

Non-Local Context-Aware Attention for Object Detection in Remote Sensing Images

Yassin Terraf¹ ^a, El Mahdi Mercha¹ ^b and Mohammed Erradi²

¹*HENCEFORTH, Rabat, Morocco*

²*ENSIAS, Mohammed V University, Rabat, Morocco*

Keywords: Object Detection, Deep Learning, Attention, Remote Sensing Images.

Abstract: Object detection in remote sensing images has been widely studied due to the valuable insights it provides for different fields. Detecting objects in remote sensing images is a very challenging task due to the diverse range of sizes, orientations, and appearances of objects within the images. Many approaches have been developed to address these challenges, primarily focusing on capturing semantic information while missing out on contextual details that can bring more insights to the analysis. In this work, we propose a Non-Local Context-Aware Attention (NLCAA) approach for object detection in remote sensing images. NLCAA includes semantic and contextual attention modules to capture both semantic and contextual information. Extensive experiments were conducted on two publicly available datasets, namely NWPU VHR and DIOR, to evaluate the performance of the proposed approach. The experimental results demonstrate the effectiveness of the NLCAA approach against various state-of-the-art methods.


1 INTRODUCTION


The rise and rapid development of Earth observation technology have resulted in the fast generation of high-quality remote sensing images. These images carry valuable information that can advance several real-life applications such as intelligent monitoring, urban planning, precision agriculture, and geographic information system updating (Li et al., 2020). However, manually processing and analyzing these large volumes of generated images to extract useful information is extremely difficult. Therefore, considerable efforts have been invested in developing automated systems such as object detection. Object detection focuses on detecting, localizing, and identifying objects in images. Given its significance in the analysis of remote sensing images, numerous studies have been carried out in this area. Nevertheless, these studies face challenges, such as the lack of important visual clues details which significantly constrain their performance.

Earlier approaches often are based on traditional methods with hand-crafted features like the histogram of orientation gradient (Zhang et al., 2013), scale-

invariant feature transform (Cheng and Han, 2016), and Hough transform (Xu et al., 2014). While these methods are easy to use and computationally efficient, they do not meet the desired accuracy. The main reasons for this include their limitations in adapting to varying conditions in remote sensing images and their dependence on hand-crafted features, which may not effectively capture the complex patterns of objects.

In recent years, deep learning techniques, especially convolutional neural networks (CNNs), have made significant improvements in object detection. A wide range of CNN-based object detection systems (Ren et al., 2015; Redmon et al., 2016) have been developed and they showed promising results on various datasets. Even though CNN-based methods have made substantial progress, finding objects in remote sensing images is still a big challenge due to the difficulty in capturing the salient features that are key for effective object detection. To tackle this challenge, many researchers have intended to create better feature representations to improve the performance of object detection in remote sensing. One of the commonly adopted techniques to obtain these features is through attention mechanisms, which help the model focus on important parts of the images. Most attention mechanism-based works focus on capturing effective features from objects based on semantic characteris-

^a  <https://orcid.org/0009-0004-4026-5887>

^b  <https://orcid.org/0000-0003-2034-1737>

tics, such as object shape and color, but they often overlook key clues, which are contextual information. Contextual information is essential in object detection as it highlights the interactions between objects and their surroundings, which improves the performance of object detection in remote sensing images (Perko and Leonardis, 2010; Oliva and Torralba, 2007).

Given the importance of contextual information in object detection, we introduce a new attention mechanism, Non-Local Context-Aware Attention (NLCAA). This mechanism effectively integrates both semantic characteristics of the object as well as contextual features within a unified framework. Our main contributions are as follows:

- We present a new attention mechanism that combines both semantic and contextual information into a unified framework to improve object detection in remote sensing images.
- We integrate NLCAA into the existing YOLOv3 architecture to enhance its feature detection capabilities.
- Extensive experiments have been conducted on object detection in remote sensing images using the DIOR and NWPU-VHR datasets.

The rest of this paper is organized as follows: Section II covers related works. Section III describes the proposed method. Section IV presents the implementation details and experimental results. Section V concludes the paper.

2 RELATED WORK

Attention mechanisms have become increasingly popular in the field of remote sensing object detection to improve performance. Notably, several approaches have developed specific attention modules to capture relationships between different regions or pixels in the image space, known as spatial dependencies, to enhance the accuracy of object detection. CAD-Net (Zhang et al., 2019) incorporated a spatial-and-scale-aware attention module to capture spatial dependencies by highlighting informative regions within the input data. Similarly, SCRDet (Yang et al., 2019) used a multi-dimensional attention network module to capture additional feature information, thereby enhancing the discriminative power of the model. FADet (Li et al., 2019) focused on enhancing object-related representations while suppressing background and noise information to improve object detection performance. MSA-Network (Zhang et al., 2021) integrates a multiscale module along with a channel and position attention module to concentrate on important regions

by extracting attention-based features. Furthermore, MTCNet (Wang et al., 2022b) employed the convolutional block attention module (Woo et al., 2018) to derive finer-grained features, thereby improving object detection effectiveness. However, these methods predominantly employ simpler attention techniques that focus on nearby features, learned through convolutions with local receptive fields, and do not account for relationships between distant regions. This limitation results in a missed opportunity to include global contextual information, thereby restricting their understanding of long-range dependencies.

To address existing limitations, a variety of non-local attention mechanisms have been proposed, drawing inspiration from the Non-Local Network (Wang et al., 2018). These mechanisms excel in extracting relationships from remote sensing images. In deep learning contexts, images are represented as a series of feature maps, emphasizing distinct attributes. These feature maps are further divided into channels, each focusing on specific image characteristics. While these attention mechanisms effectively capture relationships within a single channel, they face challenges in bridging interactions across different channels. Addressing this challenge, CGNL (Yue et al., 2018) introduced a methodology that facilitates interactions across these channels. However, these non-local networks operate at the pixel-to-pixel level, where each pixel is compared with every other pixel, potentially introducing noise into the feature representations. In response, (Zhang et al., 2021) introduced the Non-local Pyramid Attention (NP-Attention), which employs correlations between patches instead of pixels. These patches are generated by partitioning the feature maps into multiple subregions, typically achieved through pooling operations at varying scales. By focusing on patches rather than individual pixels, NP-Attention efficiently captures feature representations while reducing computational complexity. Although NP-Attention is effective in extracting features at the patch level, it primarily relies on the semantic characteristics of objects.

To address this limitation, we introduce a novel approach, NLCAA, that integrates both semantic characteristics and contextual information at the patch level. By including contextual information, our method better understands the relationships between objects and their surrounding environment, thereby enhancing object detection performance.

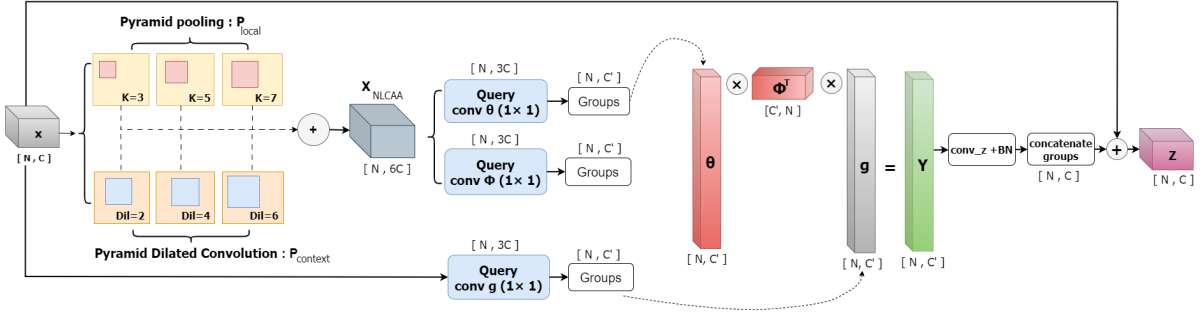


Figure 1: The architecture of the proposed NLCAA approach. The \oplus symbolizes concatenation, while the \otimes denotes the dot product operation.

3 THE PROPOSED APPROACH

In this section, we present the YOLOv3 architecture along with its components and we provide a comprehensive understanding of its operational characteristics and advantages. Furthermore, we describe the proposed NLCAA approach and we provide details about the methodology adopted to integrate it into the YOLOv3 architecture to enhance object detection in remote sensing images.

3.1 NLCAA: Yolov3-Based

In the domain of object detection in remote sensing images, selecting an appropriate framework model is crucial for validating and implementing new approaches. Widely acknowledged for its efficacy in object detection, YOLOv3 has been extensively employed in several studies, including those by Zhang et al. (Zhang et al., 2020), Wang et al. (Wang et al., 2020), and Zhang et al. (Zhang et al., 2022). The extensive use of YOLOv3 in these studies establishes it as a standard benchmark. Therefore, to evaluate the proposed NLCAA against several baseline models, we select YOLOv3 as a base framework model. This choice ensures that comparisons with the proposed NLCAA approach are consistent, providing a clear framework for evaluating the improvements of the proposed approach.

YOLOv3 introduces significant improvements over its predecessors. The key innovation lies in its ability to divide an image into a grid and predict bounding boxes and class probabilities directly, in a single pass. It consists of several key components that contribute to its object detection capabilities:

Backbone. The backbone of YOLOv3, based on the Darknet-53 architecture, is integral in extracting complex features from images. This functionality is crucial for NLCAA, which depends on detailed feature extraction to accurately detect and analyze objects in

remote sensing images. By providing an in-depth and comprehensive analysis of image data, the Darknet-53 backbone ensures that NLCAA has access to the detailed features necessary for its advanced attention mechanism.

Neck. The neck processes the features from the backbone at multiple scales, which is essential for detecting objects of different sizes. This is in line with the NLCAA objective to understand contextual relationships in images, where the scale of features plays a significant role in accurate object detection.

Head. The head of YOLOv3 is responsible for converting processed features into the final object detection outputs. This involves generating bounding boxes, class labels, and confidence scores, which are crucial for reliable detection in various remote sensing scenarios.

3.2 Non-Local Context-Aware Attention (NLCAA)

Drawing inspiration from the Non-Local Network (Wang et al., 2018), we introduce the Non-Local Context-Aware Attention (NLCAA) approach. We start with a brief review of the original non-local operation, which detects relationships between any two positions in an image. Consider a remote sensing image denoted by the feature map $X \in \mathbb{R}^{N \times C}$, where C is the number of channels and $N = HW$ is the combination of the spatial dimensions of width W and height H . The non-local operation's primary goal is to discern relationships across the entire feature map, computing the output $Y \in \mathbb{R}^{N \times C}$ as a weighted sum of features from all positions.

$$Y = f(\theta(X), \phi(X))g(X) \quad (1)$$

where $\theta(\cdot)$, $\phi(\cdot)$, and $g(\cdot)$ are learnable parameters, typically implemented using 1 by 1 convolutions. Figure 1 illustrates the global architecture of NLCAA

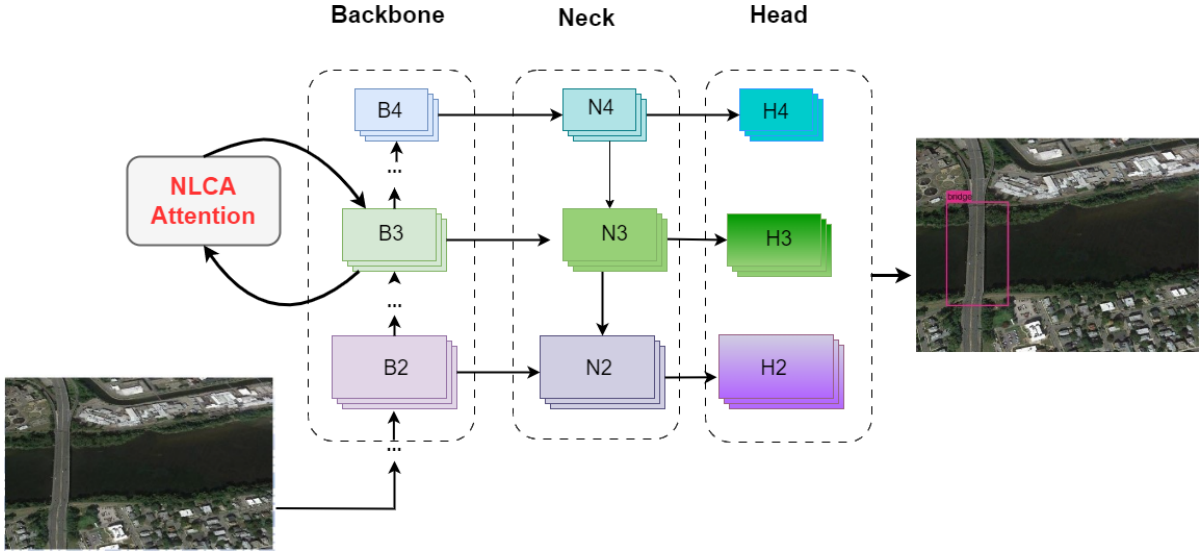


Figure 2: YOLOv3 Architecture with NLCAA Attention: Feature Map representations : Backbone B_i , Neck N_i , and Head H_i .

approach. NLCAA employs pyramid pooling with varying kernel sizes $k \in K = \{3, 5, 7\}$. Smaller kernel sizes capture fine-detail semantic information, while larger kernels capture global semantic features. These multi-scale semantic feature maps, denoted as P_{local} , are concatenated along the channel dimension:

$$P_{\text{local}} = \bigoplus_{k \in K} \text{MaxPool}_k(X) \quad (2)$$

Additionally, NLCAA uses pyramid dilated convolution on the input feature map X . Different dilation rates $d \in D = \{2, 4, 6\}$ are used, expanding the convolution filters' receptive field. This process captures multi-scale contextual information, with smaller dilation rates focusing on closer contexts and larger rates encompassing broader contexts without compromising spatial resolution. The resulting feature maps, $P_{\text{contextual}}$, are also concatenated along the channel dimension:

$$P_{\text{contextual}} = \bigoplus_{d \in D} \text{DilatedConv}_d(X) \quad (3)$$

The feature maps P_{local} and $P_{\text{contextual}}$, incorporating multi-scale semantic and contextual information, are combined to form the aggregated feature map X_{NLCAA} :

$$X_{\text{NLCAA}} = \text{Concatenate}(P_{\text{local}}, P_{\text{contextual}}) \quad (4)$$

To apply the non-local operation to X_{NLCAA} , a 1 by 1 convolution is applied to extract three fundamental components: the Query (θ), Key (ϕ), and Value (g). They are represented as:

$$\theta(X_{\text{NLCAA}})_{\text{vec}} = \text{vec}(X_{\text{NLCAA}}W_\theta) \in \mathbb{R}^{N \times 3C} \quad (5)$$

$$\phi(X_{\text{NLCAA}})_{\text{vec}} = \text{vec}(X_{\text{NLCAA}}W_\phi) \in \mathbb{R}^{N \times 3C} \quad (6)$$

$$g(X)_{\text{vec}} = \text{vec}(XW_g) \in \mathbb{R}^{N \times 3C} \quad (7)$$

Following this extraction, X_{NLCAA} is divided into separate groups using the same approach as in (Yue et al., 2018). This division enables parallel processing, thereby decreasing computational time. Each divided group performs processing through the non-local operation, generating individual outputs Y_g :

$$Y_g = f(\theta(X_{\text{NLCAA}})_g, \phi(X_{\text{NLCAA}})_g, g(X)_g) \quad (8)$$

After processing, the individual outputs Y_g are combined. The resulting merged output is normalized and then added to the original input X to generate the final feature map Y .

$$Y = \text{GroupNorm}(\text{Concatenate}(Y_1, Y_2, \dots, Y_g) + X) \quad (9)$$

The obtained Y , enriched with multi-scale semantic and contextual features details.

To effectively integrate NLCAA into the YOLOv3 architecture and improve its object detection capabilities, we have chosen a strategic approach. As shown in Figure 2, instead of applying NLCAA to the initial or final feature maps, we incorporate it into the intermediate feature maps for several reasons. Intermediate feature maps inherently capture a higher level of semantic information compared to initial feature maps, which is crucial for effective object detection. Furthermore, applying NLCAA to the initial feature maps can be computationally expensive, especially when dealing with high-dimensional input

data. Conversely, intermediate feature maps typically have lower dimensionality, allowing for more efficient computation of attention mechanisms. Additionally, the final feature maps in a neural network often exhibit high spatial resolution but may contain significant noise or irrelevant information. Applying NLCAA to these feature maps can lead to incorrect results, a loss of model stability, or overfitting. In comparison, intermediate feature maps are better poised to contain relevant information without the interference of excessive noise. By integrating NLCAA into these intermediate feature maps, we enhance the YOLOv3 architecture's contextual understanding, reduce computational complexity, and prioritize pertinent information, thereby improving its capabilities in object detection in remote sensing images.

3.3 Loss Function

The loss function used in our object detection system is a composite loss function, which is expressed through the following equation:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{loc}} \mathcal{L}_{\text{loc}} + \lambda_{\text{obj}} \mathcal{L}_{\text{obj}} + \lambda_{\text{noobj}} \mathcal{L}_{\text{noobj}} + \lambda_{\text{class}} \mathcal{L}_{\text{class}} \quad (10)$$

where \mathcal{L}_{loc} , \mathcal{L}_{obj} , $\mathcal{L}_{\text{noobj}}$, $\mathcal{L}_{\text{class}}$ represent respectively the localization loss, objectness loss, no-objectness loss, and the classification loss. \mathcal{L}_{loc} is the localization loss, this component evaluates the accuracy of the predicted bounding boxes. For every detected object, the predicted bounding box is characterized by its center coordinates (x, y) and its dimensions (w, h) . The mathematical representation is:

$$\mathcal{L}_{\text{loc}} = \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] \quad (11)$$

Here, (x, y, w, h) denote the predictions, and $(\hat{x}, \hat{y}, \hat{w}, \hat{h})$ are the actual values. The indicator 1_{ij}^{obj} confirms the presence of an object in cell i for bounding box j . \mathcal{L}_{obj} is the confidence of object presence loss, this component measures the confidence regarding the presence of objects within the bounding boxes. The formula is:

$$\mathcal{L}_{\text{obj}} = \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{\text{obj}} (C_i - \hat{C}_i)^2 \quad (12)$$

where S represents the grid size, determining the number of cells along each dimension in the image grid, and B denotes the number of predicted bounding boxes for each grid cell. For bounding boxes that

do not contain any objects, another no-objectness loss is considered:

$$\mathcal{L}_{\text{noobj}} = \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{ij}^{\text{noobj}} (C_i - \hat{C}_i)^2 \quad (13)$$

In both equations, C_i represents the predicted confidence, while \hat{C}_i denotes the true confidence. $\mathcal{L}_{\text{class}}$ is the classification loss that classifies the detected entities. The classification loss measures how accurately the approach used to classify objects manages this. It is represented as:

$$\mathcal{L}_{\text{class}} = \sum_{i=0}^{S^2} 1_i^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \quad (14)$$

Here, $p_i(c)$ is the predicted probability of class c for cell i , and $\hat{p}_i(c)$ is the true probability.

4 EXPERIMENTS

This section validates the effectiveness of the proposed NLCAA approach through experiments on two public remote sensing datasets. It demonstrates the impact of incorporating contextual and semantic information in remote sensing object detection using NLCAA.

4.1 Datasets

4.1.1 NWPU-VHR10

The NWPU-VHR10 dataset (Cheng et al., 2014), consists of a total of 800 very-high-resolution remote sensing images, of which 650 are positive samples containing various objects of interest, and 150 are negative samples that do not contain any objects of interest. The dataset provides annotations for ten types of objects, namely plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. The objects of interest are annotated using publicly accessible horizontal bounding boxes (HBB).

4.1.2 DIOR

The DIOR dataset (Li et al., 2020), is a comprehensive aerial image dataset designed for object detection, consisting of 23,463 images with 190,288 instances. All images in the dataset have a size of 800×800 pixels. The dataset is divided into three subsets: 5,862 images for training, 5,863 images for validation, and 11,738 images for testing. It encompasses 20 prevalent object categories: Airplane (c1), Airport

(c2), Baseball field (c3), Basketball court (c4), Bridge (c5), Chimney (c6), Dam (c7), Expressway service area (c8), Expressway toll station (c9), Golf course (c10), Ground track field (c11), Harbor (c12), Overpass (c13), Ship (c14), Stadium (c15), Storage tank (c16), Tennis court (c17), Train station (c18), Vehicle (c19), and Windmill (c20).

4.2 Settings

For our experiments, two datasets were primarily considered: NWPU-VHR10 and DIOR. The NWPU-VHR10 dataset does not come with a predefined training and testing split. To address this, we used a random sampling technique. Specifically, 75% of the positive images were allocated to the training set, with the remainder reserved for testing. In the case of the DIOR dataset, we conformed to its existing training/testing split. For training hyperparameters, after performing fine-tuning, we obtained an initial learning rate of 1×10^{-4} , a final learning rate of 1×10^{-6} , a cosine update strategy for the learning rate, a weight decay of 5×10^{-4} , a momentum of 0.9, a maximum training epoch of 180, an Intersection over Union (IoU) threshold of 0.5, a confidence threshold of 0.05, a Non-Maximum Suppression (NMS) threshold of 0.5, and pretraining on the ImageNet dataset. By controlling aspects such as the rate of parameter updates, regularization, convergence, and post-processing, these hyperparameters ensure consistency in the training process. All experiments in this study were conducted on a computer equipped with an Intel Xeon(R) CPU running at 2.20 GHz, 83.48 GB of memory, and an NVIDIA A100-SXM4-40GB GPU with 40 GB of memory, facilitating accelerated computations.

4.3 Baseline

In our study, we adopted the YOLOv3 model (Redmon and Farhadi, 2018) as the foundational benchmark for assessments. YOLOv3, a notable one-stage detector, stands as a reference in the object detection domain. The architecture’s backbone relies on Darknet-53, a convolutional network designed specifically for high-performance object detection tasks. The strength of Darknet-53 lies in its ability to capture hierarchical representations that are essential for detailed object detection. Following the backbone, the architecture employs the Feature Pyramid Network (FPN) as its neck. FPN enhances the model’s capability by integrating high-level semantic features with lower-level features, ensuring accurate object localization across different scales. The head component

of YOLOv3 is responsible for final object classification and bounding box regression, retaining the architecture’s inherent accuracy in object detection and localization.

4.4 Evaluation Metrics

To evaluate the effectiveness of our introduced architecture, we used the Mean Average Precision (mAP) metric, a commonly used measure in object detection. The mAP provides a comprehensive performance assessment by combining the Average Precision (AP) across various object categories. The mathematical exposition is presented as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (15)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (16)$$

$$\text{AP} = \int_0^1 \text{precision}(r) d(\text{recall}(r)) \quad (17)$$

$$\text{mAP} = \frac{1}{C} \sum_{i=1}^C \text{AP}_i \quad (18)$$

In Equations (15) and (16), Precision and Recall are detailed. Equation (17) delineates the AP, which is essentially the area under the precision-recall curve. Finally, Equation (18) computes the mAP by averaging the AP over all object classes, where C denotes the count of object classes. A high mAP score is indicative of strong detection capabilities, signifying highly effective performance across all object classes.

4.5 Experimental Results

After evaluating YOLOv3 with NLCAA on the NWPU-VHR and DIOR datasets, we have summarized the performance results in Table 1 and 2. Table 1 presents the experimental results from the DIOR dataset test set and offers a comparison with state-of-the-art methods. The achieved results show that the proposed NLCAA significantly exceeds the vanilla YOLOv3 baseline with an improvement of +21.2% in mAP. This notable gain underscores the efficacy of integrating NLCAA into the YOLOv3 baseline, markedly enhancing its object detection performance in remote sensing images. Moreover, NLCAA exceeds the performance of several state-of-the-art methods: it achieves a +24.2% mAP compared to Faster R-CNN, a +13.1% mAP against Mask R-CNN, and outperforms NPMMR-De by +0.9% mAP.

A deeper dive into individual object performances demonstrates that NLCAA exceeds other methods

Table 1: Comparative analysis on DIOR dataset.

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	mAP
Faster R-CNN (Ren et al., 2015)	53.6	53.2	78.8	66.2	28.0	70.9	62.3	69.0	55.2	68.0	56.9	50.2	27.7	73.0	39.8	75.2	38.6	23.6	45.4	54.1	
Faster R-CNN with FPN(Lin et al., 2017a)	54.0	74.5	63.3	80.7	44.8	72.5	60.0	75.6	62.3	76.0	76.8	46.4	57.2	71.8	68.3	53.8	81.1	59.5	43.1	81.2	65.1
Mask-RCNN (He et al., 2017)	53.9	76.6	63.2	80.9	40.2	72.5	60.4	76.3	62.5	76.0	75.9	46.5	57.4	71.8	68.3	53.7	81.0	62.3	43.0	81.0	65.2
SSD (Liu et al., 2016)	53.9	76.6	63.2	80.9	40.2	72.5	60.4	76.3	62.5	76.0	75.9	46.5	57.4	71.8	68.3	53.7	81.0	62.3	43.0	81.0	65.2
YOLOv3 (Redmon and Farhadi, 2018)	72.2	29.2	74.0	78.6	31.2	69.7	26.9	48.6	54.4	31.1	61.1	44.9	49.7	87.4	70.6	68.7	87.3	29.4	48.3	78.7	57.1
RetinaNet (Lin et al., 2017b)	53.3	77.0	69.3	85.0	44.1	73.2	62.4	78.6	62.8	78.6	76.6	49.9	59.6	71.1	68.4	45.8	81.3	55.2	45.1	85.5	66.1
PA-Net (Wang et al., 2019)	60.2	72.0	70.6	80.5	43.6	72.3	61.4	72.1	66.7	72.0	73.4	45.3	56.9	71.7	70.4	62.0	80.9	57.0	47.2	84.5	66.1
CornetNet (Law and Deng, 2018)	58.8	84.2	72.0	80.8	46.4	75.3	64.3	81.6	76.3	79.5	79.5	26.1	60.6	37.6	70.7	45.2	84.0	57.1	43.0	75.9	64.9
CSFF (Law and Deng, 2018)	57.2	79.6	70.1	87.4	46.1	76.6	62.7	82.6	73.2	78.2	81.6	50.7	59.5	73.3	63.4	58.5	85.9	61.9	42.9	86.9	68.0
SCRDet++ (Yang et al., 2022)	71.9	85.0	79.5	88.9	52.3	79.1	77.6	89.5	77.8	84.2	83.1	64.2	65.6	71.3	76.5	64.5	88.0	70.9	47.1	85.1	75.1
NPMMR-Det (Huang et al., 2021)	88.1	87.1	82.0	92.6	49.6	80.6	73.4	83.7	70.5	85.0	84.8	64.8	63.3	91.6	71.3	76.1	92.0	69.1	58.6	83.6	77.4
Ours(NLCAA)	85.8	88.4	80.8	92.7	53.1	82.8	73.3	91.2	79.8	85.2	83.3	66.5	63.3	90.9	70.6	77.4	92.0	62.4	57.8	89.8	78.3

Class names: Airplane (C1), Airport (C2), Baseball field (C3), Basketball court (C4), Bridge (C5), Chimney (C6), Dam (C7), Expressway service area (C8), Expressway toll station (C9), Golf course (C10), Ground track field (C11), Harbor (C12), Overpass (C13), Ship (C14), Stadium (C15), Storage tank (C16), Tennis court (C17), Train station (C18), Vehicle (C19), Windmill (C20)

Table 2: Comparative analysis on NWPU-VHR dataset.

Methods	PL	ship	ST	BD	TC	BC	GTF	harbor	bridge	vehicle	mAP
COPD (Cheng et al., 2014)	62.3	69.4	64.5	82.1	34.1	35.3	84.2	56.3	16.4	44.3	54.9
Faster RCNN(Ren et al., 2015)	90.9	86.3	90.5	98.2	89.7	69.6	100	80.1	61.5	78.1	84.5
Transferred CNN (Krizhevsky et al., 2012)	66.0	57.1	85.0	80.9	35.1	45.5	79.4	62.6	43.2	41.3	59.6
RICNN (Cheng et al., 2016)	88.7	78.3	86.3	89.1	42.3	56.9	87.7	67.5	62.3	72.0	73.1
YOLOv3 (Redmon and Farhadi, 2018)	95.4	87.1	70.9	99.1	73.2	81.2	96.2	85.6	60.6	56.1	80.5
Li et al. (Li et al., 2017)	99.7	90.8	90.6	92.2	90.3	80.1	90.8	80.3	68.5	97.1	87.1
CAD-Net (Zhang et al., 2019)	97.0	77.9	95.6	93.6	87.6	87.1	99.6	100	86.2	89.9	91.5
NPMMR-Det (Huang et al., 2021)	99.8	93.7	96.6	99.6	96.2	96.8	100	95.9	71.7	98.1	94.83
Ours(NLCAA)	99.78	94.5	96.7	99.0	96.0	91.0	100	97.6	89.1	97.7	96.15

Class names: PL (plane), ship, ST (storage tank), BD (baseball diamond), TC (tennis court), BC (basketball court), GTF (ground track field), harbor, bridge, vehicle.

across multiple categories, especially in detecting the airport, basketball court, bridge, chimney, expressway service area, expressway toll station, harbor, golf course, overpass, storage tank, and tennis court. A common thread among these objects is that they are not only characterized by semantic features such as shape, color, size, and texture but also by their contextual settings. For instance, bridges in the DIOR dataset are often pictured with rivers below, harbors are accompanied by seas, and airports often have airplanes nearby. This underlines that these objects are defined both by their semantic and contextual features. The superior performance of NLCAA in detecting these objects highlights its capability to capture both the semantic and contextual features, distinguishing it from other methods.

Table 2 continues this evaluation, detailing our tests on the NWPU-VHR10 dataset and benchmarking NLCAA against other state-of-the-art models. The results in Table 2 further validate the effectiveness of NLCAA. It outperforms the vanilla YOLOv3 baseline by +15.65% mAP, outdoes Faster R-CNN by +11.65% mAP, and edges out NPMMR-De by +1.31% mAP. Similar to its performance on the DIOR dataset, NLCAA exhibits its effectiveness at detecting and classifying several objects when benchmarked against other leading methods. For example, it excels in identifying ships, bridges, tennis courts, among others. In this context, the ability of NLCAA to capture both semantic and contextual features underscores its importance in the task of object detection in remote sensing images. Some of the results obtained

by integrating NLCAA into the YOLOv3 architecture are depicted in Figure 3.

However, while NLCAA is integrated into the YOLOv3 architecture as the base model for our approach, facilitating benchmarking and comparisons with state-of-the-art methodologies, its design is inherently modular. The pyramid pooling and dilated convolution components of NLCAA are designed to be compatible with various versions of YOLO architectures, including YOLOv4 (Bochkovskiy et al., 2020), YOLOv7 (Wang et al., 2022a), YOLOv8 (Reis et al., 2023), and YOLOX (Ge et al., 2021). These versions demonstrate enhanced detection performance compared to YOLOv3. Integrating NLCAA into these advanced architectures could further augment their capabilities in detecting objects from remote sensing images. Despite these advancements, the NLCAA approach to object detection, with its focus on extracting both semantic and contextual features, faces certain limitations. For example, in scenarios where objects such as vehicles are primarily defined by their semantic attributes rather than contextual ones, NLCAA encounters specific challenges. These challenges highlight areas for potential refinement in NLCAA performance.

5 CONCLUSION

In conclusion, this study presents NLCAA, a novel approach specifically designed for object detection in remote sensing images. The evaluations on the DIOR



Figure 3: Examples of object detection using YOLOv3 enhanced by NLCAA in diverse remote sensing images.

and NWPU VHR datasets revealed that NLCAA consistently outperformed several existing methods. The strength of NLCAA lies in its capability to integrate both detailed and broader contextual features of objects, enhancing its detection accuracy. As we progress, our objective is to further refine the proposed approach. Specifically, we will focus on optimizing the Neck and Head components of YOLOv3 to enrich its feature extraction ability.

REFERENCES

- Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.
- Cheng, G. and Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 117:11–28.
- Cheng, G., Han, J., Zhou, P., and Guo, L. (2014). Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98:119–132.
- Cheng, G., Zhou, P., and Han, J. (2016). Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415.
- Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.
- He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.
- Huang, Z., Li, W., Xia, X.-G., Wu, X., Cai, Z., and Tao, R. (2021). A novel nonlocal-aware pyramid and multiscale multitask refinement detector for object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Law, H. and Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750.
- Li, C., Xu, C., Cui, Z., Wang, D., Zhang, T., and Yang, J. (2019). Feature-attentioned object detection in remote sensing imagery. In *2019 IEEE international confer-*

- ence on image processing (ICIP), pages 3886–3890. IEEE.
- Li, K., Cheng, G., Bu, S., and You, X. (2017). Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2337–2348.
- Li, K., Wan, G., Cheng, G., Meng, L., and Han, J. (2020). Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307.
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer.
- Oliva, A. and Torralba, A. (2007). The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527.
- Perko, R. and Leonardis, A. (2010). A framework for visual-context-aware object detection in still images. *Computer Vision and Image Understanding*, 114(6):700–711.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.
- Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
- Reis, D., Kupec, J., Hong, J., and Daoudi, A. (2023). Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Wang, C., Bochkovskiy, A., and Liao, H. (2022a). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv 2022. arXiv preprint arXiv:2207.02696*.
- Wang, J., Xiao, W., and Ni, T. (2020). Efficient object detection method based on improved yolov3 network for remote sensing images. In *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 242–246. IEEE.
- Wang, K., Liew, J. H., Zou, Y., Zhou, D., and Feng, J. (2019). Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206.
- Wang, W., Tan, X., Zhang, P., and Wang, X. (2022b). A cbam based multiscale transformer fusion approach for remote sensing image change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:6817–6825.
- Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803.
- Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.
- Xu, J., Sun, X., Zhang, D., and Fu, K. (2014). Automatic detection of inshore ships in high-resolution remote sensing images using robust invariant generalized hough transform. *IEEE geoscience and remote sensing letters*, 11(12):2070–2074.
- Yang, X., Yan, J., Liao, W., Yang, X., Tang, J., and He, T. (2022). Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2384–2399.
- Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., Sun, X., and Fu, K. (2019). Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8232–8241.
- Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., and Xu, F. (2018). Compact generalized non-local network. *Advances in neural information processing systems*, 31.
- Zhang, G., Lu, S., and Zhang, W. (2019). Cad-net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):10015–10024.
- Zhang, G., Xu, W., Zhao, W., Huang, C., Yk, E. N., Chen, Y., and Su, J. (2021). A multiscale attention network for remote sensing scene images classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:9530–9545.
- Zhang, L.-g., Wang, L., Jin, M., Geng, X.-s., and Shen, Q. (2022). Small object detection in remote sensing images based on attention mechanism and multi-scale feature fusion. *International Journal of Remote Sensing*, 43(9):3280–3297.
- Zhang, S., Mu, X., Kou, G., and Zhao, J. (2020). Object detection based on efficient multiscale auto-inference in remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 18(9):1650–1654.
- Zhang, W., Sun, X., Fu, K., Wang, C., and Wang, H. (2013). Object detection in high-resolution remote sensing images using rotation invariant parts based model. *IEEE Geoscience and Remote Sensing Letters*, 11(1):74–78.