities encompass divergences in resolution, hue, and luminance, which firmly delineate endoscopy images from their natural brethren [28]. Furthermore, endoscopy images encapsulate an array of distinctive targets, exhibiting a substantial breadth of diversity. These targets are predominantly confined to the medical realm, enshrining entities such as lesions, neoplasms, ulcers, and more [29]. This diversity extends to facets like morphology, chroma, and tactile qualities. Consequently, the imperative unfurls for the development of detectors, painstakingly tailored to confront these distinctive challenges intrinsic to the domain of medical applications.

2.2. Object Detection

Deep learning, a subdivision within the realm of machine learning, proffers an array of paramount advantages [30]. Inaugurally, it excels in the acquisition of intricate features and patterns from expansive datasets, thereby culminating in substantial performance ameliorations across domains encompassing image scrutiny, speech discernment, and natural language comprehension. Secondly, the multi-tiered neural networks inherent to deep learning autonomously distill abstruse, high-level features, thus mitigating the necessity for laborious manual feature engineering, thereby simplifying the process at hand [31]. Furthermore, deep learning models demonstrate exceptional prowess in effectively tackling grandiose quandaries, spanning extensive image classification to the intricate domain of autonomous vehicular navigation [32]. Foremost among its accolades, deep learning unfolds extraordinary adeptness in surmounting intricate tasks such as natural language apprehension, speech recognition, and the explication of medical images. These attributes distinctly posit deep learning as a formidable instrument poised to unravel an extensive gamut of real-world conundrums.

The principal objective of general object detection resides in the discernment and classification of entities embedded within an image, aptly adorning them with rectangular bounding enclosures signifying their degrees of assurance. Elevating the caliber of prospective bounding enclosures and harnessing profound architectural frameworks for elevated-level feature extraction stands as a matter of paramount significance. To address these formidable quandaries, Ross Girshick unfurled R-CNN in the year 2014 [33], culminating in the attainment of a remarkable average precision (mAP) reaching 53.3% on the PASCAL VOC 2012 dataset. SPP-net adeptly confronted the predicament associated with rigid dimensions in fully connected layers by assimilating spatial pyramid matching (SPM) [34–36], which bestowed the capacity to perceive entities across diverse magnitudes without incurring forfeiture or deformation of their inherent essence, particularly when entities exhibit variances in dimensions. Fast R-CNN resolved the issue of languid region-based object detection through the introduction of a streamlined and cohesive framework that harmonizes region proposals with the art of feature extraction [37], thereby engendering notable advancements in processing velocity and precision. Faster R-CNN heightened the efficiency of candidate enclosure generation in object detection through the introduction of a Region Proposal Network (RPN) in harmonious alliance with the detection network [38], leading to a substantial enhancement in the process of proposal calculation. Mask R-CNN adroitly contended with the formidable task of instance segmentation through the unveiling of a concurrent branch dedicated to pixel-by-pixel segmentation mask prognosis [39], enabling the simultaneous discovery and segmentation of entities nestling within an image.

Object detection is a fundamental task in computer vision, yet it presents several challenges and issues. Firstly, detecting small objects is a significant problem as they are prone to being overlooked, especially in complex backgrounds. Secondly, object occlusion is a common issue, increasing detection difficulty when objects are obscured by other entities. Multiscale object detection involves objects of various sizes and scales, necessitating effective handling. Illumination variations, background interference, and noise also introduce disturbances in object detection. Additionally, the acquisition and quality of labeled data are crucial for the performance of deep learning models, but labeling data is typically expensive and time-consuming. Lastly, achieving both generality and real-time capabilities in object detection systems is challenging, as different application domains require different detection models and speeds. Consequently, re-searchers continue to focus on improving object detection performance and applicability in the domains of small object detection, occlusion handling, multiscale adaptability, robustness against interference, data annotation, and model generality.

2.3. Bounding Boxes in Clinical Endoscopy

Bounding boxes assume a paramount role in clinical endoscopy for various rationales. Foremost among these is their capacity to facilitate the meticulous localization and discrimination of pathological regions within the realm of medical imagery. This function carries profound significance in the realm of early ailment diagnosis and therapeutics, embracing the pivotal role of detecting anomalies, such as lesions or polyps in the realm of colonoscopy images and the discernment of tumors in the context of endoscopic investigations [40]. Moreover, bounding boxes serve as instrumental tools in the measurement and quantification of the dimensions of these said anomalies. Thus, they empower precise assessments, ushering in a new era of accuracy in the evaluation of ailment progression. In a culminating fashion, bounding boxes confer invaluable data for the edification of computer-aided diagnostic systems, thus underpinning the machinery of automated analysis and lending robust support to healthcare professionals in the act of arriving at judicious and well-informed determinations.

In the realm of endoscopy, we encounter enduring challenges in the domain of object detection. Firstly, the intrinsic attributes of medical imagery bestow upon us datasets that are relatively diminutive and lack the diversity that characterizes general datasets [41]. These limitations thereby place constrictions on the efficacy of deep learning models. Secondly, the profuse heterogeneity across diverse endoscopic scenarios and equipment configurations imposes impediments on model generalization, necessitating a heightened degree of adaptability tailored to specific contexts. Thirdly, the exacting nature of medical applications necessitates outcomes that are endowed with an exquisite level of precision, thereby elevating the requisites for model performance and stability [42]. Lastly, the pursuit of interpretability and comprehensibility in the realm of automated detection presents a formidable challenge, as healthcare professionals seek to possess an all-encompassing grasp of the model's cognitive processes. Our approach is anchored in the enhancement of object bounding boxes, revolving around the meticulous refinement of their boundaries. We employ the barrel scheme to disentangle the intricate process of boundary localization for each object. Furthermore, we harness the estimates from the barrel schema to augment the outcomes of classification. Rigorous testing on three distinct datasets attests to the model's robust performance and its capacity to exhibit a degree of versatility.

3. Materials and Methods

The essence of object detection resides in the exactitude of target localization. Currently, the dominant methodology hinges on predetermined anchor boxes. Nevertheless, this framework proves inadequate in furnishing pinpoint localization for diminutive targets and competently managing intersecting objects. Various approaches [43–46] have striven to augment localization precision by assimilating multi-scale features. Nevertheless, this refinement frequently accompanies heightened computational intricacy and presents dilemmas in the selection of judicious hyperparameters. Hence, there exists an imperative need for the introduction of a nimble and efficacious substitute.

We introduce a Context-enhanced Feature Fusion Module and a Boundary-Aware Con-volution Module to elevate edge features and attain a higher degree of precision in target localization. As shown in Figure 1, we commence the process of feature extraction from the input images using Res-Nets, renowned for their prowess in feature representation. The inclusion of pathways spanning from superficial strata to profound layers promotes the seamless propagation of low-level insights. Subsequently, the CEFF module amalgamates features of varying scales to engender fused features, affording more efficient utilization of contextual knowledge while mitigating parameter proliferation. Ultimately, the amalgamated features are fed into the BAC module, which consolidates features along the X and Y axes to individually extract horizontal and vertical attributes. The BAC module enhances the accuracy of boundary positions by predicting deviation values concerning the ground truth boxes. Moreover, the confidence levels of estimated barrels contribute to classification and serve to further amplify performance.

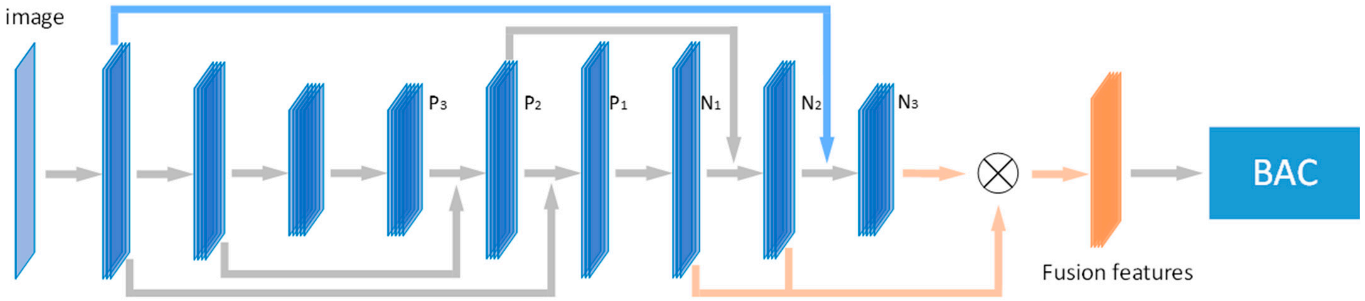

Figure 1. Illustration of our framework.

3.1. Context-Enhanced Feature Fusion

Superficial attributes capture local information and intricacies, with local textures more prone to eliciting responses in neighboring neurons. This profound insight signifies the necessity for the primary network to bolster the propagation of robust semantic features from the shallow to deep pathways, thereby enhancing all features representing classificatory information. Owing to the reality that multiple sampling operations can result in the loss of some localization information, our framework fortifies the localization prowess of the feature hierarchy by disseminating the robust responses of superficial patterns. Thus, we establish a connection pathway from the shallow to deep layers (as depicted by the solid blue line in Figure 1). It comprises fewer than ten layers spanning across these strata.

Our framework initially implements path expansion for sampling. We adhere to the FPN approach to define the generation of various network stages. The spatial size of each feature level remains consistent and corresponds to a distinct stage. ResNet serves as our backbone network, with the feature levels generated by the FPN denoted as $\{P_1, P_2, P_3\}$. Our enhancement path initiates at P_1 and progressively approaches P_3. Throughout this progression from P_1 to P_3, the spatial size decreases by a factor of 2. The newly generated feature mappings corresponding to $\{P_1, P_2, P_3\}$ are represented as $\{N_1, N_2, N_3\}$. It's worth noting that N_1 remains unchanged from P_1 and undergoes no additional processing.

In Figure 2, each connection block combines a feature map N with a feature map P from a deeper level to generate a new feature map, designated as N′. Initially, each feature map undergoes spatial dimension reduction through a 3 × 3 convolutional layer. Subsequently, each element of feature map P is integrated with the downsampled map through lateral connections. The resulting fused feature map then undergoes additional processing via a 3 × 3 convolutional layer to generate the N for the subsequent sub-network.

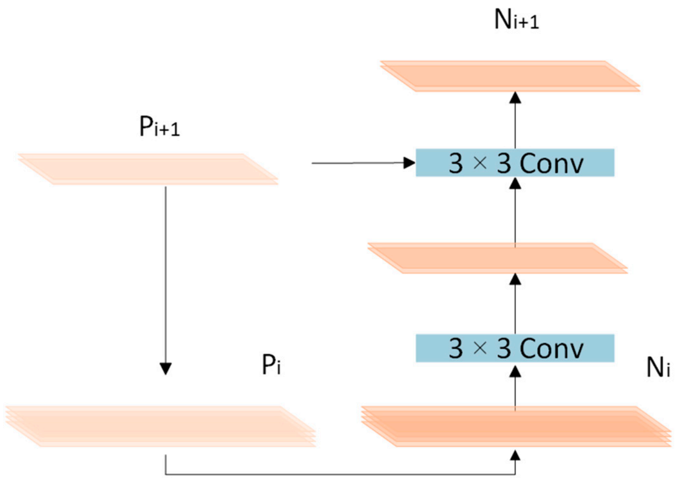

Figure 2. Illustration of our building block of path augmentation.

In contrast to approaches that rely on lateral connections for the integration of multi-level features, our method enhances multi-level features by incorporating semantic features, as depicted in orange in Figure 1. Features at layer l are denoted as N_l, and there are a total of L multi-level features. The indices for the lowest and highest levels are represented as L_{min} and L_{max}, respectively. To integrate multi-level features while maintaining their semantic hierarchy, we initially resize the multi-level features, denoted as $\{N_1,\ N_2,\ N_3\ \}$, to an intermediate size matching the size of N_2. We achieve this through interpolation and max-pooling. Following the feature resizing, we obtain the fused balanced semantic features through a straightforward averaging process:

$$N = \frac{1}{L} \sum_{l=l_{min}}^{l_{max}} N_l \tag{1}$$

3.2. Boundary Aware Convolution

The processing steps of the BAC module are shown in Algorithm 1. As shown in Figure 3, we extract lateral features, denoted as $S_{top},\ S_{down},\ S_{left},\ S_{right}$, through boundary-aware convolution, utilizing ROI features. In line with the YOLO series of detectors, we employ a Feature Pyramid Network (FPN) for extracting Region of Interest (ROI) features across various scales. Subsequently, we transform these features into S using a pair of 3 × 3 convolutional layers. To further enhance our capacity for capturing directional information within regions of interest, we concentrate on amplifying these specific features. More precisely, we employ 1 × 1 convolution to predict two distinct attention maps derived from S, which are subsequently normalized along both the x-axis and y-axis. Using the input attention maps designated as J_x and J_y, we amalgamate the S features to yield S_x and S_y, as delineated below:

$$S_x = \sum_y S(y,:) \times J_x(y,:) \tag{2}$$

$$S_y = \sum_x S(:,x) \times J_y(:,x) \tag{3}$$

Both J_x and J_y represent one-dimensional feature maps with dimensions of $1 \times k$ and $k \times 1$. They are subjected to additional refinement through 1×3 or 3×1 convolutional layers and subsequently upsampled by a factor of 2 using deconvolution layers, yielding $1 \times 2k$ and $2k \times 1$ features in both the horizontal and vertical orientations. Lastly, the upsampled features are directly partitioned into two halves, generating the lateral features, $S_{top},\ S_{down},\ S_{left},\ S_{right}$.

Algorithm 1. Boundary aware convolution

```
Extract object proposals or bounding boxes
for each bounding_box in bounding_boxes:
    Extract features from the bounding box region
    Predict side boundaries using the features
    Refine the bounding box based on side boundaries
    Replace the original bounding box with the refined one
Output the refined bounding boxes
```

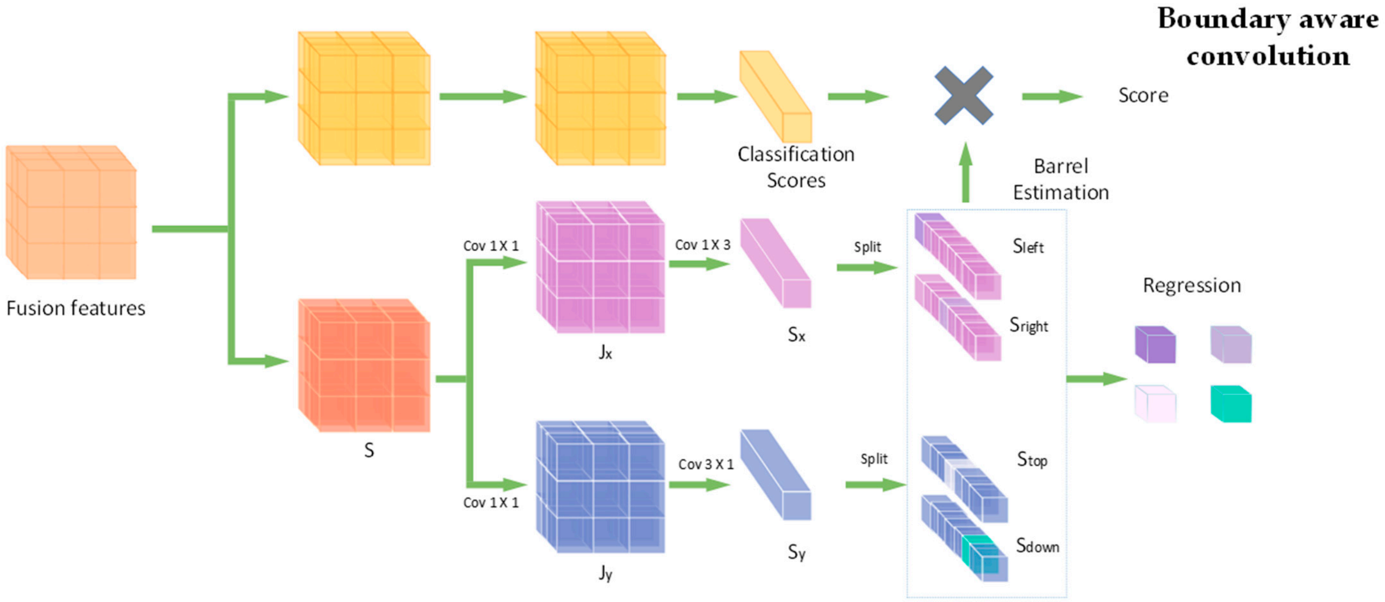

Figure 3. Architecture of the developed BAC.

The positioning process is decomposed into two stages: barrel estimation and fine regression. Candidate regions for each object boundary are partitioned into barrels along the horizontal and vertical directions. Initially, we estimate the barrel in which the boundary is situated, followed by regressing a more precise boundary location within that identified barrel. Throughout the localization process, within the context of a given predefined bounding box (denoted as $B_{top}, \ B_{down}, \ B_{left}, \ B_{right}$), the selected boundary region is magnified by a factor of ε $(\varepsilon > 1)$ to encompass the entirety of the object. This chosen region is then discretized into $2k$ barrels along both the x and y axes, with each boundary associated with k barrels. Consequently, the width of each barrel along the x and y axes is represented as

$$G_x = \frac{\left(\varepsilon B_{right} - \varepsilon B_{left}\right)}{2k} \tag{4}$$

$$G_y = \frac{\left(\varepsilon B_{down} - \varepsilon B_{top}\right)}{2k} \tag{5}$$

In the barrel estimation step, we employ a binary classifier to predict, based on boundary features, whether the boundary resides within the barrel of each side or the nearest barrel. In the fine regression step, regression techniques are employed to predict the offset from the centerline of the selected predicted barrel to the actual boundary label. As illustrated in Figure 4, on each side, the barrel closest to the ground-truth boundary is labeled as 1 (positive sample), while the remaining barrels are labeled as 0 (negative samples). To mitigate ambiguity during the training process, we, for each side, omit the second closest barrel to the ground-truth boundary, as it is challenging to differentiate from the positive barrels. Negative barrels are excluded during the training of the boundary regressor for each side. To enhance the robustness of the fine regression branch, we incorporate both the nearest barrel (marked as "positive" in the barrel estimation step) and the second nearest barrel (marked as "ignored" in the barrel estimation step) for training the regressor. The regression target represents the displacement between the barrel centerline and the corresponding true ground boundary. To alleviate training challenges for the regressor, we normalize the target by G_x and G_y along the respective axes.

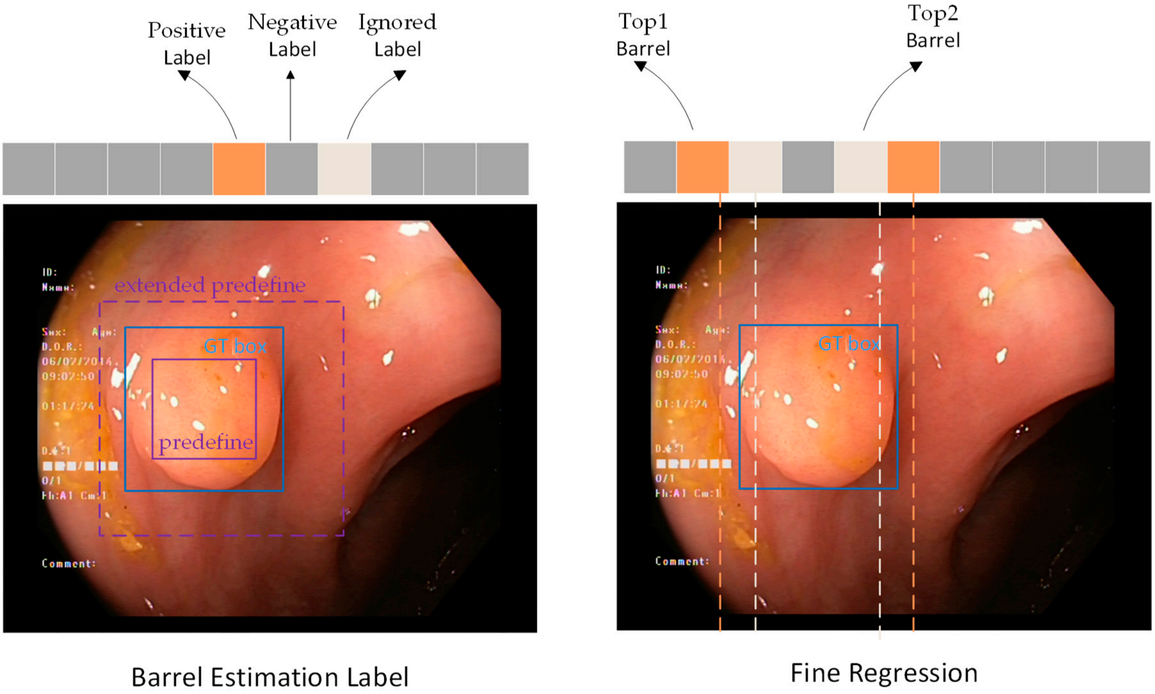

Figure 4. The localization target of BAC for barrel estimation and fine regression on x-axis. The localization target for y-axis can be calculated similarly.

To ensure more precise barrel localization during the localization process, we employ estimated barrel confidence scores for scoring assistance. Consequently, BAC computes the average of the confidence scores for barrel estimation across the four boundaries. The multi-class classification scores are then multiplied by the average localization confidence and employed for candidate sorting during NMS. Scoring plays a crucial role in preserving the best boxes, maintaining high classification confidence, and ensuring accurate localization.

3.3. Loss Design

Object detection is a multi-task learning problem that combines both classification and localization objectives. In practice, network training is typically conducted by manually tuning task weights, such as multiplying the loss during the classification process by a fixed coefficient. The specific formulations of the loss functions are detailed as follows,

$$\mathcal{L} = \mathcal{L}_{cls} + \sigma(\mathcal{L}_{barrel} + \mathcal{L}_{reg}) \tag{6}$$

where $\mathcal{L}_{cls}$ represents the classification loss function. We utilize the barrel estimation loss, denoted as $\mathcal{L}_{barrel}$, and the accurate regression loss, denoted as $\mathcal{L}_{reg}$, in lieu of the localization objective function. The term σ is employed to fine-tune the loss weights for multi-task learning.

In the context of endoscopic images, for conducting multi-task experiments, it is feasible to appropriately adjust the coefficients of the regression task loss functions. However, the impact of object texture and size in endoscopic images can result in significant loss and gradients. We have individually formulated the barrel estimation loss $\mathcal{L}_{barrel}$, and the accurate regression loss $\mathcal{L}_{reg}$. The definition of $\mathcal{L}_{barrel}$ is as follows,

$$\mathcal{L}_{barrel} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \tag{7}$$

where y represents the label of the i-th sample, and $p(y)$ represents the model's prediction for the i-th sample. $\mathcal{L}_{reg}$ is defined as follows,

$$\mathcal{L}_{reg} = \begin{cases} 0.5x^2 \; if \; |x| < 1 \\ |x| - 0.5 \; otherwise \end{cases} \tag{8}$$

where x represents the numerical disparity between the prediction and the ground truth.

The origin of cross-entropy loss is rooted in information theory concepts. Cross-entropy loss quantifies the difference between the probability distribution of true labels and the probability distribution of labels predicted by the model. Minimizing cross-entropy loss involves making the model's predicted probabilities approach the true label probabilities as closely as possible, thereby enhancing the model's classification performance. Its definition is as follows,

$$\mathcal{L}_{cls} = -\sum_{i=1}^{n} y_i log\hat{y}_i \tag{9}$$

In the equation above, y represents the true distribution, $\hat{y}$ represents the network's predicted distribution, and n is the total number of classes.

4. Results

4.1. Dataset

CVC-ClinicDB [47] is a publicly available dataset for colonoscopy polyp detection, consisting of 612 polyp images with a resolution of 384 × 288 pixels. Experts have pro-vided essential information about the polyps using advanced medical annotation tools.

The Kvasir-SEG [48] dataset is extensively utilized for the development and evaluation of methods for colon and rectal polyp detection and segmentation. This dataset comprises 1000 polyp images along with their corresponding bounding boxes and segmentation masks. These images have been meticulously annotated and verified by experienced gastroenterology experts.

EDD2020 [49] is a comprehensive dataset created for the purpose of benchmarking disease detection algorithms in endoscopy examinations. This dataset includes annotations for five distinct disease categories: BE, Suspicious, HGD, Cancer, and Polyp. Bounding box annotations for disease detection are provided within the dataset. The training set consists of a total of 386 endoscopic frames, with each frame annotated for one or multiple diseases. Regions belonging to the same category are combined into a single mask, and bounding boxes for multiple categories are treated as separate boxes located in the same position.

4.2. Performance Metrics

Intersection over Union (IOU) is a metric utilized for quantifying the extent of overlap between two bounding boxes or regions. It is computed by dividing the area of the intersection of the two regions by the area of their union. IOU is frequently employed in object detection tasks to evaluate the precision of object localization. Its definition is as follows,

$$IoU = \frac{A \cap B}{A \cup B} \tag{10}$$

where $\cap$, $\cup$ denote the intersection and union respectively. A is the detection area and B is annotated as GT.

Accuracy is a metric that evaluates the precision of a classification model. It is de-fined as the ratio of true positive predictions to the total number of positive predictions, which includes both true positives (TP) and false positives (FP). Accuracy is employed to assess the model's ability to avoid misclassifying negative samples as positive. Its definition is as follows,

$$Acc = \frac{TP}{TP + FP} \tag{11}$$

Recall, also referred to as sensitivity or the true positive rate, is a metric that quantifies a model's capability to identify all relevant instances within a dataset. It is calculated as the ratio of true positive predictions to the total number of actual positive instances (true positives (TP) + false negatives (FN)). Its definition is as follows,

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

Average Precision (AP) is a metric utilized in object detection to evaluate the precision and recall associated with object localization and classification. It is computed by determining the area under the precision-recall curve. Its definition is as follows,

$$AP = \sum\nolimits_n \{(r_{n+1} - r_n) p_{interp}(r_{n+1})\} \tag{13}$$

with $p_{interp}(r_{n+1}) = \max\limits_{\widetilde{r} \geq r_{n+1}} p\left(\widetilde{r}\right)$. Here, $p(r_n)$ denotes the precision value at a given recall value.

Mean Average Precision (mAP) is a metric utilized to assess the performance of object detection algorithms. It is computed as the average of the Average Precision (AP) for each category in the dataset. AP quantifies how effectively an algorithm ranks objects and assigns scores to each detected object. mAP offers a comprehensive evaluation of the performance of object detection models. Its definition is as follows,

$$mAP = \frac{1}{N} \sum_{i=0}^{n} AP_i \tag{14}$$

The F1 score is a metric that harmoniously combines precision and recall into a single value, offering a balanced assessment of a model's performance. It is especially valuable when the cost of false positives and false negatives is unequal. Its definition is as follows,

$$F1\ Score = 2 \times \frac{Acc \times Recall}{Acc + Recall} \tag{15}$$

4.3. Implementation Details

Our model was implemented using PyTorch on an NVIDIA 3080 GPU card equipped with 32GB of memory. To mitigate the risk of overfitting, we employed various data augmentation techniques, encompassing horizontal flips, vertical flips, and cropping. Notably, we abstained from using any pre-trained weights during the model training process. For the CVC-ClinicDB, Kvasir-SEG, and EDD2020 datasets, we diligently partitioned the data into training, testing, and validation sets following an 8:1:1 ratio. Maintaining uniformity, the input resolution and batch size were consistently set at 320 × 320 and 8, respectively, across all three datasets. During the training phase, we employed the YOLOv5 backbone for iterative training. We optimized our model using the Adam optimizer with an initial learning rate of 0.02. We conducted training on three datasets for 100 epochs.

4.4. Comparison with State-of-the-Art Methods

To facilitate a comparison between the proposed detector and various prior and contemporary methods, we have summarized their components and performance across three datasets. The reported results were either extract-ed from the original papers or obtained from publicly available implementations and models.

In the context of CVC-ClinicDB, it's noteworthy that only a small portion of each CVC-ClinicDB image contains polyps (with an average total polyp area per image of less than 15%). This can lead to a significant number of false negative regions when positive region predictions are hindered by interference. However, the proposed GFFBAC method demonstrates its effectiveness in dealing with the challenges of imbalanced data. As shown in Table 1, the proposed GFFBAC gives the most promising results with mAP_{50} and precision of 94.8% and 93.5% respectively. Notably, a substantial performance gap exists between YOLOv4 and our GFFBAC method, demonstrating the efficacy of GFFBAC in addressing imbalanced data challenges. In comparison to alternative methods, such as [48–52], GFFBAC exhibits noteworthy precision improvements of 13%, 9.9%, 0.9%, 5.2%, and 2.5% respectively. Additionally, our approach surpasses the prior state-of-the-art DC-SSDNe object detection method across all metrics, achieving improvements of 2.6% on mAP_{50}, 2.6% on AP_{50}, 2.5% on precision, 2.5% on recall, and 2.5% on F1. These results not only validate the efficacy of our proposed method but also demonstrate its focus on capturing correlations between instances.

Table 1. Comparison to mainstream methods with GFFBAC on CVC-ClinicDB dataset.

Model	Dataset	mAP_{50}	AP_{50}	P	R	F1
YOLOv4 [50]	CVC-ClinicDB	-	-	80.5 ± 0.3	73.6 ± 0.1	76.9 ± 0.1
STYOLOv5 [51]	CVC-ClinicDB	-	-	83.6 ± 0.3	73.1 ± 0.2	78 ± 0.1
ITH [52]	CVC-ClinicDB	-	-	92.6 ± 0.2	80.7 ± 0.1	86.2 ± 0.1
soet [53]	CVC-ClinicDB	89.5 ± 0.1	89.5 ± 0.1	88.3 ± 0.1	92.3 ± 0.1	89.8 ± 0.2
DC-SSDNet [54]	CVC-ClinicDB	92.2 ± 0.3	92.2 ± 0.3	91 ± 0.1	92.2 ± 0.1	88.4 ± 0.2
Ours	CVC-ClinicDB	94.8 ± 0.1	94.8 ± 0.1	93.5 ± 0.2	92.7 ± 0.1	90.9 ± 0.1

The Kvasir-SEG dataset contains images with relatively larger polyp areas, with an average total tumor area exceeding 70%. This setting is conducive to the outstanding performance of BAC. Corresponding results are shown in Table 2. GFFBAC surpasses all competing methods, achieving an increase of 0.9% in mAP and 0.4% in accuracy compared to the second-best method. A significant advantage of GFFBAC lies in its ability to capture global contextual information, which YOLO-based methods struggle to achieve. The experimental results affirm the superiority of boundary-aware feature convolution in object detection.

Table 2. Comparison to mainstream methods with GFFBAC on Kvasir-SEG dataset.

Model	Dataset	mAP_{50}	AP_{50}	P	R	F1
YOLOv4 [55]	Kvasir-SEG	71.0 ± 0.1	71.0 ± 0.1	65.0 ± 0.2	66.0 ± 0.2	63.0 ± 0.1
YOLOv5l [55]	Kvasir-SEG	81.0 ± 0.1	68.0 ± 0.1	65.0 ± 0.2	65.0 ± 0.1	64.0 ± 0.1
YOLOv5m [55]	Kvasir-SEG	81.0 ± 0.2	80.0 ± 0.2	65.0 ± 0.0	65.0 ± 0.1	64.0 ± 0.3
YOLOv5n [55]	Kvasir-SEG	75.0 ± 0.0	75.0 ± 0.0	64.0 ± 0.1	64.0 ± 0.1	62.0 ± 0.2
YOLOv5s [55]	Kvasir-SEG	74.0 ± 0.1	74.0 ± 0.1	63.0 ± 0.2	62.0 ± 0.1	61.0 ± 0.1
DETR [55]	Kvasir-SEG	80.0 ± 0.3	80.0 ± 0.3	65.0 ± 0.1	69.0 ± 0.2	66.0 ± 0.1
soet	Kvasir-SEG	92.6 ± 0.1	92.6 ± 0.1	95.1 ± 0.1	93.1 ± 0.1	94.0 ± 0.2
Ours	Kvasir-SEG	93.5 ± 0.1	93.5 ± 0.1	95.5 ± 0.2	93.2 ± 0.1	94.7 ± 0.1

In the EDD2020 dataset, we conducted a comparative analysis involving our model, the EDD2020 Detection Challenge team, and YOLO series models in Table 3. Our model exhibits exceptional performance, achieving a 44.1% mAP_{50} using BAC, outperforming YOLOv5 by 1.4% mAP_{50}. By leveraging the more potent feature-enhanced fusion CEFF, GFFBAC attains an overall mAP of 46.5%. As illustrated in Table 4, GFFBAC demonstrates strong performance across all categories, with particular excellence in detecting polyps and related classes.

Table 3. Comparison to mainstream methods with GFFBAC on EDD2020 dataset.

Team Names	Dataset	mAP_{25}	mAP_{50}	mAP_{75}	Overall mAP
sahadate [56]	EDD2020	37.6 ± 0.1	23.3 ± 0.1	15.8 ± 0.1	26.8 ± 0.1
VinBDI [56]	EDD2020	43.2 ± 0.1	27.0 ± 0.1	17.0 ± 0.1	30.2 ± 0.1
adrian [56]	EDD2020	48.3 ± 0.1	33.6 ± 0.1	27.1 ± 0.2	37.6 ± 0.1
YOLOv4	EDD2020	53.1 ± 0.1	41.2 ± 0.1	32.3 ± 0.2	42.2 ± 0.2
YOLOv5	EDD2020	54.7 ± 0.1	42.7 ± 0.1	32.9 ± 0.1	43.4 ± 0.3
Ours	EDD2020	59.7 ± 0.1	44.1 ± 0.1	35.6 ± 0.2	46.5 ± 0.1

Table 4. Per class evaluation results with GFFBAC for the detection task of the EDD2020 dataset.

Class	Dataset	mAP_{50}	$mAP_{50\text{-}95}$	Precision	Recall
BE	EDD2020	66.3 ± 0.2	48.4 ± 0.1	59.4 ± 0.2	68.2 ± 0.1
suspicious	EDD2020	25.5 ± 0.1	16.9 ± 0.2	35.6 ± 0.3	21.9 ± 0.1
HGD	EDD2020	35.3 ± 0.2	22.9 ± 0.2	47.4 ± 0.2	27.8 ± 0.2
cancer	EDD2020	34.8 ± 0.2	16.3 ± 0.2	64.0 ± 0.1	25.0 ± 0.1
polyp	EDD2020	58.4 ± 0.3	40.2 ± 0.1	62.9 ± 0.2	57.7 ± 0.2

Qualitative results on the CVC-ClinicDB dataset, Kvasir-SEG dataset, and EDD2020 dataset are showcased in Figure 5. These results only display bounding boxes with an IoU greater than 0.5, and they underscore the precision of the proposed GFFBAC method.

4.5. Ablation Study

As presented in Table 5, to further assess the significance of each proposal com-ponent, we conducted a series of ablation studies on the CVC-ClinicDB dataset. Whether balanced feature fusion is applied or not, augmenting the path enhancement from shallow to deep levels consistently enhances both AP_{50} and model precision, surpassing 1.7% and 0.3%, respectively. This affirms the utility of information from lower-level features. Balanced feature fusion consistently enhances performance, with or without the path enhancement from shallow to deep levels. The results exhibited consistent improvements across small, medium, and large scales, affirming that balanced semantic features balanced low-level and high-level information at each level and generated consistent enhancements. Our observation indicates the utility of features from various layers in the final prediction, affirming the effectiveness of CEFF in overall performance improvement at all scales. As illustrated in Figure 6, we applied BAC to bring the detection boxes closer to reality. This approach elevated the performance from 89.7% to 94.1%, under-scoring the significance of each boundary feature in object detection. The inclusion of all these components in GFFBAC results in a 5.1% improvement in mAP_{50} compared to the baseline. The notable enhancement in detection precision primarily stems from the contributions of BAC, emphasizing content boundaries, and CEFF, dedicated to enhancing semantic features.

Table 5. The effects of each module in our design. BAC, BSF, PA denote Boundary-aware Convolution, balanced semantic features and path augmentation, respectively.

BAC	BSF	PA	mAP_{50}	AP_{50}	P	R	F1
			89.7 ± 0.2	89.7 ± 0.2	90.5 ± 0.2	86.3 ± 0.2	87.2 ± 0.2
	√		90.9 ± 0.2	90.9 ± 0.2	91.4 ± 0.2	87.2 ± 0.2	87.9 ± 0.2
		√	91.4 ± 0.2	91.4 ± 0.2	90.8 ± 0.2	87.4 ± 0.2	87.6 ± 0.2
	√	√	92.4 ± 0.2	92.4 ± 0.2	91.7 ± 0.2	88.2 ± 0.2	88.6 ± 0.2
√			94.1 ± 0.2	94.1 ± 0.2	92.6 ± 0.2	92.2 ± 0.2	89.4 ± 0.2
√	√		94.3 ± 0.2	94.3 ± 0.2	93.4 ± 0.2	92.7 ± 0.2	90.2 ± 0.2
√	√	√	94.8 ± 0.1	94.8 ± 0.1	93.5 ± 0.2	92.7 ± 0.1	90.9 ± 0.1

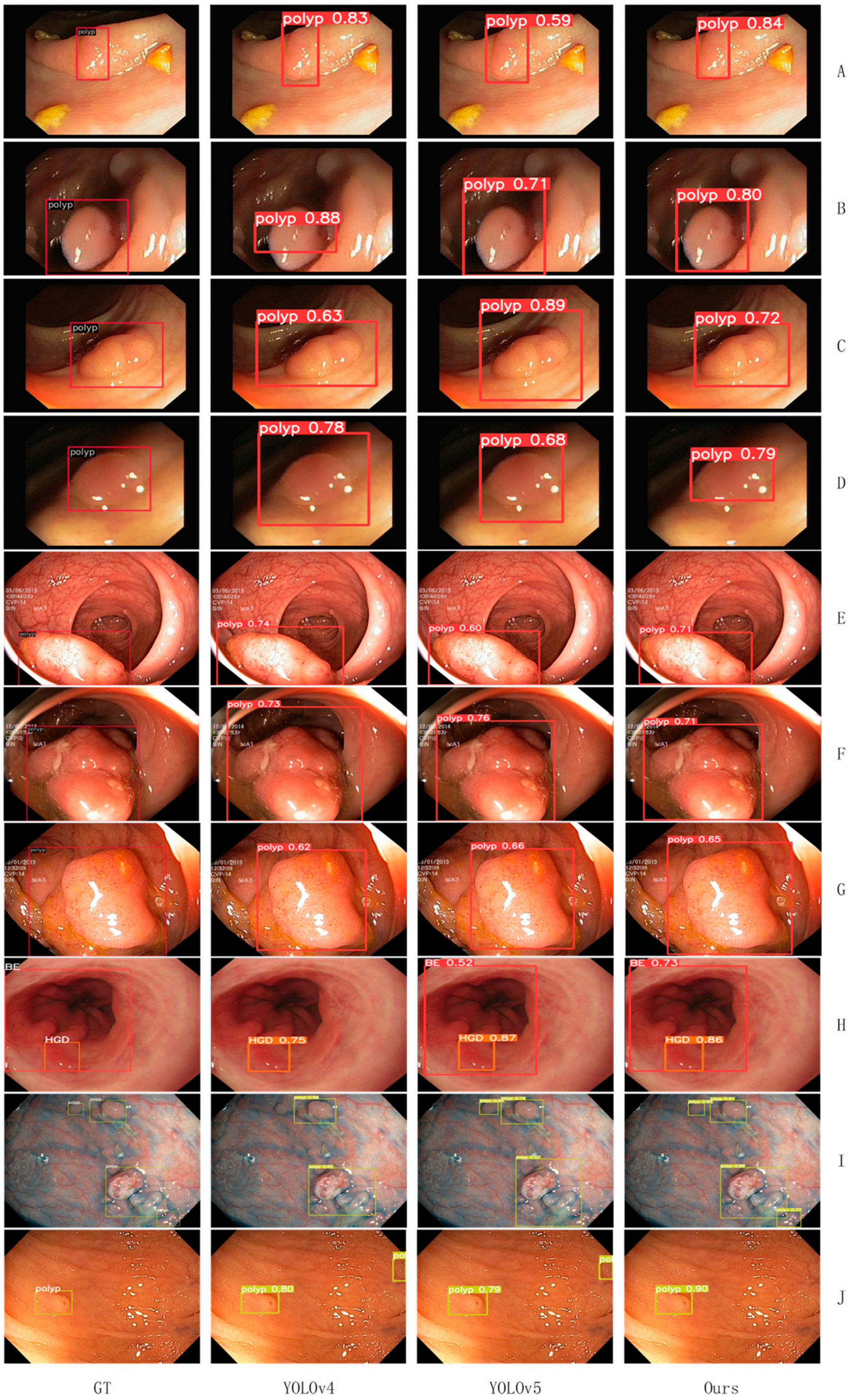

Figure 5. Comparison of object detection results on CVC-ClinicDB (**A**–**D**), Kvasir-SEG (**E**–**G**) and EDD2020 (**H**–**J**), respectively.

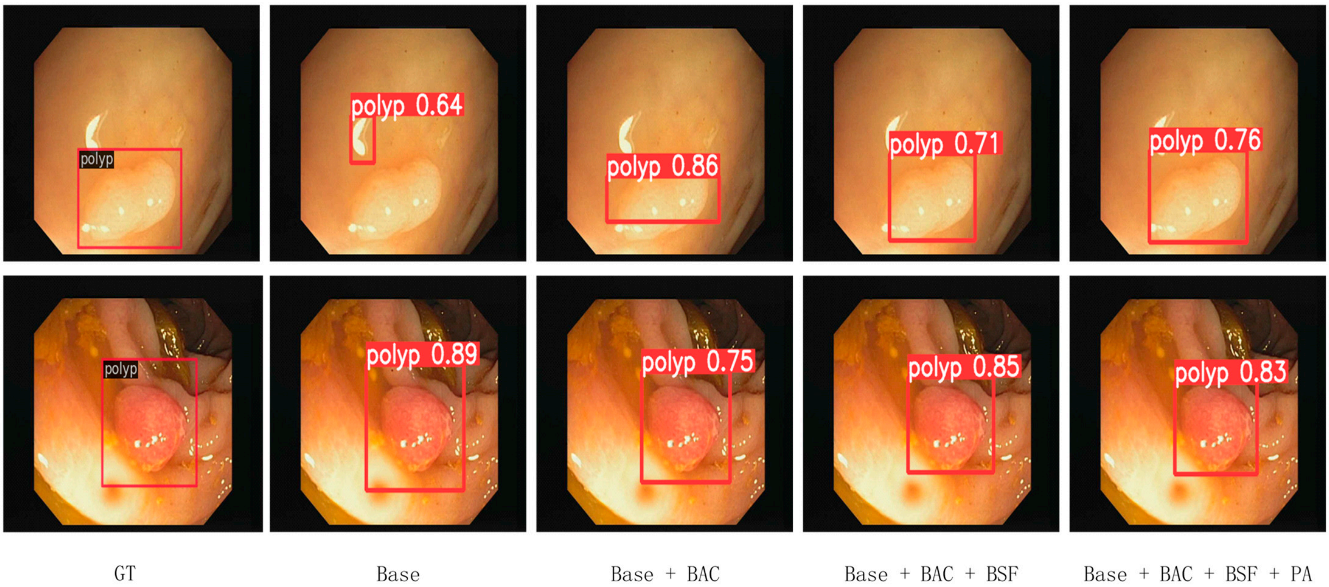

Figure 6. Qualitative results comparison between BAC, BSF and PA on the CVC-ClinicDB dataset.

5. Discussion

Deep learning is rapidly gaining traction in applications such as computer-aided detection and disease diagnosis within complex clinical settings, including endoscopy examinations. In comparison to traditional machine learning, deep learning offers numerous advantages due to its multi-layered architecture, enabling training on raw data and the acquisition of abstract features. The capacity of neural networks to derive meaningful representations from data is a fundamental aspect of deep learning. Medical image processing aids in identifying and extracting features that may elude hu-man perception. Hence, deep learning excels in object detection tasks. Moreover, data augmentation techniques and transfer learning methods can enhance the generalization of object detection models to novel data and scenarios. The swift advancement of hardware components, such as GPUs, has rendered the training of large-scale deep learning models more accessible, expediting the model development and optimization process. Through continual technological innovations, we are empowered to achieve superior models.

However, in clinical practice, endoscopic images often present complex back-grounds and exhibit significantly higher data variability compared to natural scenes. This necessitates a closer examination of image details, textures, and structures. Additionally, lesion or polyp regions in endoscopic images tend to display varying scales and shapes across different patients, making smaller targets particularly susceptible to being overlooked. Moreover, endoscopic images often contain sensitive patient information, and the acquisition of large-scale annotated endoscopic image datasets is a costly and time-consuming endeavor. Consequently, privacy concerns and labeling challenges can result in insufficient data, potentially affecting the performance of deep learning models. Up to this point, we have taken these challenges into account. We propose a shallow-to-deep approach that focuses on the boundaries of lesion regions and improves classification by balancing semantic features. In the EDD dataset, a major challenge is class confusion, particularly among the suspicious, HGD, and cancer categories. The model achieves accuracies of 59.4%, 35.6%, 47.4%, 64.0%, and 62.9% for BE, Suspicious, HGD, Cancer, and Polyp, respectively, with an mAP0.5 1.4 higher than YOLOv5. As depicted in Figure 5, the proposed model yields detection boxes closer to the labels. This is partly attributed to balanced semantic features, but more significantly, BAC focuses on the boundaries of lesion regions and introduces confidence scores to facilitate the selection of better-fitting detection boxes.

The strengths of our model render it especially well-suited for the following scenarios: (1) addressing the significant challenge of missed detections resulting from small polyps,

(2) emphasizing the need for robust multi-class performance, and (3) fulfilling the demand for intuitive interpretability. Nevertheless, the model inherently comes with its set of constraints: firstly, although it exhibits commendable performance in polyp recognition, there exists considerable scope for improvement. Secondly, the model's performance enhancement relies heavily on the availability of high-quality and extensive datasets, underlining the significance of the dataset and data augmentation techniques.

6. Conclusions

We introduce GFFBAC for object detection in endoscopic examinations. To enhance information propagation within the representative pipeline, we design the CEFF module. This module aggregates features from all layers and reduces the gap between shallow and deep feature layers, thereby promoting reliable information transmission. Additionally, we employ the same-depth integration to bolster balanced semantic features across multiple levels. Furthermore, we introduce BAC to improve the classification and localization abilities of detection boxes. We utilize boundary features to focus on content boundaries for precise localization and introduce a confidence score to maintain high-quality detection boxes. GFFBAC yields substantial improvements on challenging datasets, including CVC-ClinicDB, Kvasir-SEG, and EDD2020. Comprehensive experiments demonstrate GFFBAC's competitive accuracy in assisting medical professionals with diagnostic tasks. Our future work will be to extend our method to video and other fields.

Author Contributions: Conceptualization, W.F. and J.Y.; methodology, W.F. and J.Y.; software, W.F.; validation, Z.J., J.Y. and W.F.; formal analysis, W.F.; investigation, J.Y.; resources, Z.J.; data curation, W.F.; writing—original draft preparation, W.F.; writing—review and editing, W.F. and J.Y.; supervision, Z.J.; project administration, J.Y.; funding acquisition, Z.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant No. 52075530), the Zhejiang Provincial Natural Science Foundation of China (LQ23F030001), Hangzhou Innovation Team under No. TD2022011, and the AiBle project co-financed by the European Regional Development Fund.

Data Availability Statement: Data is contained within the article.

Acknowledgments: The authors would like to acknowledge the support from the National Natural Science Foundation of China (Grant No. 52075530), the Zhejiang Provincial Natural Science Foundation of China (LQ23F030001), Hangzhou Innovation Team under No. TD2022011, and the AiBle project co-financed by the European Regional Development Fund.

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

References

1. Min, J.K.; Kwak, M.S.; Cha, J.M. Overview of deep learning in gastrointestinal endoscopy. *Gut Liver* **2019**, *13*, 388. [CrossRef] [PubMed]
2. Jain, S.; Seal, A.; Ojha, A.; Yazidi, A.; Bures, J.; Tacheci, I.; Krejcar, O. A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images. *Comput. Biol. Med.* **2021**, *137*, 104789. [CrossRef] [PubMed]
3. Hashimoto, R.; Requa, J.; Dao, T.; Ninh, A.; Tran, E.; Mai, D.; Lugo, M.; Chehade, N.E.H.; Chang, K.J.; Karnes, W.E.; et al. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). *Gastrointest. Endosc.* **2020**, *91*, 1264–1271.e1. [CrossRef]
4. Li, K.; Boyd, P.; Zhou, Y.; Ju, Z.; Liu, H. Electrotactile feedback in a virtual hand rehabilitation platform: Evaluation and implementation. *IEEE Trans. Autom. Sci. Eng.* **2018**, *16*, 1556–1565. [CrossRef]
5. Liu, H.; Ju, Z.; Ji, X.; Chan, C.S.; Khoury, M. *Human Motion Sensing and Recognition*; Springer: Berlin, Germany, 2017.
6. Yu, J.; Gao, H.; Chen, Y.; Zhou, D.; Liu, J.; Ju, Z. Deep object detector with attentional spatiotemporal LSTM for space human–robot interaction. *IEEE Trans. Hum. Mach. Syst.* **2022**, *52*, 784–793. [CrossRef]
7. Montero-Valverde, J.A.; Organista-Vázquez, V.D.; Martínez-Arroyo, M.; de la Cruz-Gámez, E.; HernándezHernández, J.L.; Hernández-Bravo, J.M.; Hernández-Hernández, M. Automatic Detection of Melanoma in Human Skin Lesions. In *Proceedings of the International Conference on Technologies and Innovation*; Guayaquil, Ecuador, 13–16 November 2023, Springer Nature: Cham, Switzerland, 2023; pp. 220–234.

8. Sarda, A.; Dixit, S.; Bhan, A. Object detection for autonomous driving using yolo [you only look once] algorithm. In *Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*; Tirunelveli, India, 4–6 February 2021, IEEE: Piscataway, NJ, USA, 2021; pp. 1370–1374.
9. George, J.; Skaria, S.; Varun, V.V. Using YOLO based deep learning network for real time detection and localization of lung nodules from low dose CT scans. In *Medical Imaging 2018: Computer-Aided Diagnosis*; SPIE: Bellingham, DC, USA, 2018; Volume 10575, pp. 347–355.
10. Mirzaei, B.; Nezamabadi-Pour, H.; Raoof, A.; Derakhshani, R. Small Object Detection and Tracking: A Comprehensive Review. *Sensors* **2023**, *23*, 6887. [CrossRef]
11. Simony, M.; Milzy, S.; Amendey, K.; Gross, H.M. Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
12. Poon, Y.S.; Lin, C.C.; Liu, Y.H.; Fan, C.P. YOLO-based deep learning design for in-cabin monitoring system with fisheye-lens camera. In Proceedings of the 2022 IEEE International Conference on Consumer Electronics (ICCE), Virtual, 7–9 January 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–4.
13. Pathak, A.R.; Pandey, M.; Rautaray, S. Application of deep learning for object detection. *Procedia Comput. Sci.* **2018**, *132*, 1706–1717. [CrossRef]
14. Bharati, S.P.; Wu, Y.; Sui, Y.; Padgett, C.; Wang, G. Real-time obstacle detection and tracking for sense-and-avoid mechanism in UAVs. *IEEE Trans. Intell. Veh.* **2018**, *3*, 185–197. [CrossRef]
15. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 July 2015; pp. 3431–3440.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
17. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
18. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1520–1528.
19. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–16 July 2017; pp. 2117–2125.
20. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 354–370.
21. Kong, T.; Yao, A.; Chen, Y.; Sun, F. Hypernet: Towards accurate region proposal generation and joint object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 845–853.
22. Ghiasi, G.; Fowlkes, C.C. Laplacian pyramid reconstruction and refinement for semantic segmentation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 519–534.
23. 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, 27–30 June 2016; pp. 779–788.
24. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
25. Simon, M.; Amende, K.; Kraus, A.; Honer, J.; Samann, T.; Kaulbersch, H.; Milz, S.; Michael Gross, H. Complexer-yolo: Real-time 3d object detection and tracking on semantic point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–20 June 2019.
26. Han, X.; Chang, J.; Wang, K. Real-time object detection based on YOLO-v2 for tiny vehicle object. *Procedia Comput. Sci.* **2021**, *183*, 61–72. [CrossRef]
27. Chen, W.; Huang, H.; Peng, S.; Zhou, C.; Zhang, C. YOLO-face: A real-time face detector. *Vis. Comput.* **2021**, *37*, 805–813. [CrossRef]
28. Jang, J.Y. The past, present, and future of image-enhanced endoscopy. *Clin. Endosc.* **2015**, *48*, 466–475. [CrossRef] [PubMed]
29. Banerjee, S.; Cash, B.D.; Dominitz, J.A.; Baron, T.H.; Anderson, M.A.; Ben-Menachem, T.; Fisher, L.; Fukami, N.; Harrison, M.E.; Ikenberry, S.O.; et al. The role of endoscopy in the management of patients with peptic ulcer disease. *Gastrointest. Endosc.* **2010**, *71*, 663–668. [CrossRef] [PubMed]
30. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A survey of modern deep learning based object detection models. *Digit. Signal Process.* **2022**, *126*, 103514. [CrossRef]
31. Yu, J.; Ma, T.; Chen, H.; Lai, M.; Ju, Z.; Xu, Y. Marrying Global–Local Spatial Context for Image Patches in Computer-Aided Assessment. *IEEE Trans. Syst. Man Cybern. Syst.* **2023**, *53*, 7099–7111. [CrossRef]
32. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*. [CrossRef]
33. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

34. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
35. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; IEEE: Piscataway, NJ, USA, 2006; Volume 2, pp. 2169–2178.
36. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In Proceedings of the Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 143–156.
37. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
38. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. Available online: https://proceedings.neurips.cc/paper_files/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html (accessed on 15 November 2023). [CrossRef] [PubMed]
39. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
40. Chen, S.; Urban, G.; Baldi, P. Weakly Supervised Polyp Segmentation in Colonoscopy Images Using Deep Neural Networks. *J. Imaging* **2022**, *8*, 121. [CrossRef]
41. Fan, W.; Ma, T.; Gao, H.; Yu, J.; Ju, Z. Deep Learning-Powered Multiple-Object Segmentation for Computer-Aided Diagnosis. In Proceedings of the 2023 42nd Chinese Control Conference (CCC), Tianjin, China, 24–26 July 2023; pp. 7895–7900.
42. Yu, J.; Ma, T.; Fu, Y.; Chen, H.; Lai, M.; Zhuo, C.; Xu, Y. Local-to-global spatial learning for whole-slide image representation and classification. *Comput. Med. Imaging Graph.* **2023**, *107*, 102230. [CrossRef] [PubMed]
43. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
44. Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; Jiang, Y. Acquisition of localization confidence for accurate object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 784–799.
45. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region proposal by guided anchoring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2965–2974.
46. Yu, J.; Gao, H.; Zhou, D.; Liu, J.; Gao, Q.; Ju, Z. Deep temporal model-based identity-aware hand detection for space human–robot interaction. *IEEE Trans. Cybern.* **2021**, *52*, 13738–13751. [CrossRef] [PubMed]
47. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. [CrossRef]
48. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-seg: A segmented polyp dataset. In Proceedings of the MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, Republic of Korea, 5–8 January 2020; pp. 451–462.
49. Ali, S.; Ghatwary, N.; Braden, B.; Lamarque, D.; Bailey, A.; Realdon, S.; Cannizzaro, R.; Rittscher, J.; Daul, C.; East, J. Endoscopy disease detection challenge 2020. *arXiv* **2020**, arXiv:2003.03376.
50. Carrinho, P.; Falcao, G. Highly Accurate and Fast YOLOv4-Based Polyp Detection. Available at SSRN 4227573. 2022. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4227573 (accessed on 15 November 2023).
51. Ma, C.; Jiang, H.; Ma, L.; Chang, Y. A Real-Time Polyp Detection Framework for Colonoscopy Video. In Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Shenzhen, China, 4–7 November 2022; Springer International Publishing: Cham, Switzerland, 2022; pp. 267–278.
52. Yu, T.; Lin, N.; Zhang, X.; Pan, Y.; Hu, H.; Zheng, W.; Liu, J.; Hu, W.; Duan, H.; Si, J. An end-to-end tracking method for polyp detectors in colonoscopy videos. *Artif. Intell. Med.* **2022**, *131*, 102363. [CrossRef] [PubMed]
53. Lima, A.C.D.M.; De Paiva, L.F.; Bráz, G.; De Almeida, J.D.S.; Silva, A.C.; Coimbra, M.T.; De Paiva, A.C. A two-stage method for polyp detection in colonoscopy images based on saliency object extraction and transformers. *IEEE Access* **2023**, *11*, 2169–3536. [CrossRef]
54. Souaidi, M.; Lafraxo, S.; Kerkaou, Z.; El Ansari, M.; Koutti, L. A Multiscale Polyp Detection Approach for GI Tract Images Based on Improved DenseNet and Single-Shot Multibox Detector. *Diagnostics* **2023**, *13*, 733. [CrossRef]
55. Neto, A.; Couto, D.; Coimbra, M.; Cunha, A. Colonoscopic Polyp Detection with Deep Learning Assist. In Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023), Virtual, 8–10 February 2023.
56. Ali, S.; Dmitrieva, M.; Ghatwary, N.; Bano, S.; Polat, G.; Temizel, A.; Krenzer, A.; Hekalo, A.; Guo, Y.B.; Matuszewski, B.; et al. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Med. Image Anal.* **2021**, *70*, 102002. [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.