Vol. 11, No. 2, Oktober 2026, Page. 521-538
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

An Empirical Analysis of Anchor Optimization and RPN Sampling in Faster R-CNN for Pleural Effusion Detection

Muhammad Jafar Sakone[1], Izzati Muhimmah[*2]

[1,2]Magister Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia, Indonesia
Email: [1]mjafar@students.uii.ac.id, [2]izzati@uii.ac.id

Abstract

Pleural effusion detection from chest X-ray images remains challenging due to variations in lesion scale and severe foreground-background imbalance within two-stage object detectors. This study empirically investigates the effects of anchor representation and Region Proposal Network (RPN) sampling on the detection process of Faster R-CNN for pleural effusion detection. A Faster R-CNN with ResNet-50 and Feature Pyramid Network (FPN) was evaluated using representative anchors derived from ground-truth object characteristics and adjusted RPN Intersection over Union (IoU) assignment thresholds. The analysis examined anchor representation, RPN sample distribution, proposal quality, detection confidence, and stage-wise detection outcomes. Representative anchors improved geometric correspondence with the target objects, while RPN threshold adjustment substantially increased positive RPN samples from approximately 4 to 110 per image. Proposal analysis showed high spatial accessibility, with proposal IoU coverage at IoU ≥ 0.50 reaching up to 99.35% across the evaluated configurations. However, stage-wise analysis showed that classification was the dominant source of detection failure, accounting for 15.03–16.99% of the evaluated objects, while proposal and localization failures remained comparatively limited. These findings demonstrate that optimization in anchor representation and RPN sampling can enhance upstream proposal characteristics, but such improvements do not necessarily translate proportionally into final detection performance.

Keywords: anchors optimization, faster R-CNN, object detection, pleural effusion, RPN sampling.

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)

1. INTRODUCTION

Pleural effusion is a common thoracic abnormality [1] characterized by the accumulation of fluid in the pleural space and is routinely assessed using chest X-ray (CXR) imaging [2], [3], [4]. The widespread use and accessibility of CXR have also encouraged the development of artificial intelligence (AI)-based approaches to support the detection of thoracic abnormalities, including pleural effusion [5], [6], [7]. While image-level classification can determine the presence of an abnormality [8], object detection provides additional spatial information by localizing the suspected region using bounding boxes [9], [10]. Among two-stage object detectors, Faster R-CNN employs a Region Proposal Network (RPN) to generate candidate object regions that are subsequently refined and classified by the detection stage [5]. Therefore, the final detection outcome depends not only on the classification process but also on the ability of preceding components to generate proposals that adequately represent the target objects [11].

In anchor-based detectors such as Faster R-CNN, predefined anchors serve as spatial references with different scales and aspect ratios for matching ground-truth objects [12]. This matching is governed by Intersection over Union (IoU)-based assignment criteria, which determine the positive, negative, or ignored labels used to train the Region Proposal Network (RPN). Consequently, anchor configuration and assignment thresholds influence the distribution of training samples and the generation of candidate object regions. Previous studies have addressed different limitations of conventional anchor-based detection. Anchor Box Optimization [9] and Adaptive Anchor Box Optimization [10] investigate how anchor configurations can be better aligned with object characteristics, while Guided Anchoring adapts anchor locations and shapes to image content [13]. Other approaches target proposal generation and sample assignment: IoU-Uniform R-CNN [14] addresses the distribution of training samples across IoU levels, whereas Cascade RPN [15] progressively refines region proposals. These approaches address

complementary aspects of detection, indicating that anchor representation, RPN sample assignment, and proposal quality should be considered as related but distinct factors.

In medical imaging, object-detection methods have been investigated for tasks such as pulmonary nodule detection, pneumonia detection, and the identification and localization of multiple abnormalities in chest X-ray images. Zhao et al. [16] combined multiscale feature fusion with a separate 3D CNN-based candidate-filtering stage for pulmonary nodule detection in CT images and used repeated sampling of positive nodules to address the imbalance between positive and negative samples. For pneumonia detection in chest X-ray images, Yao et al. [17] employed K-Means++ to derive dataset-specific anchor sizes and incorporated Soft-NMS into an improved Faster R-CNN framework to refine detection results. In contrast, Fan et al. [18] developed an anchor-free YOLOX-based framework for localizing multiple abnormalities in chest X-ray images and reported an imbalance in the distribution of ground-truth abnormality annotations.

These studies address distinct aspects of medical object detection, including sample imbalance, anchor adaptation, candidate filtering, and abnormality localization, but do not directly establish how changes in anchor representation and RPN sampling thresholds affect successive detection stages in pleural effusion detection. Related work has also examined the role of negative samples in medical object detection, including polyp detection in colonoscopy images [19], while research on detector failures has investigated factors associated with missed detections in general object-detection systems [20]. Although these studies provide important evidence on medical detection applications, training sample composition, and detection failures, they address different tasks and methodological questions. Final performance measures such as precision, recall, and average precision remain essential for evaluating detection systems [21], but they do not, on their own, identify how changes to anchor representation and RPN sampling affect intermediate detection outcomes. However, it remains unclear whether changes in anchor representation and RPN sampling consistently improve intermediate detection outcomes and, more importantly, whether these improvements translate into better final detection performance for pleural effusion on chest X-ray images. A stage-wise analysis can address this question by examining proposal coverage, localization, and classification alongside final detection performance.

This study presents an empirical analysis of anchor optimization and RPN sampling in Faster R-CNN for pleural effusion detection from chest X-ray images. First, representative anchor configurations are evaluated based on their correspondence with ground-truth bounding-box characteristics. Second, the effects of adjusting RPN IoU assignment thresholds on foreground–background sample distribution and proposal generation are examined. Third, stage-wise detection outcomes are analyzed to distinguish proposal related limitations from localization and classification failures. Finally, the four configurations: baseline, baseline with adjusted RPN sampling, representative anchors, and representative anchors with adjusted RPN sampling are compared using standard COCO evaluation metrics across five validation folds. Rather than introducing a new detection architecture, this study investigates whether changes to anchor representation and RPN sampling are consistently associated with improvements across the detection pipeline and in final detection performance.

2. METHODS

2.1 Dataset and Experimental Framework

This study used chest X-ray (CXR) images from NIH CXR14 [22] to investigate pleural effusion detection using an anchor-based two-stage object detector. The experimental dataset comprised 306 images, including 153 images with pleural effusion bounding-box annotations and 153 images labeled as No Finding. The annotated positive images were used to train and evaluate the object detector, while the No Finding images were included to assess false-positive predictions in the absence of the target finding. Model evaluation employed five-fold cross-validation to reduce dependence on a single training and validation partition.

The detection framework was based on Faster R-CNN [12] from Detectron2 [23] with a ResNet-50 backbone [24] and Feature Pyramid Network (FPN) [25]. The backbone extracted hierarchical image features, which were aggregated by the FPN to provide multi-scale representations. The Region Proposal Network (RPN) generated candidate regions from these features, and the region-of-interest (RoI)

detection head subsequently refined the candidate bounding boxes and classified them as pleural effusion or background. The detector was configured for a single target class.

Four experimental configurations were designed to investigate the effects of anchor representation and RPN sampling thresholds. The first configuration used the baseline anchor sizes and aspect ratios with the baseline RPN sampling. The second used representative anchors derived from the annotated bounding-box dimensions while retaining the baseline RPN sampling. The third retained the baseline anchors but used adjusted RPN sampling, and the fourth combined representative anchors with the adjusted RPN sampling. The baseline RPN sampling assignment IoU thresholds were 0.3 and 0.7, whereas the adjusted thresholds were 0.3 and 0.5. This design enabled the isolated effects of anchor representation and RPN sampling to be examined through controlled pairwise comparisons.

The model used a base learning rate of 0.02 as default from Detectron2, a batch size of 16 images per iteration, the WarmupCosineLR learning-rate scheduler, a maximum of 2,000 training iterations, and 200 warm-up iterations. The ResNet-50 backbone used res2, res3, res4, and res5 feature stages, with the first two stages frozen. The FPN used these four backbone stages as inputs. The RPN processed the p2–p6 feature levels, used a per-image anchor sampling budget of 256, and employed a positive-anchor fraction of 0.5. The RoI detection head used a per-image sampling budget of 512 proposals and a positive fraction of 0.25. The RoI matching IoU threshold was set to 0.5.

The RPN proposal limits were configured as 2,000 pre-NMS and 1,000 post-NMS proposals per feature level during training, and 1,000 pre-NMS and 100 post-NMS proposals during inference, according to the experimental configuration. The RPN NMS threshold was left at its Detectron2 default setting. These proposal limits were used in the small-dataset training setup and were distinguished from the IoU thresholds used for anchor assignment. Image preprocessing, pretrained weights, optimizer settings, and random seeds must be reported according to the exact settings used in the experiments to ensure reproducibility.

2.2 Anchor Representation

Anchor representation was analyzed by examining the geometric characteristics of ground-truth bounding boxes and their correspondence with predefined anchors. For each ground-truth object g_i, the best-matching anchor was determined based on the maximum Intersection over Union (IoU) value among the available anchors:

$$IoU_i^{best} = max_j IoU(g_i, a_j) \qquad (1)$$

Where a_j denotes the *j*-th anchor. The quality of anchor representation was evaluated using the average best-matching IoU across all ground-truth objects and the proportion of objects covered by an anchor with an IoU equal to or greater than a predefined threshold. The resulting object-scale characteristics were used to derive representative anchor sizes and aspect ratios to compare their representation quality with the default anchor configuration.

Representative anchors were derived using an IoU-based K-Means clustering procedure applied to the width and height of the ground-truth bounding boxes. Nine initial cluster centers were randomly selected from the observed bounding-box dimensions without replacement, using a random seed of 42. During each iteration, the distance between each bounding box and each cluster center was defined as one minus their IoU. Each bounding box was assigned to the cluster with the minimum distance, and the cluster centers were updated using the component-wise median width and height of the assigned bounding boxes. The procedure was terminated when cluster assignments no longer changed or when the maximum of 100 iterations were reached.

The nine resulting cluster centers were sorted by bounding-box area in ascending order. Five representative anchor sizes were obtained by calculating the square root of the product of width and height for each cluster center, sorting the resulting values, and selecting five values at approximately evenly spaced positions in the ordered sequence. Similarly, three representative aspect ratios were obtained by calculating the width-to-height ratio of each cluster center, sorting the ratios, and selecting three values at approximately evenly spaced positions. The resulting sizes and aspect ratios were used to configure the anchor generator across the FPN feature levels. The baseline configuration used anchor

sizes of [32, 64, 128, 256, 512] and aspect ratios of [0.5, 1.0, 2.0], whereas the representative configuration used sizes of [108, 175, 236, 335, 414] and aspect ratios of [0.39, 0.67, 2.96].

The anchor configurations were evaluated before examining their influence on the detection pipeline. This procedure separated the geometric suitability of an anchor configuration from its subsequent effect on model training and detection performance. By measuring anchor-ground-truth correspondence, the analysis determined whether the proposed representative anchor provided a closer representation of pleural effusion objects. The representative configuration was then used as the basis for the subsequent RPN sampling analysis, allowing changes in sample assignment and proposal generation to be examined in relation to the initial anchor representation.

2.3 RPN Sampling Configuration

The RPN sampling analysis examined how anchor representation and IoU-based assignment affected the distribution of foreground and background samples during training. Four experimental configurations were considered: the baseline anchor configuration, the representative anchor configuration, and both baseline and representative anchor configuration combined with adjusted RPN IoU thresholds. During anchor assignment, anchors were categorized according to their IoU with the ground-truth objects. Given lower and upper thresholds T_{low} and T_{high}, the assignment can be expressed as:

$$y_i = \begin{cases} 1, & IoU(a_i, g) \geq T_{high} \\ 0, & IoU(a_i, g) \geq T_{low} \\ -1, & otherwise \end{cases} \qquad (2)$$

Where $y_i = 1$, $y_i = 0$, and $y_i = -1$ represent positive, negative, and ignored samples, respectively. The representative anchor configuration was first evaluated to determine its effect on the availability of positive and negative samples. The RPN IoU thresholds were subsequently adjusted to $[0.3, 0.5]$ to examine how a lower positive assignment threshold changed the foreground-background composition.

As the RPN is designed to generate candidate regions for subsequent RoI head processing, an excessively strict positive threshold can limit the availability of foreground anchors, particularly when object boundaries and anchor geometry do not align precisely. This consideration is supported by medical object-detection studies that use an IoU of [0.3, 0.5] to define positive candidate regions in pulmonary nodule detection and pulmonary embolism detection using Faster R-CNN-based frameworks [16]. In chest X-ray imaging, localization performance has also been evaluated across IoU thresholds including 0.3, 0.5, and 0.7, demonstrating the increasing difficulty of localization as the IoU requirement becomes stricter [17], [18].

Therefore, [0.3, 0.5] was selected as a compromise between maintaining sufficient spatial correspondence and increasing the availability of positive anchors under the limited-annotation setting. The lower threshold of 0.3 was retained to exclude clearly mismatched background anchors, while 0.6 was considered relatively restrictive and 0.4 potentially too permissive for positive assignment. This configuration was subsequently evaluated empirically through the resulting RPN sample distribution and proposal quality.

For each configuration, the numbers of positive and negative samples generated during training were recorded and compared. This analysis was intended to identify whether improvements in anchor-object correspondence and changes in IoU assignment increased the availability of foreground samples for RPN learning before evaluating their effects on proposal generation and final detection.

2.4 Proposal Quality and Stage-Wise Detection

The quality of the generated proposals and the subsequent detection process were evaluated progressively from the region proposal stage to the final detection stage. Each ground-truth object was traced through confidence scores, successive stages consisting of proposal generation, localization, classification, and final detection relative to the total number of ground-truth objects, allowing the progression of detections to be expressed as:

$$Ground\ Truth \rightarrow Proposal \rightarrow Localization \rightarrow Classification \rightarrow Detection \quad (3)$$

For each ground-truth bounding box, the quality of the corresponding RPN proposal was first assessed using the maximum IoU between the ground-truth and the generated proposals. A proposal was considered successful when its IoU with the ground-truth was at least 0.5. This criterion was formulated as:

$$s_p(g_i) = \begin{cases} 1, & max_j IoU(g_i, p_j) \geq T_p \\ 0, & otherwise \end{cases} \quad (4)$$

where g_i denotes the *i*-th ground-truth bounding box, p_j denotes an RPN proposal, and T_p=0.5. The subsequent detection stages were then evaluated using the final bounding box associated with each ground-truth. Localization success was defined when the IoU between the ground-truth and the final predicted bounding box reached at least 0.5:

$$S_L(g_i) = \begin{cases} 1, & if\ IoU(g_i, d_i) \geq T_L \\ 0, & otherwise \end{cases} \quad (5)$$

where d_i denotes the corresponding final detection and T_L=0.5. Classification success was operationally defined based on the confidence score of a spatially valid detection, with a confidence threshold of 0.5:

$$S_C(g_i) = \begin{cases} 1, & if\ s_i \geq T_C \\ 0, & otherwise \end{cases} \quad (6)$$

where s_i is the detection confidence score and T_C=0.5. Thus, final detection success represents the cumulative success of the preceding stages and is defined as:

$$S_D(g_i) = \begin{cases} 1, & if S_P(g_i) = 1\ AND\ S_L(g_i) = 1\ AND\ S_C(g_i) = 1 \geq T_C \\ 0, & otherwise \end{cases} \quad (7)$$

For each stage $k \in$ {P, L, C, D} the success rate was calculated with respect to the total number of ground-truth objects:

$$R_k = \left(\frac{N_k}{N_{GT}}\right) \times 100\% \quad (8)$$

The failure stage was assigned hierarchically according to the first unsuccessful stage in the sequence. Consequently, an object was categorized as a proposal failure when no proposal satisfied the proposal IoU criterion; as a localization failure when a valid proposal existed but the final detection did not satisfy the localization criterion; and as a classification failure when localization was successful, but the detection confidence remained below 0.5. An object satisfying all three conditions was categorized as a detection success. This stage-wise formulation allows the reduction in successful objects to be traced progressively from proposal generation to final detection.

In addition to the cumulative stage-wise evaluation, the final predictions were analyzed diagnostically according to their localization and confidence characteristics. The prediction was categorized as poor localization when its spatial overlap with the ground-truth was below the IoU threshold of 0.5, whereas low classification confidence was assigned when the prediction achieved an IoU of at least 0.5, but its confidence score was below 0.5. Predictions satisfying both criteria were categorized as true positives. These diagnostic categories were used to identify whether unsuccessful detections were primarily associated with inadequate spatial localization or insufficient classification confidence.

The standard COCO evaluation protocol was applied to the validation sets during model training to assess object-detection performance using ground-truth bounding-box annotations. The evaluation reported average precision (AP) at specified intersection-over-union (IoU) thresholds, including AP50

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

and AP75, to assess detection performance under different localization criteria. In addition, precision, recall, and F1-score were calculated at an IoU threshold of 0.50 and a detection confidence threshold of 0.50. True positives (TP), false positives (FP), and false negatives (FN) were determined by matching predicted bounding boxes to ground-truth boxes according to the specified IoU threshold. Each ground-truth box could be matched to at most one prediction, and duplicate predictions for an already matched ground-truth box were counted as false positives. False positives on negative images were also recorded to assess the occurrence of spurious pleural effusion detections in images without annotated pleural effusion. These metrics were used to compare the detection performance of the baseline, baseline with adjusted RPN sampling, representative anchor, and representative anchor with adjusted RPN sampling configurations across the five validation folds.

3. RESULT AND DISCUSSION

3.1. Anchor Representation through Object Characterization

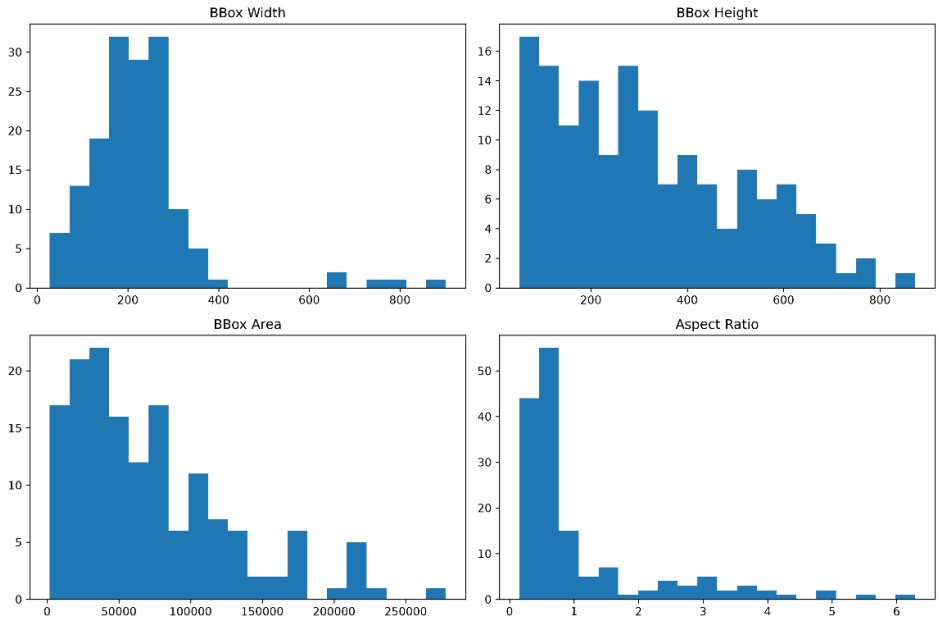
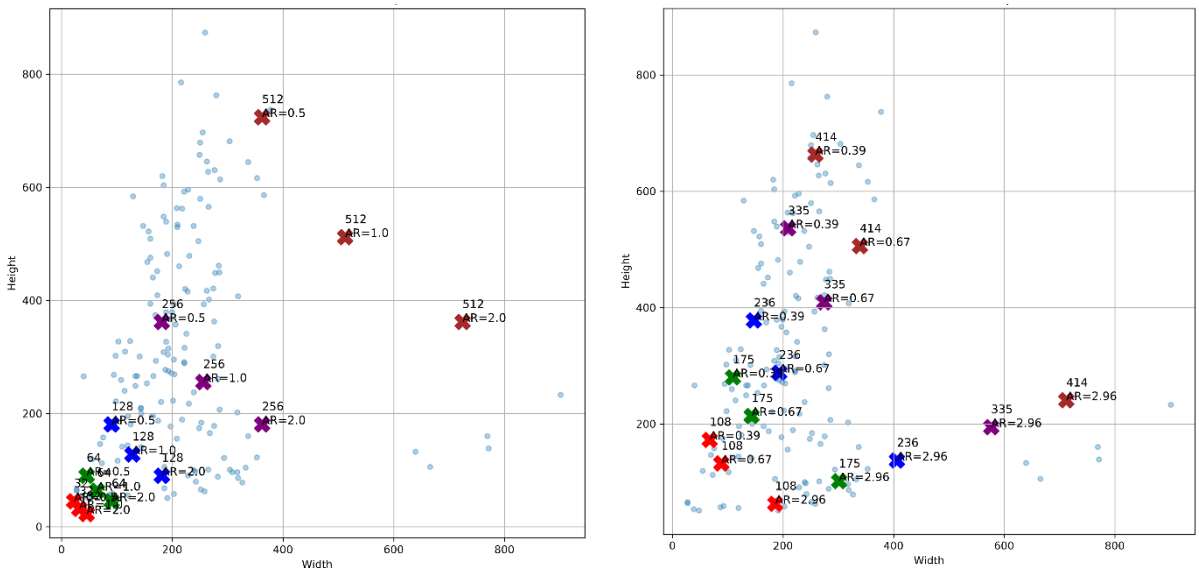


Figure 1. Ground Truth Bounding Boxes Distribution Histogram

The distribution of ground-truth bounding-boxes with width (upper left), height (upper right), area (lower left) and aspect ratio (lower right) are shown in Figure 1 above. The geometric characteristics of the pleural effusion annotations showed that most bounding boxes were concentrated within width and height ranges below 400 pixels (mid to low object sizes), while a smaller number of objects extended toward approximately 600–1000 pixels.

Based on the ground-truth width and height distribution, IoU-based clustering produced representative anchor sizes of $[108, 175, 236, 335, 414]$ with aspect ratios of $[0.39, 0.67, 2.96]$ as opposed to baseline configuration anchor sizes of $[32, 64, 128, 256, 512]$ and aspect ratios $[0.5, 1.0, 2.0]$. The resulting 5×3 configuration was subsequently selected for evaluation because it provided anchor scales and shapes that were more concentrated around the dominant object characteristics in the dataset. The comparative anchors between default and representative configuration across ground-truth distributions as shown in Figure 2 below.

Vol. 11, No. 2, Oktober 2026, Page. 521-538
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI



(a) Baseline Anchor (b) Representative Anchor

Figure 2. Anchor Configurations across Ground Truth Bounding Boxes

The markers with different colors illustrated each of five combinations of anchor sizes and aspect ratios, while the coordinate (x, y) represented a pair of localized covered areas. The baseline anchor configuration provided uneven coverage of the observed object distribution. Several anchor combinations were associated with only a small number of objects or received no corresponding object allocation within range of 256 to 512 anchor sizes, particularly at the extreme scales. The representative configuration further validated its most matched localization covered area of distributed ground-truth bounding-boxes.

Table 1. IoU Coverage Between Baseline and Representative Anchors

IoU Metrics	Anchors	
	Baseline	Representative
Average IoU	65.8%	**76.9%**
Median IoU	64.8%	**79.2%**
Min IoU	**36.4%**	14.9%
Max IoU	96.4%	**99.6%**
Coverage IoU ≥ 0.50	**96.1%**	95.4%
Coverage IoU ≥ 0.60	60.8%	**87.6%**
Coverage IoU ≥ 0.70	35.3%	**79.1%**
Coverage IoU ≥ 0.80	15.0%	**47.1%**

The improvement in geometric representation was confirmed through the best-matching IoU analysis as shown in Table 1 above. The average best IoU increased from 0.658 for the default configuration to 0.769 for the representative 5×3 configuration, corresponding to an improvement of approximately 16.7%. More pronounced differences were observed at higher IoU thresholds: coverage at IoU ≥ 0.60 increased from 60.8% to 87.6%, at IoU ≥ 0.70 from 35.3% to 79.1%, and at IoU ≥ 0.80 from 15.0% to 47.1%.

However, coverage at IoU ≥ 0.50 decreased slightly from 96.1% to 95.4%, indicating that the improved average compatibility was primarily achieved for most objects rather than uniformly across the entire distribution. This result suggests that the representative configuration better matches the central distribution of pleural effusion objects while providing slightly weaker coverage for objects located at the distribution tails.

3.2. RPN Sampling Assignment on Sample Distribution

The effect of anchor representation and IoU threshold adjustment on RPN sampling was examined by comparing the average numbers of positive and negative anchors across the four experimental configurations. The positive samples proportion defines:

$$Positive\ Proportion = \frac{N_{pos}}{N_{pos}+N_{neg}} \times 100\% \qquad (9)$$

Under the baseline configuration, the RPN generated approximately 4 positive and 252 negative samples per image, indicating a strongly background-dominated sample distribution. Replacing the default anchors with the representative configuration increased the average number of positive samples to approximately 31 while reducing negative samples to 225. This change was further amplified after adjusting the RPN IoU thresholds to $[0.3, 0.5]$, which increased the number of positive samples to approximately 107 and reduced negative samples to approximately 149.

Table 2. RPN Sampling Assignment IoU Thresholds

Configuration	Thresh [0.30, 0.70]		Ratio	Thresh [0.30, 0.50]		Ratio
	Negative	Positive		Negative	Positive	
Baseline	252	4	1.6%	237	19	7.6%
Representative	225	31	**12.2%**	149	107	**41.8%**

The progressive changes in sample composition indicate that both anchor representation and IoU-based assignment affected the availability of foreground samples during RPN training. Compared with the baseline, the representative anchor configuration increased the number of positive samples by more than seven times, whereas the combination of representative anchors and the adjusted IoU thresholds produced approximately 27 times more positive samples. At the same time, the number of negative samples progressively decreased across the four configurations. These results demonstrate a substantial redistribution of the RPN training samples, shifting the sampling composition from a predominantly background-oriented distribution toward a more balanced foreground-background representation.

The redistribution observed at the RPN stage was also reflected in the samples forwarded to the RoI head. As shown in Table 3 below, under the baseline configuration, each fold contained only 14 foreground samples and 498 background samples, corresponding to approximately 2.7% foreground representation while adjusted threshold increased to 3.7%. After replacing the default anchors with the representative configuration, the number of foreground samples increased to 25–29 per fold, while the number of background samples decreased to 483–487, increasing the foreground proportion to approximately 4.9–5.7%. This result indicates that the improved geometric correspondence between anchors and ground-truth objects was propagated beyond anchor-level assignment, resulting in a larger number of foreground proposals available for second-stage learning.

Table 3. RoI Head Sample Distribution

Configuration	Thresh [0.30, 0.70]		Ratio	Thresh [0.30, 0.50]		Ratio
	Negative	Positive		Negative	Positive	
Baseline	498	14	2.7%	493	19	3.7%
Representative	485	27	**5.7%**	**477**	**35**	**7.0%**

A further increase was observed after adjusting the RPN sampling assignment thresholds from $[0.3, 0.7]$ to $[0.3, 0.5]$. The RoI head received 33–36 foreground samples across the folds, compared with 14 samples in the baseline configuration and 25–29 samples with representative anchors alone. Except for the recorded Fold 4 value, the corresponding foreground proportion increased to approximately 6.4–7.0%, while the number of background samples was reduced to 476–479.

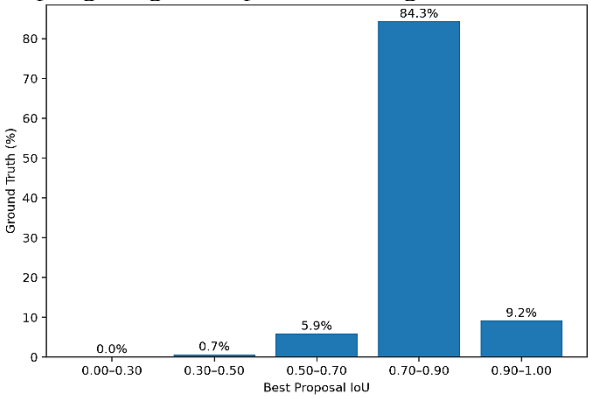
Overall, the results demonstrate a progressive shift in the RoI training population, in which representative anchors increased the availability of foreground proposals and the adjusted RPN sampling assignment further amplified this effect. However, background samples remained dominant in all configurations, indicating that the second-stage detector continued to operate under a substantially background-oriented sampling distribution.

Vol. 11, No. 2, Oktober 2026, Page. 521-538
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

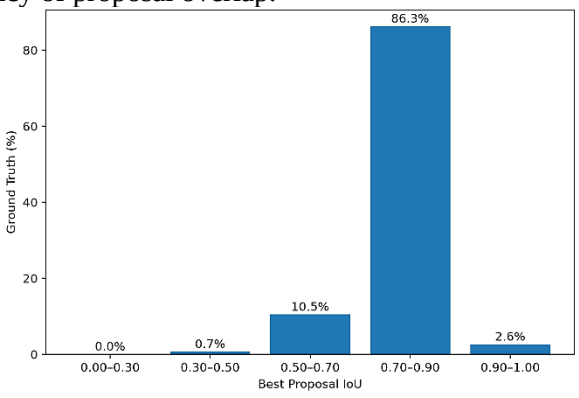
3.3. RPN Proposal Quality

The quality of RPN-generated proposals was evaluated by selecting the proposal with the highest IoU for each of the 153 ground-truth objects. Under the default RPN sampling assignment thresholds $[0.3, 0.7]$, the representative anchor configuration produced a mean best-proposal IoU of 0.7943, slightly lower than the 0.8087 obtained with the baseline anchors, although the median IoU was comparable (0.8159 vs. 0.8125). The coverage at IoU ≥ 0.50 also remained high for both configurations, at 98.69% for representative anchors and 99.35% for baseline anchors.

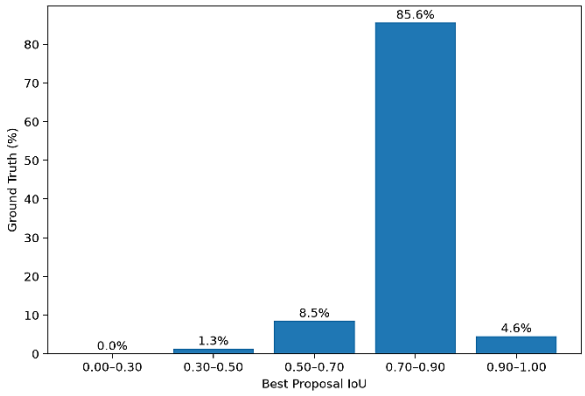
After adjusting the RPN sampling assignment thresholds to $[0.3, 0.5]$, the representative configuration achieved the highest mean and median IoU, increasing to 0.8135 and 0.8410, respectively. Its IoU ≥ 0.50 coverage remained high at 98.04%. These results indicate that representative anchors alone did not uniformly improve proposal quality, whereas their combination with adjusted RPN sampling assignment produced a higher central tendency of proposal overlap.



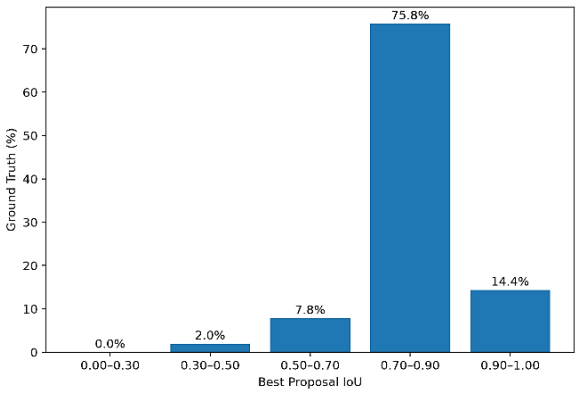
(a) Baseline Thresh [0.3, 0.7]



(b) Baseline Thresh [0.3, 0.5]



(c) Representative Thresh [0.3, 0.7]



(d) Representative Thresh [0.3, 0.5]

Figure 3. RPN Best Proposal Quality Distribution

The distribution of best proposal IoU as shown in Figure 3 above further illustrates the shift in proposal quality. The IoU distributions further illustrate this difference. With the default assignment, both configurations were dominated by proposals in the 0.70–0.90 range, accounting for 84.31% of objects for the baseline and 85.62% for representative anchors. After the RPN threshold adjustment, the representative configuration shifted further toward high-quality proposals, with 14.38% of objects reaching IoU ≥ 0.90 compared with 4.58% before adjustment, while the proportion in the 0.30–0.50 range remained low at 1.96%.

In contrast, the adjusted baseline configuration showed a reduction in the ≥ 0.90 group from 9.15% to 2.61%, despite maintaining a high mean IoU of 0.7994. Overall, the results suggest that RPN sampling adjustment interacted with anchor representation rather than producing a uniform improvement across configurations, with the representative anchor configuration benefiting most clearly from the adjusted assignment in terms of median IoU and the proportion of highly overlapping proposals.

3.4. Stage-wise Detection Outcomes

3.4.1 Detection Confidence Score

The detection confidence analysis was conducted to identify whether unsuccessful detections were primarily associated with inadequate localization or insufficient classification confidence. Based on the criteria defined in Section 2.4, a prediction with an IoU below 0.5 was categorized as poor localization, whereas a spatially valid prediction with an IoU of at least 0.5 but a confidence score below 0.5 was categorized as low classification confidence. Therefore, the confidence distribution was interpreted together with the corresponding localization quality rather than from the confidence score alone.

Across the four configurations, the detector produced consistently high confidence scores for its predictions as shown in Table 4 below. The mean confidence ranged from 0.7886 to 0.8055, indicating that most retained detections were assigned very high classification confidence. The baseline configuration yielded a mean confidence of 0.7886, which increased slightly to 0.7901 with representative anchors and to 0.7985 after adjusting the RPN IoU thresholds. The highest mean confidence of 0.8055 was obtained with the representative anchors combined with the adjusted RPN sampling. However, this configuration did not produce the highest number of true-positive detections, with 121 true positives compared with 122 for the baseline and representative configurations and 123 for the adjusted baseline.

Table 4. Detection Confidence Scores

Configuration	$\bar{x}$ confidence	TP	Low confidence	Poor localization
	Thresh [0.30, 0.70]			
Baseline	0.7886	122	25	6
Representative	0.7901	122	26	5
	Thresh [0.30, 0.50]			
Baseline	0.7985	**123**	24	6
Representative	**0.8055**	121	23	9

The results in Table 4 show that low classification confidence accounted for 23–26 objects across the evaluated configurations, while poor localization accounted for 5–9 objects. Low-confidence predictions indicate cases in which candidate regions were localized adequately but did not receive sufficiently high classification confidence, whereas poor-localization cases represent predictions that failed to achieve the required spatial overlap with the ground-truth bounding boxes. Notably, the Representative with adjusted RPN sampling configuration achieved the highest mean confidence (0.8055) and the fewest low-confidence cases (23), yet it also exhibited the largest number of poor-localization cases (9) and the lowest number of true positives (121). In contrast, Baseline with adjusted RPN sampling, achieved the highest number of true positives (123), despite a lower mean confidence (0.7985) and six poor-localization cases.

This discrepancy suggests that the representative anchor configuration may have improved classification confidence without producing a corresponding improvement in bounding-box localization. One possible explanation is that changes in anchor representation affected the spatial alignment of candidate regions and subsequent box regression, allowing predictions to receive higher confidence scores while still failing the localization criterion. However, the aggregate results do not establish whether the additional localization failures directly account for the reduction in true positives. Overall, these findings demonstrate that classification confidence alone is insufficient to characterize detection effectiveness and that confidence and localization quality must be considered jointly when evaluating the effects of anchor representation and RPN optimization.

3.4.2 Success Rate at each Detection Stage

Following the Faster R-CNN pipeline described earlier in Equation 3, the progression of object detection success from ground-truth toward the final detection stage can be systematically tracked as presented in Table 5 below. Proposal success represents the availability of an RPN proposal with an IoU of at least 0.5 with the ground-truth, followed by localization success when the final predicted bounding box reaches the same IoU criterion. Classification success is then determined from the confidence score,

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

with a minimum threshold of 0.5, while detection success requires all preceding conditions to be satisfied.

Table 5. Proposal Success at each Detection Stage

Configuration	Thresh [0.30, 0.70]			
	Proposal	Localization	Classification	Final Detection
Baseline	152	147	121	121
Representative	151	148	122	122
	Thresh [0.30, 0.50]			
Baseline	152	147	123	123
Representative	150	144	124	121

Results of Table 5 above demonstrate a progressive reduction in the number of successful objects across the detection pipeline. The baseline configuration achieved 99.35% proposal success and 96.08% localization success, followed by 79.08% classification and detection success. The representative anchor configuration produced a slightly lower proposal success rate of 98.69%, but the highest localization success among the configurations without RPN adjustment (96.73%) and a classification success of 79.74%. After applying the adjusted RPN IoU thresholds, the baseline configuration maintained 99.35% proposal success and 96.08% localization success, while classification and detection success increased to 80.39%. These results indicate that the first two stages of the detection pipeline were relatively stable, with most ground-truth objects successfully represented by proposals and subsequently localized.

A different pattern was observed when representative anchors were combined with the adjusted RPN sampling. Although proposal and localization success decreased to 98.04% and 94.12% respectively, classification success increased to the highest value of 81.05%. However, only 121 objects (79.08%) were retained as final detections, indicating a further loss of three objects between classification and final detection. Across all configurations, the largest reduction consistently occurred between localization and classification, ranging from approximately 16.34 to 17.65 percentage points, while the proposal-to-localization reduction remained comparatively small.

The difference between proposal and localization success represents objects for which the RPN generated a sufficiently overlapping candidate, but the subsequent bounding-box prediction did not maintain adequate spatial correspondence. A further reduction from localization to classification success indicates objects that were spatially localized but received confidence scores below the predefined threshold. Finally, detection success represents the objects that successfully passed all evaluated stages. This progressive analysis therefore identifies the stage at which successful objects are lost and provides a direct assessment of how proposal quality, localization, and confidence contribute to the final detection outcome.

3.4.3 Failure Distribution

The stage-wise success rates indicate that the major reduction in object retention occurred during the transition from localization to classification. This pattern is further confirmed by the failure distribution across the four configurations shown in Table 6 below. In the baseline configuration, 26 objects (16.99%) were categorized as classification failures, representing the largest source of failure, followed by 5 objects (3.27%) with localization failure and 1 object (0.65%) with proposal failure.

Table 6. Proportion Of Failure Distribution

Configuration	Detection Success	Thresh [0.30, 0.70]		
		Proposal	Localization	Classification
Baseline	121	1	5	26
Representative	122	2	4	25
		Thresh [0.30, 0.50]		
Baseline	123	1	5	24
Representative	121	3	6	23

The representative anchor configuration showed a similar pattern, with classification failure decreasing slightly to 25 objects (16.34%), while localization and proposal failures accounted for only 4 (2.61%) and 2 (1.31%) objects, respectively. After RPN adjustment, the baseline configuration produced the lowest overall failure outside the final detection group, with 24 classification failures (15.69%), 5 localization failures (3.27%), and only 1 proposal failure (0.65%). These results consistently identify classification as the dominant failure category, while proposal generation contributed the smallest proportion of failures.

The representative anchor configuration combined with adjusted RPN sampling exhibited a slightly different pattern. Although it produced the lowest classification-failure rate of 15.03% (23 objects), it simultaneously showed higher proposal and localization failures, at 1.96% (3 objects) and 3.92% (6 objects), respectively. Consequently, its final detection success decreased to 79.08%, compared with 80.39% for the adjusted baseline. This result is important because it shows that increasing the availability of positive RPN samples does not necessarily improve the complete detection pipeline uniformly.

The changes introduced at the RPN stage may improve classification outcomes while also altering proposal and localization behavior. Overall, the failure distribution demonstrates that classification remained the dominant source of detection failure across all configurations, whereas proposal and localization failures were comparatively limited. This finding suggests that the effect of RPN optimization should be interpreted as an improvement in the upstream proposal process rather than as a direct guarantee of improved final detection performance.

3.4.4 Performance Evaluation during Training

Table 7. Mean COCO evaluation results across five validation folds

Configuration	Thresh [0.30, 0.70]		
	AP	**AP50**	**AP75**
Baseline	2.53	9.96	0.63
Representative	2.02	7.09	0.68
	Thresh [0.30, 0.50]		
Baseline	1.84	8.22	0.52
Representative	1.66	6.74	0.14

Table 8. Mean detection metrics across five validation folds at confidence threshold 0.50 and IoU threshold 0.50. TP, FP, and FN represent the cumulative counts across the five validation folds. In contrast, precision, recall, and F1-score are reported as the arithmetic means of the corresponding metrics calculated independently for each fold. Therefore, these mean metrics are not recalculated from the cumulative TP, FP, and FN counts.

Configuration	Thresh [0.30, 0.70]					
	TP	**FP**	**FN**	**Precision**	**Recall**	**F1-score**
Baseline	25	101	128	19.6	16.3	17.7
Representative	15	83	138	13.2	9.8	11.0
	Thresh [0.30, 0.50]					
Baseline	25	99	128	20.5	16.3	18.0
Representative	16	69	137	18.4	10.3	13.1

Table 9. False-positive detection negative validation images. The total of 153 negative-image evaluations represents the cumulative number of negative-image evaluations across the five validation folds. Images are counted separately for each fold in which they are evaluated; therefore, this total does not necessarily represent 153 unique images across the entire cross-validation procedure. The FP image rate is calculated as the percentage of evaluated negative images containing at least one false-positive detection, rather than the percentage or number of false-positive bounding boxes.

Configuration	Thresh [0.30, 0.70]		
	Negative images evaluated	**Negative images with FP**	**FP image rate (%)**
Baseline	153	19	12.4
Representative	153	14	9.2

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

	Thresh [0.30, 0.50]		
Baseline	153	20	13.1
Representative	153	19	12.4

Across the five validation folds as shown on Table 7 above, the baseline configuration achieved the highest mean COCO AP (2.53%) and AP50 (9.96%), followed by baseline with adjusted RPN sampling (1.84% and 8.22%), representative anchors (2.02% and 7.09%), and representative anchors with adjusted RPN (1.66% and 6.74%), respectively. The mean AP75 values were 0.63%, 0.52%, 0.68%, and 0.14%, respectively, indicating consistently limited performance at the stricter IoU threshold.

Based on the detection counts at a confidence threshold of 0.50 and IoU threshold of 0.50 shown on table 8 above, the baseline and baseline with adjusted RPN sampling each achieved 25 TP and 128 FN across the folds, while the representative anchor configurations achieved 15 TP and 138 FN and 16 TP and 137 FN, respectively. Mean precision, recall, and F1-score across folds were 19.6%, 16.3%, and 17.7% for the baseline; 20.5%, 16.3%, and 18.0% for baseline with adjusted RPN sampling; 13.2%, 9.8%, and 11.0% for representative anchors; and 18.4%, 10.3%, and 13.1% for representative anchors with adjusted RPN.

The baseline also exhibited substantial fold-to-fold variation, with AP50 ranging from 7.81% to 13.81%. Across the 153 negative-image evaluations accumulated over the five folds, false-positive detections occurred in 19, 20, 14, and 19 image-fold instances for the four configurations, respectively. Overall, the baseline provided the strongest average COCO AP and AP50, whereas adjusted RPN sampling produced only a marginal improvement in mean F1-score when combined with the baseline anchors.

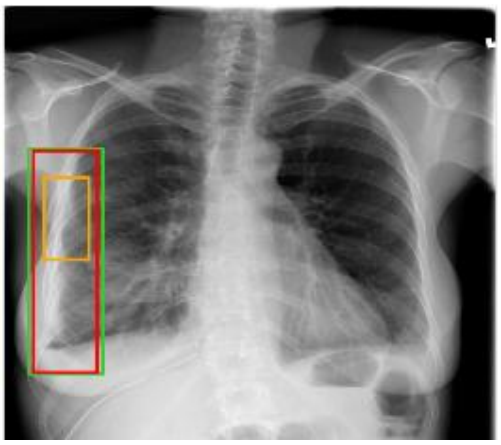
These results suggest that the limited validation performance was influenced by the interaction between anchor configuration, proposal sampling, and the scarcity of annotated training examples, although the aggregate metrics alone cannot establish a single causal mechanism. The baseline anchors may have provided a more favorable overall balance between object-scale coverage and localization for this dataset, as reflected in their higher mean AP and AP50. In contrast, the representative anchor configurations did not consistently improve detection despite being designed to better reflect the observed bounding-box distribution, suggesting that anchor matching quality alone was insufficient to ensure of effective learning of pleural effusion features.

Adjusting RPN sampling IoU thresholds and proposal settings yielded a small increase in mean precision and F1-score with the baseline anchors, but TP and FN totals remained unchanged. This pattern indicates that the adjustment primarily altered the balance of false-positive detections rather than improving the number of successfully detected lesions. The representative anchors combined with adjusted RPN also reduced the total number of false-positive bounding-box detections compared with the representative anchors alone, but this reduction was accompanied by only one additional TP across all five folds and persistently low recall. The generally low AP75 values further suggest that accurate localization at stricter IoU requirements remained difficult. Taken together, the findings indicate that neither anchor redesign nor RPN adjustment alone was sufficient to overcome the learning limitations associated with the small annotated dataset, and that improvements in proposal generation or sampling did not consistently translate into better final detection performance.

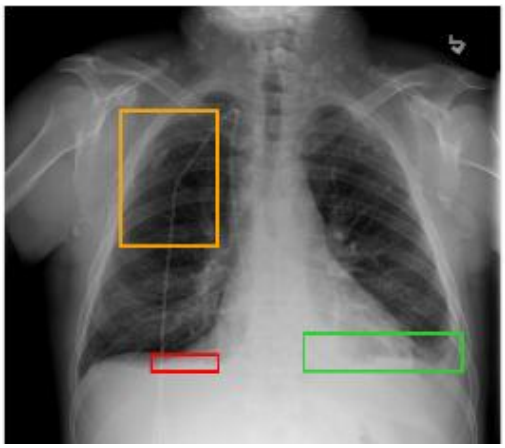
3.4.5 Error Analysis

The combined analysis of confidence, stage-wise success, and failure distribution provides a consistent explanation of the observed detection behavior. The high median confidence across configurations indicates that the detector was generally highly confident in its retained predictions, but the relatively stable number of true positives shows that confidence alone did not determine successful detection. The stage-wise results further indicate that proposal generation and localization were comparatively effective, with success rates consistently above 94%, while the largest reduction occurred at the classification stage. This pattern is consistent with the failure distribution, in which classification failure accounted for 15.03–16.99% of the ground-truth objects across configurations, substantially exceeding proposal failure (0.65–1.96%) and localization failure (2.61–3.92%). Representative examples of these error patterns are presented in Figure 4 below

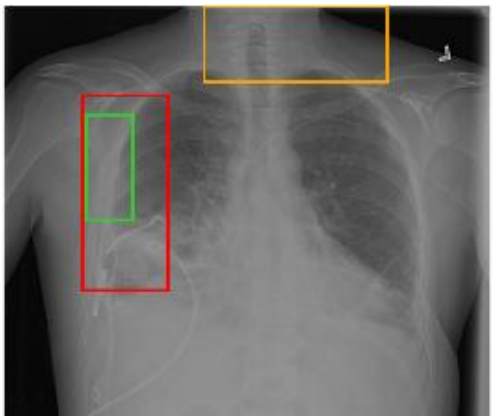
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI



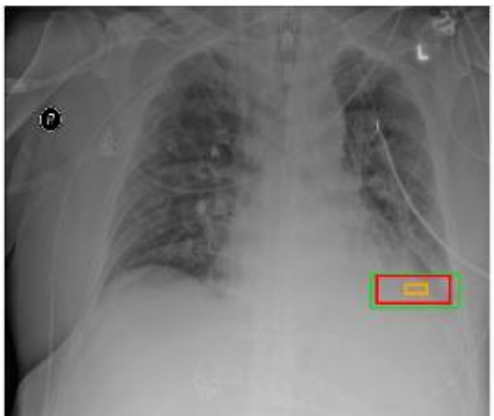
True Positive Case (Image 84)
RPN IoU: 0.219 | objectness: 0.997
RoI IoU: 0.851 | confidence: 0.994



Proposal Failure Case (Image 45)
RPN IoU: 0.0 | objectness: 0.071
RoI IoU: 0.343 | confidence: 0.0



Poor Localization Case (Image 15)
RPN IoU: 0.0 | objectness: 0.001
RoI IoU: 0.272 | confidence: 0.740



Low Confidence Case (Image 7)
RPN IoU: 0.065 | objectness: 0.995
RoI IoU: 0.633 | confidence: 0.125

Figure 4. Case Visualizations across Detection Stages

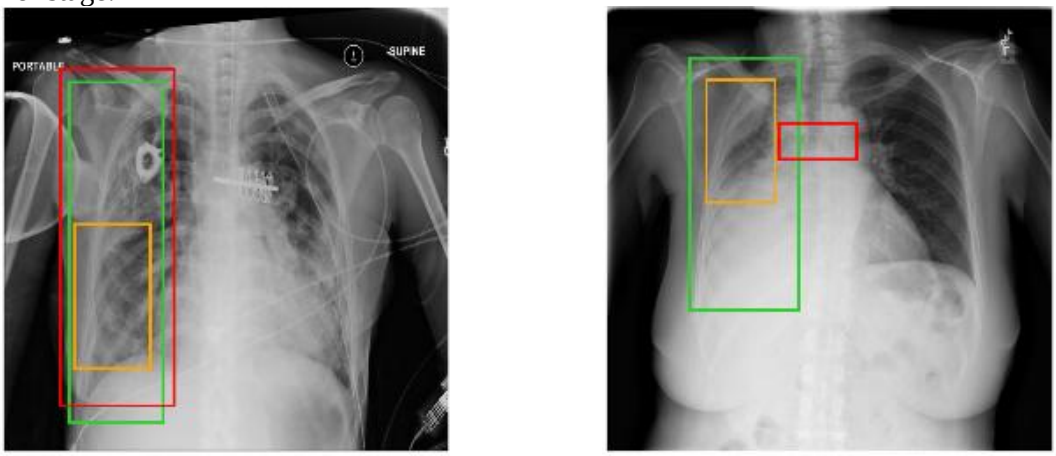
The representative anchor and adjusted RPN configurations therefore improved aspects of the upstream proposal process and, in some cases, reduced classification failures, but these changes were not uniformly translated into higher final detection rates. The findings suggest that the principal limitation observed in this experimental setting was not the availability of candidate regions, but the ability of the subsequent detection process to consistently convert sufficiently localized proposals into correct final detections.

The visualizations of representative detection cases further illustrate the distinct failure characteristics across the Faster R-CNN pipeline. In these visualizations, the green bounding box represents the ground-truth, the orange bounding box represents the RPN proposal, and the red bounding box represents the final RoI detection. In the true-positive case, although the RPN proposal has a relatively low IoU of 0.219, its high objectness score (0.997) is followed by a well-localized RoI detection with an IoU of 0.851 and a high classification confidence of 0.994.

In contrast, the proposal-failure case exhibits no meaningful overlap between the RPN proposal and the ground-truth (IoU = 0.000), accompanied by a low objectness score (0.071), resulting in a poorly localized RoI detection (IoU = 0.343) with zero confidence. The poor-localization case shows a different failure pattern, where the RoI produces a relatively high confidence of 0.740 despite a low IoU of 0.272, indicating that strong classification confidence does not necessarily correspond to accurate localization. Meanwhile, the low-confidence case demonstrates the opposite behavior: the RPN produces a high-objectness proposal (0.995), while the RoI achieves a reasonably good localization (IoU = 0.633) but assigns only 0.125 confidence to the detection.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

The low-confidence cases further demonstrate that the remaining detection errors cannot be attributed solely to insufficient RPN proposal quality. As shown in Figure 3 earlier, the RPN is generally capable of generating proposals with sufficiently high overlap with the ground-truth regions. Nevertheless, this proposal quality does not consistently translate into high-confidence final detections at the RoI stage.



Low Confidence Case (Image 113)
RPN IoU: 0.344 | objectness: 0.920
RoI IoU: 0.936 | confidence: 0.0

Low Confidence Case (Image 89)
RPN IoU: 0.304 | objectness: 0.870
RoI IoU: 0.720 | confidence: 0.0

Figure 5. Low Confidence Cases of Different Condition

Two representative cases illustrate this behavior as shown in Figure 5 above. In image 113, the RPN proposal achieves an IoU of 0.344 with an objectness score of 0.920, while the final RoI detection attains a substantially higher IoU of 0.936 but produces a confidence score of 0.0. Although the final detection has excellent overlaps with the ground truth, its bounding box extends beyond the ground-truth region, indicating that the detector can produce a spatially overlapping region that is geometrically different from the annotated lesion extent. In contrast, image 89 presents the opposite pattern. The RPN proposal obtains an IoU of 0.304 with an objectness score of 0.870, and the final detection achieves an IoU of 0.720, yet its confidence remains 0.0. In this case, the ground-truth lesion is considerably larger than the predicted bounding box.

Thus, despite the relatively good IoU resulting from partial spatial overlap, the predicted region represents only a smaller portion of the annotated lesion. These cases indicate that IoU alone may not fully characterize the spatial adequacy of the final detection, particularly when substantial differences exist in bounding-box extent and scale. More importantly, the consistently high RPN objectness scores in both cases suggest that the proposal stage can identify relevant candidate regions, whereas the subsequent RoI stage remains conservative in assigning foreground confidence. Prevalence the prevalence of low-confidence detections is more plausibly associated with the interaction between localization extent and RoI classification confidence rather than with a fundamental inability of the RPN to generate suitable proposals.

Taking together, these observations suggest a shift in the primary bottleneck of the detection pipeline. While the RPN proposal analysis in Figure 3 indicates that proposal quality is already adequate, the low-confidence examples reveal that the subsequent RoI stage may remain sensitive to discrepancies between the predicted and annotated lesion extents. Consequently, improving proposal quality alone is unlikely to fully resolve the low-confidence problem; instead, the remaining limitation appears to lie in how the RoI head converts sufficiently relevant proposals into confident final detections under the limited-annotation setting.

3.5. Discussion

The stage-wise analysis indicates that the dominant loss of detection performance occurs at the classification stage, with approximately 15–17% of ground-truth objects failing after the proposal and localization stages have already been successfully passed. This finding suggests that the remaining limitation is less related to the availability or overall spatial quality of candidate regions and more related

to the detector's ability to assign sufficient confidence to already localized objects. Complementary evaluation using the standard COCO evaluation protocol during training provides an additional perspective on overall detection performance through established object-detection metrics, including average precision at different IoU thresholds. While these metrics facilitate consistent performance assessment across model configurations, they do not directly identify the stage at which detection failures occur.

The COCO-based evaluation therefore complements rather than replaces the stage-wise analysis, which provides a more detailed diagnosis of the underlying detection errors. This interpretation is supported by the RPN proposal quality analysis, which shows that the RPN is generally capable of generating sufficiently relevant candidate regions. However, the subsequent low-confidence cases demonstrate that adequate proposal quality does not necessarily translate into confident final detections. For example, the final detection in image 113 achieves a high IoU of 0.936 but extends beyond the ground-truth region, whereas in image 89 the predicted region is considerably smaller than the ground-truth despite achieving an IoU of 0.720. These cases indicate that differences in bounding-box extent and scale may remain influential at the RoI stage, even when the resulting localization achieves a relatively good IoU.

A practical mitigation strategy is therefore to improve the learning of difficult foreground instances rather than further increasing proposal generation capacity. Hard-positive mining can be employed to give greater training emphasis to positive RoIs that are spatially valid but receive relatively low classification confidence. Focal loss is another potential strategy because it reduces the contribution of easily classified samples and places greater emphasis on difficult examples, which may be beneficial under the limited-annotation setting used in this study.

Based on these findings, subsequent optimization should prioritize the RoI classification stage while retaining the representative anchor and RPN configuration established in this study as the preceding baseline. The recommended direction is to evaluate hard-positive sampling and confidence-aware classification losses through controlled ablation experiments, while monitoring not only final detection performance but also the stage-wise success rates and the proportion of Low Classification Confidence cases. This evaluation would determine whether improvements in classification learning can reduce the observed 15–17% failure while preserving the localization quality already achieved by the preceding stages. Such an approach provides a targeted continuation of the present empirical analysis, where the identified classification-stage limitation, rather than proposal generation capacity, serves as the basis for selecting and evaluating the next optimization strategy.

Beyond two-stage detectors, the difficulty of diagnosing performance bottlenecks is equally relevant for single-stage detectors trained on limited data. In a recent application of EfficientDet with a Bidirectional Feature Pyramid Network (Bi-FPN) for small-object detection of vaname shrimp [26], detection accuracy of only 77.3% using 178 annotated images. Although the Bi-FPN design improved multi-scale feature fusion, the study evaluated only final detection accuracy and could not determine whether the remaining errors originated from object representation, sample imbalance, or the classification stage.

This mirrors the limitation addressed in the present study: final metrics alone obscure where a detector fails. Moreover, the severe foreground-background imbalance reported in our RPN analysis (4 positive versus 252 negative samples per image under the baseline configuration) is typically aggravated in small datasets such as the 158-image training set used [26], where minority-object anchors are even more scarcely represented. The stage-wise diagnostic approach demonstrated here therefore offers a transferable framework not only for medical imaging but also for resource-constrained detection tasks in aquaculture and other domains, where identifying the dominant failure stage is a prerequisite for deciding whether to invest in additional data, anchor reconfiguration, or classifier refinement.

4. CONCLUSION

This study presented an empirical analysis of anchor representation and RPN sampling in Faster R-CNN for pleural effusion detection from chest X-ray images. Ground-truth object characterization enabled the construction of representative anchors that better matched the geometric distribution of the annotated objects. However, improved anchor correspondence did not consistently translate into better proposal quality or final detection performance across configurations. Similarly, adjusting the RPN IoU

assignment thresholds substantially changed the training sample distribution and the characteristics of the generated proposals, but these changes did not produce consistent improvements in final detection outcomes. These findings indicate that anchor representation and RPN sampling should be evaluated using multiple geometric and detection metrics rather than assuming that improvements at an intermediate stage will necessarily improve overall detection performance.

The stage-wise analysis provided a more detailed understanding of the detector's limitations by examining proposal generation, localization, and classification separately. Although a high proportion of ground-truth objects received sufficiently overlapping proposals, some detections still failed to satisfy the localization or classification criteria. In particular, the observed classification-related failures indicate that adequate proposal availability and spatial correspondence alone are insufficient to ensure successful final detections. The main contribution of this analysis is therefore its diagnostic value: it helps identify where detection failures occur within the two-stage detection pipeline and provides an empirical basis for directing subsequent optimization efforts toward the stages that remain limiting factors.

These findings should be interpreted considering the study's limitations. The experiments were conducted on a relatively small dataset of 306 images, including only 153 pleural effusion images with bounding-box annotations. This limited number of annotated positive cases restricts the diversity of lesion appearances represented during training and evaluation and may affect the stability of the observed performance across data partitions. In addition, uncertainty in the placement and extent of bounding-box annotations may influence IoU-based assessments of anchor correspondence, localization quality, and detection success. Although five-fold cross-validation was used, the evaluation remained within the same source dataset and did not include external validation on an independent dataset. Consequently, the results should be regarded as preliminary empirical evidence under a specific experimental setting, rather than proof of clinical reliability or generalizability to other datasets or imaging populations.

Future work should therefore investigate these findings using larger datasets with more extensively annotated positive cases, assess annotation consistency, and conduct external validation on independent chest X-ray datasets. Further experiments should also determine whether improvements in anchor representation and RPN sampling assignment lead to reproducible gains in final detection performance across different data distributions. Until such evidence is available, the proposed configurations should be considered experimental alternatives rather than established improvements for clinical deployment. Overall, this study demonstrates the value of stage-wise evaluation as a diagnostic framework for understanding the limitations of two-stage object detectors and guiding subsequent model development under annotation scarcity.

5. REFERENCE

[1] E. Weldon and J. Williams, “Pleural Disease in the Emergency Department,” *Emerg. Med. Clin. North Am.*, vol. 30, no. 2, pp. 475–499, May 2012, doi: 10.1016/J.EMC.2011.10.012.

[2] D. Bates, N. Yang, M. Bailey, and R. Bellomo, “Prevalence, characteristics, drainage and outcome of radiologically diagnosed pleural effusions in critically ill patients,” *Critical Care and Resuscitation*, vol. 22, no. 1, pp. 45–52, Mar. 2020, doi: 10.51893/2020.1.OA5.

[3] E. DeBiasi and J. Puchalski, “Pleural effusions as markers of mortality and disease severity: A state-of-the-art review,” *Curr. Opin. Pulm. Med.*, vol. 22, no. 4, pp. 386–391, Jul. 2016, doi: 10.1097/MCP.0000000000000278.

[4] C. A. Aprilia, R. D. Aulia, A. E. Wibowo, E. N. Permatasari, S. Bakhri, and Abdurrahman, “View of Massive pleural effusion due to heart failure in an undetected diabetic hypertensive patient: A case report.” Accessed: Dec. 18, 2025. [Online]. Available: https://journal.uii.ac.id/JKKI/article/view/19749/pdf

[5] F. Karataş and Ö. Dikensoy, “Artificial Intelligence in Pleural Diseases: Current Applications and Next Steps,” Jan. 01, 2026, *AVES*. doi: 10.4274/ThoracResPract.2025.2025-6-2.

[6] Y. Y. Huang *et al.*, “Automated Computer-Assisted Diagnosis of Pleural Effusion in Chest X-Rays via Deep Learning,” *Diagnostics*, vol. 15, no. 18, Sep. 2025, doi: 10.3390/diagnostics15182322.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

[7] D. C. Lepcha *et al.*, "Deep Learning in Medical Image Analysis: A Comprehensive Review of Algorithms, Trends, Applications, and Challenges," *Computer Modeling in Engineering & Sciences*, vol. 145, no. 2, pp. 1487–1573, Nov. 2025, doi: 10.32604/CMES.2025.070964.

[8] S. Matta *et al.*, "A systematic review of generalization research in medical image classification," Dec. 01, 2024, *Elsevier Ltd*. doi: 10.1016/j.compbiomed.2024.109256.

[9] Y. Zhong, J. Wang, J. Peng, and L. Zhang, "Anchor Box Optimization for Object Detection," Jan. 2020, [Online]. Available: http://arxiv.org/abs/1812.00469

[10] W. Ma, T. Tian, H. Xu, Y. Huang, and Z. Li, "AABO: Adaptive Anchor Box Optimization for Object Detection via Bayesian Sub-sampling," Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.09336

[11] D. Miller, G. Goode, C. Bennie, P. Moghadam, and R. Jurdak, "Why Object Detectors Fail: Investigating the Influence of the Dataset."

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Jan. 2016, Accessed: Dec. 08, 2025. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/results

[13] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region Proposal by Guided Anchoring," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1901.03278

[14] L. Zhu, Z. Xie, L. Liu, B. Tao, and W. Tao, "IoU-uniform R-CNN: Breaking Through the Limitations of RPN," Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.05190

[15] T. Vu, H. Jang, T. X. Pham, and C. D. Yoo, "Cascade RPN: Delving into High-Quality Region Proposal Network with Adaptive Convolution," Dec. 2019, [Online]. Available: http://arxiv.org/abs/1909.06720

[16] Y. Zhao *et al.*, "Pulmonary Nodule Detection Based on Multiscale Feature Fusion," *Comput. Math. Methods Med.*, vol. 2022, 2022, doi: 10.1155/2022/8903037.

[17] S. Yao, Y. Chen, X. Tian, and R. Jiang, "Pneumonia Detection Using an Improved Algorithm Based on Faster R-CNN," *Comput. Math. Methods Med.*, vol. 2021, 2021, doi: 10.1155/2021/8854892.

[18] W. Fan *et al.*, "A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest X-ray," *Nature Communications* , vol. 15, no. 1, Dec. 2024, doi: 10.1038/s41467-024-45599-z.

[19] A. Nogueira-Rodríguez, D. Glez-Peña, M. Reboiro-Jato, and H. López-Fernández, "Negative Samples for Improving Object Detection—A Case Study in AI-Assisted Colonoscopy for Polyp Detection," *Diagnostics*, vol. 13, no. 5, Mar. 2023, doi: 10.3390/diagnostics13050966.

[20] D. Miller, G. Goode, C. Bennie, P. Moghadam, and R. Jurdak, "Why Object Detectors Fail: Investigating the Influence of the Dataset."

[21] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," Apr. 2017, [Online]. Available: http://arxiv.org/abs/1611.10012

[22] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," Dec. 2017, doi: 10.1109/CVPR.2017.369.

[23] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," 2019, *Facebook AI Research*: 0.6.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," Apr. 2017.

[26] R. Faisal, R. A. Kambau, and M. H. H, "Perhitungan Otomatis Ukuran Udang Vaname Menerapkan Object Detection dengan Convolutional Neural Network Menggunakan Metode Bidirectional Feature Pyramid Network," *INSTEK*, no. Volume 9, Apr. 2024, Accessed: Sep. 05, 2026. [Online]. Available: https://doi.org/10.24252/instek.v9i1.46420