Mathematical Biosciences and Engineering

MBE, 20(8): 15219–15243.
DOI: 10.3934/mbe.2023681
Received: 06 February 2023
Revised: 09 May 2023
Accepted: 31 May 2023
Published: 20 July 2023

http://www.aimspress.com/journal/MBE

Research article

Detection and localization of multi-scale and oriented objects using an enhanced feature refinement algorithm

Deepika Roselind Johnson[1,*] and Rhymend Uthariaraj Vaidhyanathan[2]

[1] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamilnadu, India
[2] Ramanujan Computing Centre, Anna University, Chennai, Tamilnadu, India

* **Correspondence:** Email: deepikaroselind.j@vit.ac.in.

Abstract: Object detection is a fundamental aspect of computer vision, with numerous generic object detectors proposed by various researchers. The proposed work presents a novel single-stage rotation detector that can detect oriented and multi-scale objects accurately from diverse scenarios. This detector addresses the challenges faced by current rotation detectors, such as the detection of arbitrary orientations, objects that are densely arranged, and the issue of loss discontinuity. First, the detector also adopts a progressive regression form (coarse-to-fine-grained approach) that uses both horizontal anchors (speed and higher recall) and rotating anchors (oriented objects) in cluttered backgrounds. Second, the proposed detector includes a feature refinement module that helps minimize the problems related to feature angulation and reduces the number of bounding boxes generated. Finally, to address the issue of loss discontinuity, the proposed detector utilizes a newly formulated adjustable loss function that can be extended to both single-stage and two-stage detectors. The proposed detector shows outstanding performance on benchmark datasets and significantly outperforms other state-of-the-art methods in terms of speed and accuracy.

Keywords: object detection; single-stage rotation detection; feature refinement; oriented object detection; progressive approach; loss discontinuity

1. Introduction

Object detection is a trending research area used in various applications. It is widely used for surveillance, scene analysis, autonomous driving, real-time tracking, etc. An efficient detection, tracking, and recognition framework consist of the following components—detection, localization, classification, tracking and action detection. Object localization and classification are challenging as it is difficult to locate objects present in dense and cluttered scenes. Also, many of the benchmark datasets restrict object orientations. For example, in sports videos, an athlete's position (such as diving or gymnastics) is oriented either vertically or horizontally. To overcome these disadvantages, the model is trained with different rotation angles to learn the distinguishing features to improve its localization and classification accuracy. Therefore, an efficient technique is required to localize and detect both arbitrarily oriented and multi-scale objects. In this paper, an improved single-stage rotation detector for fast and accurate detection of oriented and multi-scale objects is proposed. The proposed single-stage rotation detector aims to address the following challenges – arbitrary orientations, densely arranged objects, and discontinuity of loss. The contribution of the proposed work aims to address the feature angle problems in single-stage detectors, which have a significant impact on the accuracy of classification and regression.

Existing detectors such as DFN [1], S-SRNN [2], PA-SSD [3], SRN [4] and RAMS-CNN [5] are designed to detect objects in a specific orientation, usually upright. When objects are presented in arbitrary orientations, it can be more challenging for the model to detect them accurately. Some models are designed to handle a range of orientations, but even these can struggle with extreme cases. In general, object detection models perform better when objects are presented in a consistent orientation. When objects are densely packed together, it can be difficult for object detection models [1–5] to distinguish them from one another. This is especially true when the objects are similar in shape or color. The model may mistake one object for another or fail to detect some objects altogether. Object detection models are typically trained using a loss function [6] that penalizes the model for making incorrect predictions. In some cases, this loss function can be discontinuous, meaning that a small change in the model's parameters can result in a huge loss. This can make it difficult for the model to learn effectively, as small changes in the parameters may not result in a corresponding improvement in the loss function. As a result, the model may struggle to converge to an optimal solution.

To overcome the existing challenges, the proposed single-stage rotation detector introduces a feature refinement module. The feature refinement module is designed to minimize feature angle problems encountered in current detectors. It uses a feature interpolation technique to obtain positional data corresponding to refined anchors. By reconstructing the entire feature map pixel by pixel, it reduces feature angulation issues and the number of bounding boxes generated. Also, this work is the first attempt to address the problem of feature angulation in rotated detectors. To effectively handle different scenarios, the proposed detector uses a progressive regression approach from coarse to fine-grained. The first phase uses horizontal anchors for faster object detection and a higher recall rate in cluttered backgrounds. During the second phase, the refined rotating anchors are used in subsequent refinement levels to adapt to more intense scenarios for oriented object detection. This approach leverages the strengths of both horizontal and rotating anchors. Another important contribution is the formulation of a tunable loss function to address the loss disruption caused by rotational sensitivity error (RSE), which is specific to rotation-based detectors. The proposed loss function is extended to both single-stage and two-stage detectors, thus improving their performance. Finally, the proposed

single-stage rotation detector is evaluated on publicly available benchmark datasets and demonstrates improved performance. By effectively addressing the feature angulation problems, utilizing a progressive regression approach, and formulating an adjustable loss function, the proposed detector achieves improved accuracy in object detection tasks.

2. Related work

Object detection is a fundamental aspect of computer vision, with numerous generic object detectors proposed by various researchers. Current object detectors are categorized as single-stage and two-stage object detectors. A single-stage detector predicts all bounding boxes at once and requires only one pass through the neural network. It has a high reference speed because it overrides the range suggestion in two-stage detectors. A two-stage detector consists of two stages—region of interest (RoI) extraction and classification. First, the detector proposes an RoI using a selective search approach, and the regions are pooled. Second, a classifier processes the candidate regions for accurate classification and recognition.

Visual object detection is a popular topic that witnessed immense progress in recent years. Some of the prominent single-stage approaches are Overfeat [10], single stage detector (SSD) [11] and YOLO [12]. Similarly, R-CNN [6], fast RCNN [7], R-FCN [8] and R-CNN [9] are some of the prominent two-stage approaches. Multiscale feature fusion techniques are widely used in both single-stage and two-stage approaches, including feature pyramid network [13], RetinaNet [2] and DSSD [14]. A few cascaded or sophisticated detectors have recently been proposed. For instance, two-stage detectors such as cascade RCNN [15], HTC [16] and FSCascade [17] perform numerous classifications and regressions, resulting in noticeable improvements in accuracy in terms of localization and classification. In addition, anchor-free methods such as FCOS [18], FoveaBox [19] and RepPoints [20] are gaining popularity. By removing anchors, the structures of these detectors can be simplified, and anchor-free techniques have thus opened new possibilities for object detection. However, the above detectors only produce horizontal bounding boxes, which limits their usefulness in many situations encountered in the real world. Objects that are often closely spaced and have large aspect ratios in scene text and aerial photography require more accurate localization. As a result, rotated object detection has gained popularity in recent years.

Rotated object detectors gained popularity as they were needed for detecting objects in real scenarios such as natural scenes [21], aerial photos [22] and videos. These detectors typically describe the positions of objects using rotated bounding boxes, which are more accurate than horizontal boxes. Numerous detectors have been proposed for detecting text, medical images, and emotions [48–57]. For example, RRPN uses rotating anchors to refine region suggestions. Similarly, R^2CNN [23] is a detector used to detect both horizontal and rotated text appearing in natural scenes. To accommodate elongated texts, TextBoxes++ [25] used an approach that increased the number of region suggestions and used a long convolution kernel. The ICN approach [25] combines several modules including image pyramid, feature pyramid network, and deformable inception sub-networks to achieve satisfactory results on the DOTA benchmark dataset. Likewise, rotationally invariant features are extracted by the RoI Transformer [44] to improve subsequent classification and regression. The SCRDet [23] proposes a smooth IoU-based L1 loss function to handle the sudden loss change caused by feature angulation issues when it comes to handling small, dense, and rotated objects.

Moreover, the detection and localization of multi-scale and oriented structures is a key challenge

in many medical image analysis tasks, such as the detection of tumors, segmentation of blood vessels, and localization of anatomical landmarks. In recent years, there has been a lot of research on developing methods that can automatically detect and localize such structures in medical images. Many deep learning-based methods have been widely used for the detection and localization of multi-scale and oriented structures in medical images. These methods use convolutional neural networks (CNNs) to learn features from images at different scales and orientations. Some popular deep learning-based methods for this medical image analysis include Faster R-CNN, A-CNN [52], T-GAN [53] and YOLO [55] with the research area gaining more prominence in recent years.

However, the above-mentioned approaches do not consider the problem of loss discontinuity. Discontinuity of loss can affect the stability of the learning model and influence the detection results. The fundamental issue that drives this proposed work has not yet been addressed by any existing studies. Furthermore, object detection in sports images or images captured in diverse camera angles is more challenging. Similarly, the main difficulties are reflected in complex backgrounds, camera angles, dense backgrounds, and the presence of numerous small objects. However, no research has been proposed in detecting oriented objects (poses such as gymnastics and diving) obtained from sports videos, camera angles (CCTV or overhead fisheye camera), etc.

3. Proposed work

The architecture of the proposed single-stage rotation detector is shown in Figure 1 and the overview of the process is shown in Algorithm 1. To continuously improve the features of the estimated bounding boxes to improve detection and localization accuracy, multiple levels of feature filtering are added to the network. Finally, the enhanced feature filtering module is introduced in the last filtering stage for an effective feature map reconstruction.

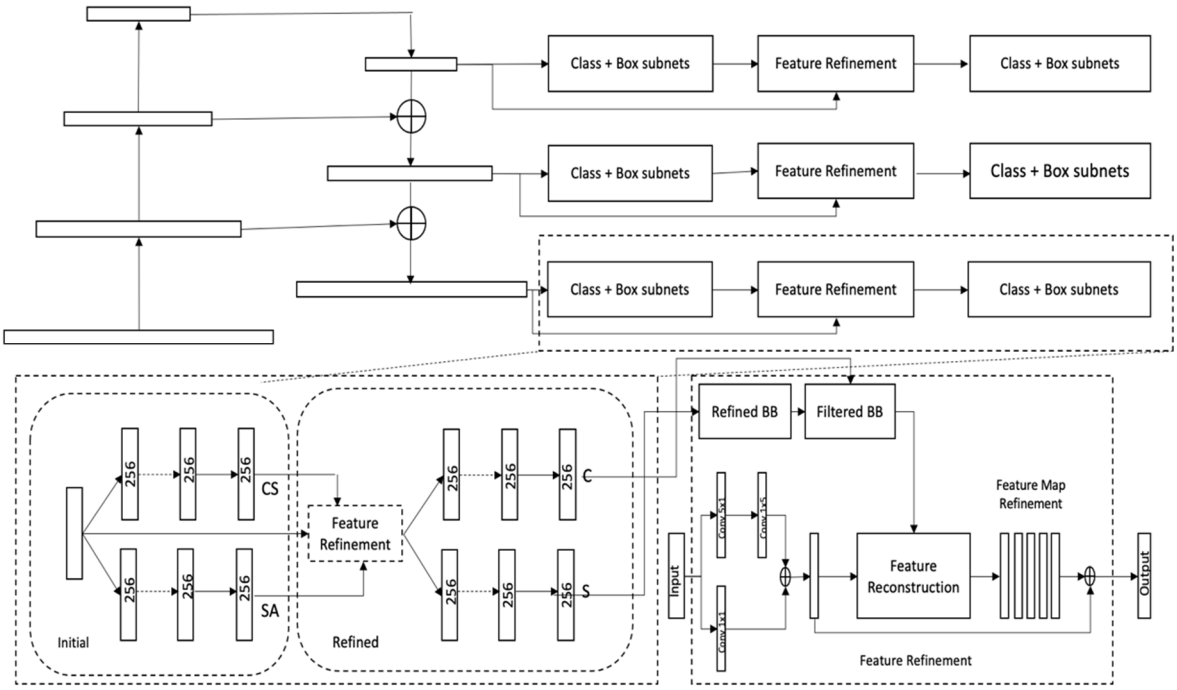


Figure 1. The architecture of the proposed single-stage rotation detector.

3.1. Parametrization for oriented object estimation

To perform oriented object detection, the following five parameters(x, y, h, w, θ) are introduced to denote the oriented bounding box. The value of θ is between $[-\frac{\pi}{2}, 0)$ and used to predict the offset angle (regression subnetwork). The bounding boxes for detecting oriented objects are specified as given in Eq (1), it is used to calculate the coordinates, width, height and angle of the ground truth box.

$$b_x = (\frac{x-x_a}{w_a}),\ \ b_y = (\frac{y-y_a}{h_a}),\ b_w = \log\left(\frac{w}{w_a}\right), b_h = \log\left(\frac{h}{h_a}\right), b_\theta = (\theta - \theta_a) \quad (1)$$

Equation (2) is used to calculate the coordinates, width, height, and angle of the estimated bounding box.

$$b'_x = (\frac{x'-x_a}{w_a}), b'_y = (\frac{y'-y_a}{h_a}),\ \ b'_w = \log\left(\frac{w'}{w_a}\right),\ \ b'_h = \log\left(\frac{h'}{h_a}\right), b'_\theta = (\theta' - \theta_a) \quad (2)$$

where x, y are the centre coordinates and the ground truth box, w denotes width, h denotes height, and θ is the offset angle. Also, x_a denotes the ground truth box and x' denotes the estimated bounding box.

3.2. Feature angulation loss function estimation

Each bounding box consists of a centre, width, and height, as shown in Figure 2. The angle θ varies between the ground truth and the estimated bounding box for the same aspect ratio. Thus, the smooth *L1* loss (combination of *L1* loss and *L2* loss) of both bounding boxes (ground truth and prediction) remains the same, but the angular loss varies. As shown in Figure 2, the green and orange colours denote the inconsistency between the smooth *L1* loss and the angle loss (A_{IoU}), making the previous loss function unsuitable for detecting oriented objects. For example, oriented objects with a large aspect ratio are very sensitive to changes in angle.

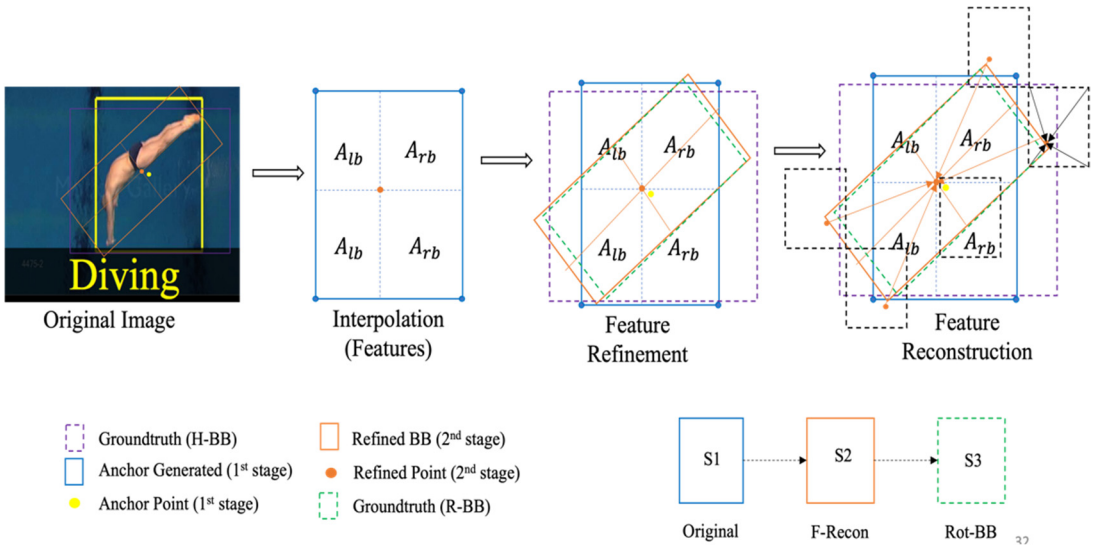


Figure 2. Enhanced feature filtering module with feature angulation analysis.

Some of the loss functions used in traditional bounding boxes are G_{IoU} [26] and D_{IoU} [27]. These loss parameter uses a regression function to overcome the limitations of a traditional bounding box. In order to estimate the angle loss (A_{IoU}) for oriented objects, a new loss function is proposed. SCRDet [23] serves as inspiration for obtaining the derivable loss function as follows.

$$Loss = \frac{\alpha_1}{N}\sum_{n=1}^{N} o_n \frac{Loss_{reg}(r'_n, r_n)}{|Loss_{reg}(r'_n, r_n)|}.|f(A_{IoU})| + \frac{\alpha_2}{N}\sum_{n=1}^{N} Loss_{cls}(d_n, b_n) \quad (3)$$

$$Loss_{reg}(r', r) = Loss_{smooth-l1}(r'_\theta, r_\theta) - IoU(r'_{(x,y,w,h)}, r_{(x,y,w,h)}) \quad (4)$$

where N is the anchor, o_n specifies the foreground and background objects, r denotes the ground truth target vector, r' specifies the offset vector for the predicted box. The probability of the distance between classes is provided by d_n and b_n specifies the label of the object. A_{IoU} represents the overlap between the estimated field and the actual data indicated by A_{IoU}, α_1 and α_2 specify the control trade-off between the hyper parameters and their values are set to 1. Finally, the term $f(\cdot)$ specifies the loss function related to A_{IoU} and $IoU(\cdot)$ denotes the IoU calculation function of the traditional bounding box.

From Eq (3) it can be observed that $Loss_{reg}$ (regression loss) is further categorized into two components that are used to estimate the gradient direction (derivable term). Similarly, $|f(A_{IoU})|$ is used to calculate the loss term associated with the gradient and amplitude (non-derivable term). Using Eq (3), the $Loss_{reg}$ is used as the dominant function to address the instability between the smooth $L1$ and A_{IoU}. As already mentioned, A_{IoU} is sensitive to changes in angle. For example, a minimal change in angle affects the IoU (intersection over union) score, as shown in Figure 2. Thus, by optimizing the estimated bounding box, the recall rate of the proposed single-stage rotation detector is improved. Multiple feature refinement phases with different IoU thresholds are used during training to improve detection accuracy. Additionally, the F_{IoU} (foreground threshold) is fixed at 0.5 and B_{IoU} is set to 0.4 during the initial testing phase. Subsequently, as more refinement phases are added, the F_{IoU} is fixed at 0.6 and 0.7, while B_{IoU} set to 0.5 and 0.6, respectively. Thus, the overall loss function for the proposed single stage detector is defined in Eq (5), where, $Loss_i$ provides the loss term defined for the i^{th} stage, while β_i denotes the bias coefficient.

$$Loss_t = \sum_{i=1}^{N}(\beta_i, Loss_i) \quad where \;\; \beta_i = 1 \quad (5)$$

3.3. Improved feature filtering mechanism

Current detectors [1,3–5,29] uses a single feature map to perform multi-level classification and regression. However, these detectors do not consider the angulation problems of the features that arise due to the inconsistency between the ground truth and the estimated boxes, as shown in Figure 2. The feature angle problem affects the detection accuracy of the detector due to detection of false features. These erroneous features affect the detection accuracy for objects with large aspect-ratio. The proposed single-stage rotation detector constructs the entire feature map to minimize feature angulation problems. As shown in Figure 2, the feature map is constructed by re-encoding the position of the estimated bounding box obtained during the feature filtering phase using feature interpolation technique in Eq (6).

$$fm = (fm_{lt} * A_{rb} + fm_{rt} * A_{lb} + fm_{rb} * A_{lt} + fm_{lb} * A_{rt}) \quad (6)$$

where the term fm indicates a feature vector and A denotes the areas on the feature map, as shown in Figure 2. In Eq (6), a feature refinement mechanism is proposed and the pseudocode is specified in Algorithm 1. Using the convolution operation, the feature maps are added to obtain new features. During the refinement phase, the estimated bounding box with the highest confidence value is considered for processing. The detector must also ensure that each feature point on the map corresponds only to one bounding box obtained during the feature refinement phase. For each point (feature) on the feature map, the corresponding feature vector is extracted along with its coordinates (lt, rb, rt, lb, c) to perform feature recovery. The variables lt, rb, rt and lb denote the four vertices and c denotes the midpoint of the estimated bounding box. Finally, the points (features) are traversed sequentially to regenerate the refined feature map.

Algorithm 1. Improved feature refinement

- **Input:** Feature map (fm), bounding box (b), and confidence (c)
- **Output:** Reconstructed feature map (fm')

$B' \leftarrow BF(b, c)$ // BB with the highest score for each trait point is kept in the refinement phase to increase speed
$h, w \leftarrow ShapeFeatures(fm), fm' \leftarrow zeros_{like(fm)}$
$fm \leftarrow conv_{1\times1}(fm) + conv_{1\times5}(conv_{5\times1}(fm))$ //Two-way convolution to get a new feature
for $i \leftarrow 0$ **to** $h - 1$ **do**
  for $j \leftarrow 0$ **to** $w - 1$ **do**
    $E \leftarrow ExtractCoords(fm'(i, j))$; //5 feature vectors on FM
    for $e \in E$ **do**
      $e_x \leftarrow min(e_x, w - 1)$ *where* $e_x \leftarrow \max(e_x, 0)$;
      $e_y \leftarrow min(e_y, h - 1)$ *where* $e_y \leftarrow \max(e_y, 0)$;
      $fm'(i, j) \leftarrow fm'(i, j) + BI(fm, e)$; //Exact feature vector from BI
    end for
  end for
end for
$fm' \leftarrow fm' + fm$; //Reconstruct FM
return fm'

To obtain accurate features, bilinear interpolation is performed on the five previously derived feature vectors and the result is added to the current feature vector fm'. The resulting feature vector fm' replaces the existing feature vectors. The entire feature map is reconstructed by going through all feature points. Once the traversal is complete, the reconstructed feature map is added to the existing feature map. Feature refinement is performed multiple times during the feature reconstruction procedure specified in Eq (7), where fm_{i+1} shows the feature map of $i + 1$ level, b_i, c_i denotes the BB and confidence value of the i^{th} prediction.

$$fm_{i+1} = FR(b_i, c_i, \{P_2, \ldots, P_7\}) \quad (7)$$

To handle A_{IoU}problems arising during experimental analysis, a feature reconstruction method is used. The feature reconstruction procedure is used to minimize feature angulation issue. Initially, A_{IoU}

consists of numerous sampling points, and reducing the sample sizes affects the performance of the detector. Whereas, during the feature reconstruction process, sampling is performed by considering only the extracted feature points (five-parameters). Due to the fact that only a few points are considered for sampling, the computational effort of the detector is drastically reduced. Second, A_{IoU} generates instance-level features corresponding to a RoI before performing classification and regression. However, during the feature reconstruction process, the entire feature map is reconstructed in a pixel-by-pixel (image plane) manner. Thus, this reconstruction process is efficient and involves a smaller number of parameters compared to the previous process.

3.4. Rotation sensitivity error for oriented detection

The existing rotated detectors use a five-parameter [23,30,31,44] or an eight-parameter regression approach [32–34]. Either one of these parameter regression approaches are used to describe the rotated bounding boxes and its corresponding *L1* loss function. Though both these approaches have provided considerable results, they suffer loss discontinuity issue.

In five-parameter methods, the angle parameter is primarily responsible for the discontinuity of loss (DoL) issue. Once the angle reaches the limit of its range, the loss value increases. To obtain the ground truth and the prediction box, a horizontal rectangle is turned one degree clockwise and counterclockwise, respectively. The reference box position is only slightly changed, but the angular periodicity has significantly altered the angle of the rectangle. Additionally, according to the OpenCV-standard five-parameter definition method, the height and width are also switched. Furthermore, the five parameters of the system (angle, height, width and centers) have different units and show different IoU values. Thus, the performance may be negatively impacted by merely adding them up for inconsistent regression. Second, though the parameters in an eight-parameter method clearly denote the coordinate value, the discontinuity of loss also occurs in this method. This phenomenon is known as RSE.

Rotation sensitivity error occurs primarily due to the sudden loss change (increase) in the boundary case and is usually caused by the adoption of the angle parameter and switching of width and height. It also occurs due to inconsistency in regression of the five-parameter approach. Loss discontinuity is brought about by the angle parameter θ. The horizontal reference bounding box is rotated counterclockwise to produce the predicted box that matches the ground truth box. To address the DoL issue, an adjustable rotation loss function is proposed. This adjustable loss function follows symmetry of *L1* loss corresponding to its location. It achieves minimum *L1* loss with a continuous loss curve and does not reach the boundary range of the angle since it is larger than *L1* loss. The equation for the proposed adjustable rotation loss L_{ar} function is formulated as follows.

$$L_{ar} = min\begin{cases} L1\ \ (five - parameter) \\ L1\ \ (adjusted\ \ parameter) \end{cases} \tag{8}$$

The rotation sensitivity error occurs only when the boundaries are discontinuous. Equation (9) shows the boundary constraints of the adjustable rotation loss L_{ar} function.

$$L_c = |x_1' - x_2'| + |y_1' - y_2'| \tag{9}$$

$$L_{ar} = \ min\begin{cases} L_c + |w_1' - w_2'| + |h_1' - h_2'| + |\theta_1' + \theta_2'| \\ L_c + |w_1' - h_2'| + |w_1' - h_2'| + \left|90° - |\theta_1' + \theta_2'|\right| \end{cases} \tag{10}$$

where L_c is the loss function of the center point, Eq (9) specifies L_{ar} as the extension of $L1$ loss and Eq (10) is used for modelling a continuous loss function by removing angular periodicity. Therefore, the adjustable angle parameter L_{ar} is large than the normal $L1$ loss parameter if it fails to reach the boundary conditions specified by the angle parameter and ceases if the $L1$ loss is discontinuous.

4. Implementation details

4.1. Dataset

The datasets ImageNet, Olympic Sports, Sports Videos in the Wild, HABBOF and DOTA were used for the assessment due to their modularity and variety of actions as shown in Table 1. The ImageNet [35] consists of image dataset with more than 14 million images out of which 5000 images are used for training and testing. It consists of varied images with multiple object classes. The Olympic sports dataset [36] contains short sequenced YouTube videos of athletes playing 16 different sports in an uncontrolled environment. The Sports Videos in the Wild dataset [37] includes a collection of 4200 video images of amateur gamers. The data set consists of 30 sports with 44 different actions. It is considered the most difficult data set to annotate due to the presence of playing amateurs. The HABBOF dataset [38] consists of 4 videos recorded with mounted fisheye cameras and consists of 5847 frames. It consists of people performing normal activities such as walking, sitting, and standing. Some frames also consist of people making complex poses in close proximity. Finally, DOTA [52] is a large-scale aerial image dataset designed specifically for object detection and instance segmentation tasks. The dataset contains more than 2800 high-resolution aerial images covering different geographic locations and diverse object categories.

The single-stage rotation detector is modelled based on RetinaNet [39] and during training ResNet50, ResNet101 and ResNet 152 is used for experimental analysis. The ResNet50 backbone [32] is used for network initialization and is pre-trained on the ImageNet dataset using TensorFlow. The model is trained for 30 epochs for each dataset and the number of iterations is varied according to the number of epochs. The initial learning rate is set to $5^{e\text{-}4}$, momentum as 0.9 and weight decay as 0.0001. The learning rate is varied from $5^{e\text{-}5}$ to $5^{e\text{-}6}$ for 18 and 24 epochs respectively. Rotating non-maximum suppression (R-NMS) inference technique is used in post-processing the results of the final detection.

Additionally, the image iterations is set as 60,000 (ImageNet), 54,000 (Olympic Sports), 10,000 (Sports Videos in the Wild), 5,000 (HABBOF), and 45,000 (DOTA) respectively. The iteration is doubled when multi-scale training and data augmentation is introduced. A momentum optimizer over 1 GPU with a total of 4 frames per mini-batch is allocated for processing. For the pyramid levels (P_3 to P_7), the anchors range from 32^2 to 512^2 with seven aspect ratios {1, 1/2, 2, 1/3, 3, 5, 1/5} and three scales $\{2^0, 2^{1/3}, 2^{2/3}\}$. Additionally, six angles $\{-90^\circ, -75^\circ, -60^\circ, -45^\circ, -30^\circ, -15^\circ\}$ are added for the rotating anchor-based method (R-RetinaNet).

Table 1. Benchmark datasets used for evaluation of the proposed detector.

Dataset	Type	Modality	Resolution	Environment type	Features
ImageNet	Diverse	RGB	720 × 1280	Uncontrolled	Diverse images
Olympics Sports	Sports	RGB	640 × 360	Uncontrolled	16 Sports Activities
Sports Videos in the Wild	Sports	RGB	640 × 360	Uncontrolled	30 Sports Activities
HABBOF	Fish Eye	RGB	2048 × 2048 × 30	Controlled	Varied poses with occlusion
DOTA	Aerial images	RGB	1024 × 1024	Uncontrolled	Varied scale, orientation and shape

4.2. Baseline methods

Robust baseline models with different anchor settings are used to analyse the performance of the proposed work. Horizontal RetinaNet (H-RetinaNet) takes advantage of horizontal anchoring. Although fewer anchors are used, the IoU is computed as a horizontally defining ground truth rectangle that fits more positive models but also covers areas with many objects that are irrelevant. For objects with high aspect ratio, the predicted rotation limit field is usually not accurate. Similarly, Rotated RetinaNet (R-RetinaNet) leverages rotating anchors by adding an angle parameter to prevent the introduction of noise regions and offers better detection performance in dense scenarios. However, as the number of anchors increases, the efficiency of the model decreases.

Table 2. Ablative study of each component of the proposed detector on all datasets with and without the feature refinement module. It also explores various techniques such as data augmentation, sampling, box filtering, feature refinement and angle loss function.

Baseline	Specifications					Datasets (mAP %)				
	Data Aug.	Sample balance	Box filtering	Feature refinement	Angle loss	ImageNet	OSD	SVW	HABBOF	DOTA
ResNet50	Yes	Yes	No	No	No	66.7	68.1	63.7	65.9	70.1
ResNet50	No	No	Yes	Yes	Yes	70.7	71.0	64.7	67.3	69.5
ResNet50	Yes	Yes	Yes	Yes	Yes	71.3	73.9	67.3	68.6	73.5
ResNet101	Yes	Yes	Yes	Yes	Yes	72.4	77.1	71.0	70.1	78.2
ResNet152	Yes	Yes	Yes	Yes	Yes	73.6	78.5	73.7	70.2	75.8
R-RetinaNet	Yes	Yes	Yes	Yes	Yes	74.9	80.9	76.6	72.3	78.6
H-RetinaNet	Yes	Yes	Yes	Yes	Yes	79.2	81.4	78.7	76.0	78.0
Proposed	No	No	No	No	No	83.2	84.7	80.2	79.1	80.7
Proposed	No	No	Yes	Yes	Yes	85.8	87.8	83.2	81.3	83.6
Proposed	Yes	Yes	Yes	Yes	Yes	87.9	89.8	86.8	85.3	85.2

Table 2 shows the performance of the proposed detector and its comparison with other base models. The baseline models considered for assessment are ResNet-50, ResNet-101, ResNet-152, H-RetinaNet and R-RetinaNet evaluated across all datasets. The performance of the baseline models is

evaluated using the following specifications such as data augmentation, sample balancing, box filtering, feature refinement, and angle loss. Among the base models, H-RetinaNet has an overall mAP of 76.20% and R-RetinaNet has an overall mAP of 78.86%. From the results, it can be concluded that horizontal anchors perform efficiently in terms of speed, while rotated anchors have better regression functionality. It is suitable for detecting objects in cluttered or dense environments and objects with a large aspect ratio. The proposed detector shows an overall mAP of 81.82% without the inclusion of the feature refinement module. Similarly, it shows an overall mAP of 87.50% with inclusion of the feature refinement module. Comparing the performance of the proposed detector with baseline models is important for several reasons. First, it provides a benchmark to evaluate the effectiveness of the proposed detector. By comparing the performance of the proposed detector with that of existing baseline models, improvement in detection accuracy achieved by the proposed detector is assessed. Second, it helps to identify the strengths and weaknesses of the proposed model. From Table 2, it can be observed that the detection accuracy of the proposed detector can be improved with more training in terms of data augmentation (random cropping, translation, scaling rotation) and by increasing the diversity and quality of the training samples.

5. Ablation studies

5.1. Enhanced feature refinement module

Table 3 shows the comparison of the proposed detector with and without the addition of feature refinement module on all datasets. It can be observed that the performance of the proposed detector shows an accuracy of 81.6, 83.2, 85.1, 86.9 and 89.5% respectively without the addition of the feature refinement module. However, with the addition of feature refinement module, it shows a significant accuracy of 90.5(+8.9%), 93.2(+10%), 93.8(+8.7%), 92(+5.1%) and 92.3(+2.8%) respectively. From the results, it can be observed that precision increases with incorporation of the feature refinement module. Since, the feature refinement module reconstructs the entire feature map (pixel-by-pixel) based on anchor refinement technique, detection accuracy of the proposed detectors increases.

Table 3. Comparison of proposed detector with and without feature refinement (RF) module.

Method	FR	ImageNet		OSD		SVW		HABBOF		DOTA	
		P	R	P	R	P	R	P	R	P	R
Proposed Detector	No	81.6	84.9	83.2	90.1	85.1	87.0	86.9	89.1	89.5	91.9
Proposed Detector	Yes	90.5	90.4	93.2	89.2	93.8	91.9	92.0	94.9	92.3	92.3

5.2. Refinement strategy

Table 4 analyses the impact of the number of stages used for refinement and construction of the entire feature map. It also explores the relationship between the number of refinement stages and performance of the model. From Figure 3, it can be observed that as the number of stages increases, the performance of the model also increases since more robust features are identified via feature map reconstruction and hierarchical representations of the input-data are captured more accurately. The

accuracy of the refinement module is obtained as 71.3, 67.2, 70.9, 67.7 and 71.7% respectively, in the same experimental setup. Compared to the other deformable learning method, the enhanced feature refinement module is more efficient and accurate. This highlights the location responsiveness of the feature points, when the features are correctly refined, the performance of the model increases.

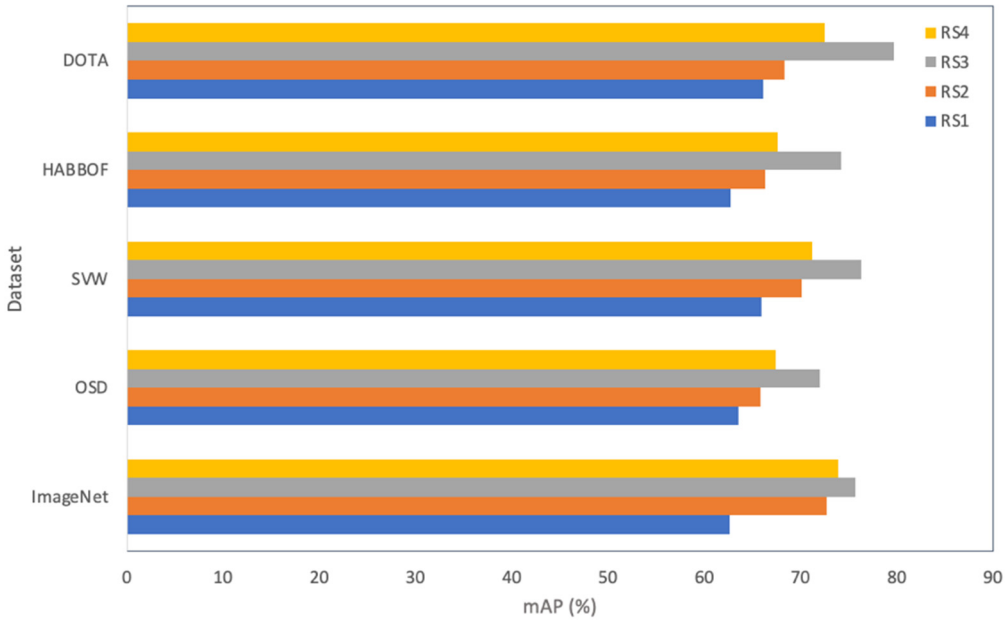


Figure 3. Relation between number of refinement stages and precision accuracy.

Table 4. Study of number of stages used on the benchmark datasets (RS-Refinement Stage).

Dataset	**Refinement Stages**				**mAP (%)**
	RS1	**RS2**	**RS3**	**RS4**	
ImageNet	62.6	72.7	75.7	73.9	71.3
OSD	63.5	65.8	72.0	67.4	67.2
SVW	65.9	70.1	76.3	71.2	70.9
HABBOF	62.7	66.3	74.2	67.6	67.7
DOTA	66.1	68.3	79.7	72.5	**71.7**

It is observed that RS3 provides improved results when compared to other stages. This is attributed to the fact that RS3 captures high-level and complex representations in a detailed manner due to an increased receptive field. Also, the inclusion of a feature refinement module enhances the process of feature composition to provide discriminative representation and non-linear transformations (different object classes) to improve the performance of the model.

5.3. Validation of detector using skew function

Additionally, the performance of the proposed detector is compared with various detectors using an approximated skew function. RetinaNet-based detectors are more likely to experience non-

convergence during training due to the occurrence of many low A_{IoU} (skew) prediction bounding boxes in the early stages of training. The derivative function is related to A_{IoU} in comparison to other linear function, meaning that it has a higher performance improvement because more attention is paid for training challenging samples.

5.4. Adjustable rotation loss function

Horizontal RetinaNet (H-RetinaNet) with ResNet50 backbone is used as the baseline model to verify the effectiveness of the loss function. As shown in Table 5, an accuracy of 3.25% (ImageNet), 5.12% (OSD), 5.63% (SVW), 3.02% (HABBOF) and 3.87% (DOTA) is achieved when the loss function transitions from smooth *L1* loss to adjustable loss function L_{ar}. These ablation experiments demonstrate that adjustable loss function L_{ar} is effective at enhancing the accuracy of the detector and also minimizes the loss associated with feature angulation issues. Additionally, these two techniques add a very small amount of parameters and computation, it does not increase the computational overhead of the model.

Table 5. Comparison of proposed adjustable loss function with other loss functions.

Loss Function	**Regression Type & Range**	**mAP (%)**				
		ImageNet	**OSD**	**SVW**	**HABBOF**	**DOTA**
Smooth L1	$[-\frac{\pi}{2}, 0)$	66.77	79.41	72.16	69.59	77.10
Smooth L1 [6]	$[-\pi, 0)$	68.41	80.62	73.45	72.41	78.24
IoU Smooth L1 [46]	$[-\frac{\pi}{2}, 0)$	69.99	81.22	75.59	73.97	78.53
Proposed Adjustable Loss (L_{ar})	$[-\frac{\pi}{2}, 0)$	73.24	86.34	81.22	76.99	82.11

5.5. Study on rotation sensitivity error

The concept of rotation sensitivity error has already been explored in previous works [41,42]. The previous works explored the concept of eliminating burst loss and incorporates trigonometric functions to eliminate the effect of angular periodicity. However, these methods fail to provide the solution for solving RSE. When compared to the previous methods, the proposed approach provides promising results.

5.6. Data augmentation and sampling

The performance of the proposed single-stage rotation detector can be increased with data augmentation. A variety of augmentation operations, such as random rotation, image greying, horizontal flipping, and vertical flipping is applied. For datasets that show severe imbalance, the samples in each category is increased by random copying. This random copying technique boosts the operation by 0.43%. Furthermore, the impact of different backbones is studied to enhance the performance of the detectors. Similarly, the performance of ResNet50 is 70.29%, RestNet101 is 72.70% and ResNet152 is 74.03% respectively as shown in Table 2.

6. Performance evaluation and comparison

6.1. Comparison with state-of-the-art detectors

The performance of the proposed single-stage rotation detector is compared with other state-of-the-art single- and two-stage detectors. Figure 4 shows the performance of both type of detectors (single- and two-stage) tested on benchmark datasets.

Figure 4. Performance comparison of proposed detector with state-of-the art detectors.

6.1.1. Comparison with single-stage detectors

The results on the datasets are shown in Table 6. The proposed detector is compared with eight state-of-the-art single-stage detectors such YOLOv2 [12], DRBox [40], IENet [41], DRN [42], P-RSDet [44], PIoU [43] and O^2-Dnet [28]. For the ImageNet dataset, the YOLOv2 detector showed the highest performance at 80.21%. For Olympic Sports Dataset, DRN showed an accuracy of 83.27%, R^3Det shows an accuracy of 82.89% for SVW, and 81.22% for HABBOF. However, the proposed detector showed a significant performance in both categories – with and without multi-scale training and testing. It outperforms the superior detectors such as YOLOv2, DRN and R^3Det for all datasets by 87.68% (7.47), 90.99% (7.72), 89.70% (6.81) and 89.15% (7.93) without multi-scale training and testing. It outperforms YOLOv2, DRN and R^3Det by 90.54% (10.33), 93.21% (9.94), 93.89% (11) and 92.04% (10.82) with multi-scale training and testing. The higher performance is attributed to the fact that the proposed single-stage detector is capable of identifying various classes and also classifies the significant differences between the classes (people, horses, bicycles, cars, dogs, athletes etc.).

6.1.2. Comparison with two-stage detectors

The performance of the detector is also compared with state-of-the-art two-stage detectors as shown in Table 6. Among these, RoI Transformer [46], SCRDet [23], Gliding Vertex [34], FFA [47] and CenterMap [5] have performed well. However, their detection accuracy was less when compared with the proposed detector as shown in Table 6. This is attributed to the fact that current two-stage detectors use complex model architectures that influence their performance. Additionally, the two-stage detectors increasingly depend on multi-stage regression and generation of low-level feature maps for detecting small objects present in dense environments. The proposed detector outperforms these detectors significantly due to its light-weight and strong backbone architecture without compromising on speed and accuracy.

Table 6. Performance comparison of proposed detector with state-of-the-art single- and two-stage detectors.

Method	Detector	Backbone	MS	Image Size	Datasets					Speed (fps)
					ImageNet	OSD	SVW	HABBOF	DOTA	
YOLOv2 [13]	SS	Darknet-19	Y	448 × 448	80.21	77.12	77.70	76.55	68.05	**64.6**
DRBox [40]	SS	Hourglass104	Y	511 × 511	73.07	75.17	77.16	71.90	73.01	76.1
IENet [41]	SS	ResNet101	Y	800 × 800	65.54	72.03	44.66	52.58	72.98	73.9
PIoU [43]	SS	DLA-34	N	511 × 511	69.70	60.21	52.58	64.83	75.21	83.5
P-RSDet [46]	SS	ResNet101	Y	800 × 800	73.65	78.58	69.75	77.26	72.68	94.9
O^2-Dnet [28]	SS	Hourglass104	Y	511 × 511	72.76	70.62	77.20	72.99	73.19	87.1
DRN [42]	SS	Hourglass104	Y	511 × 511	64.40	83.27	78.62	80.12	74.32	91.4
R^3Det [58]	SS	ResNet101	Y	800 × 800	77.67	81.72	82.89	81.22	82.87	81.5
AlignConv [54]	SS	S^2A-Net	Y	800 × 800	74.12	78.11	78.39	79.44	90.03	159.2
G-Rep [18]	SS	Cas-RetinaNet	Y	800 × 800	78.52	84.92	81.64	81.30	83.71	210.1
RIDet [59]	SS	ResNet50	Y	800 × 800	79.45	81.40	81.12	84.11	79.58	163.9
GFNet [51]	SS	EAST+ResNet50	Y	800 × 800	85.14	83.33	86.92	85.61	80.81	215.2
R^2CNN [24]	TS	ResNet101	Y	800 × 800	65.17	66.92	72.48	60.73	72.76	145.8
R-DFPN [31]	TS	ResNet101	N	800 × 800	73.70	70.61	72.83	71.03	77.10	135.1
R^2PN [22]	TS	VGG16	N	511 × 511	79.83	78.91	80.01	78.29	78.32	–
ICN [26]	TS	ResNet101	Y	800 × 800	64.90	67.80	70.04	69.05	72.03	154.5
FADet [40]	TS	ResNet101	Y	800 × 800	68.27	73.18	76.41	79.56	84.89	162.2

Continued on next page

Method	Detector	Backbone	MS	Image Size	Datasets					Speed (fps)
					ImageNet	OSD	SVW	HABBOF	DOTA	
RoI Trans. [44]	TS	ResNet101	Y	800 × 800	75.92	78.52	73.68	68.81	74.72	180.8
CAD-Net [2]	TS	ResNet101	N	800 × 800	73.50	71.10	63.50	73.21	79.81	170.9
SCRDet [23]	TS	ResNet101	Y	800 × 800	68.36	60.32	65.02	72.41	81.32	127.4
Cascade-FF [18]	TS	ResNet152	N	800 × 800	75.50	51.73	68.26	75.61	81.36	**110.1**
RADet [39]	TS	ResNeXt101	N	800 × 800	79.45	65.83	74.40	68.86	78.62	146.3
Gliding Vertex [35]	TS	ResNet101	N	800 × 800	72.94	79.02	70.91	72.33	81.46	134.5
CenterMap [49]	TS	ResNet101	Y	800 × 800	70.25	66.06	69.23	66.46	78.31	249.1
FFA [1]	TS	ResNet101	Y	800 × 800	61.20	71.11	64.63	64.20	70.91	302.9
APE [25]	TS	ResNeXt101	N	800 × 800	74.01	76.03	77.27	74.99	79.56	180.5
WSRD-Net [46]	TS	ResNet-FPN	Y	800 × 800	73.14	79.56	83.40	81.46	82.54	178.3
MSFC-Net [51]	TS	ResNeSt-101	N	800 × 800	77.34	79.58	81.43	83.05	84.49	189.2
FSODM [54]	TS	ResNet101	Y	800 × 800	79.58	77.81	86.93	84.62	81.30	156.1
Proposed	SS	RetinaNet	N	800 × 800	87.68	90.99	89.70	89.15	87.52	175.0
Proposed	SS	RetinaNet	Y	800 × 800	**90.54**	**93.21**	**93.89**	**92.04**	**90.33**	192.6

6.2. Comparison with benchmark datasets

In order to examine the effectiveness of the proposed detector, the detection results of the base model and the proposed detector are analysed using five benchmark datasets. Figures 5–9 show the comparison and qualitative illustration of the proposed detector on the benchmark datasets—ImageNet, OSD, SVW, HABBOF (overhead fish-eye dataset) and DOTA captured in both outdoor and indoor setting. The "red" boxes indicate the original fitted bounding boxes (Horizontal) and "green" boxes indicate the rotated bounding boxes.

The bounding boxes with a confidence value greater than 0.5 are considered. The proposed method fits the rotated bounding boxes to the aligned objects more precisely than other methods. The methods considered for the assessment had generated multiple bounding boxes, resulting in over-detection. The resulting over-detection occurs because the detector must find objects in overlapping areas throughout the image. Objects that occupy only half of the bounding boxes will degrade recognition accuracy because they are caused by the original bounding boxes rotated during training. The proposed detector overcomes this problem by closely fitting the boxes to the oriented objects present without using the images obtained during training. Compared to the detected boxes of the baseline methods, those of the proposed method are rotated according to the angle of appearance of the objects.

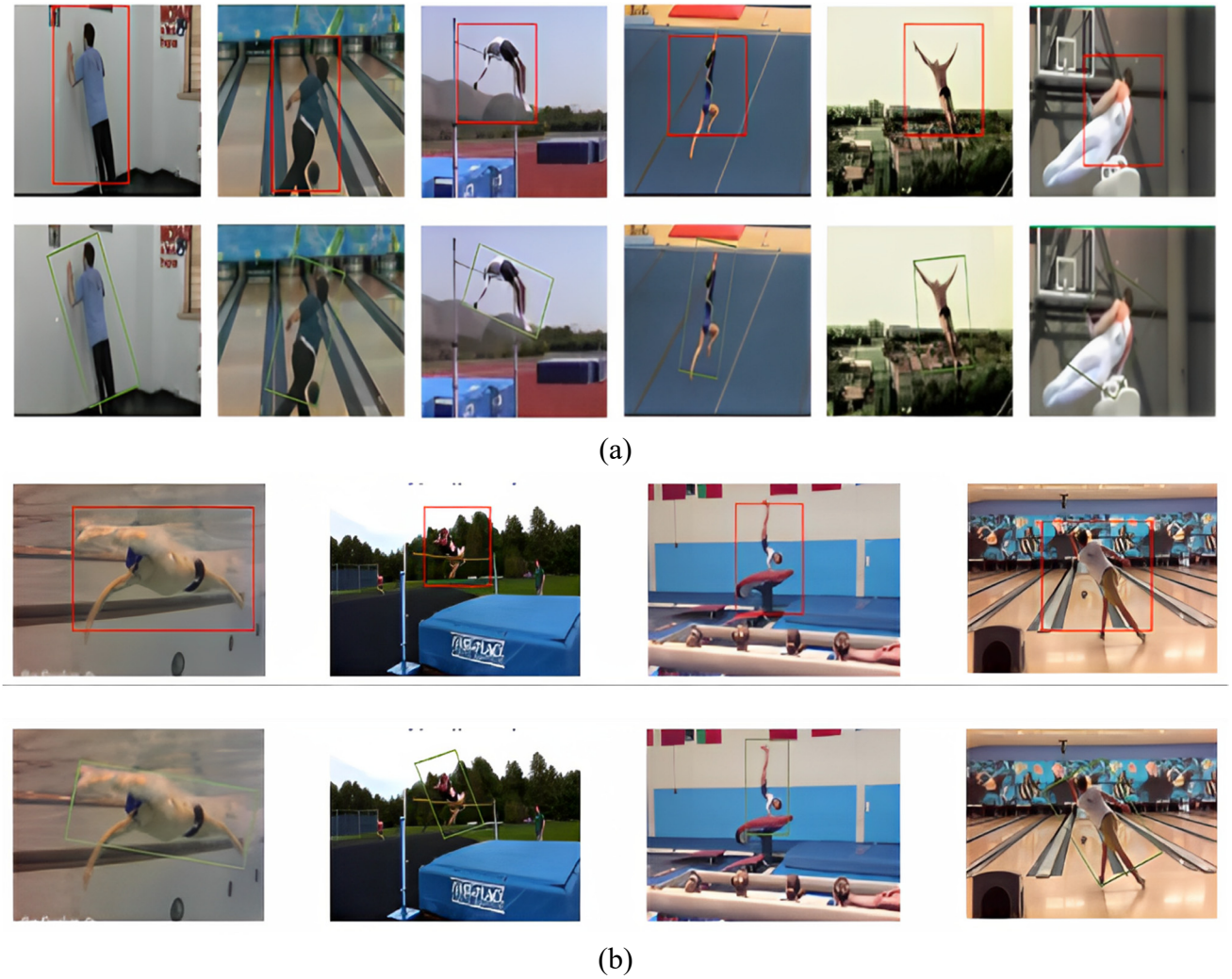


Figure 5. (a) Comparison and qualitative illustration of the proposed detector on OSD. Red colour indicates the original bounding box and green colour indicates the rotated bounding box. (b) Comparison and qualitative illustration of the proposed detector on OSD. Red colour indicates the original bounding box and green colour indicates the rotated bounding box.

To further analyze the baseline detection results and the proposed methods, an analysis of the relationship between the average accuracy and the positions of the objects in the images obtained from ImageNet, OSD, SVW and HABBOF datasets is shown in Table 4. The distortion of the records is removed based on the number of objects and their difficulty of detection. The object's position and its detection capability are varied for each angle by rotating the images $[-\frac{\pi}{2}, 0)$ and calculating their mAP (mean average precision) for each interval. It was found that the other detectors showed relatively low accuracy in the center and at the limit of the field of view. The reason for the low values is that only a part (top, bottom, or part of the objects) is considered. Because this type of phenomenon is rare in the selected datasets, the detectors cannot see objects present in the center.

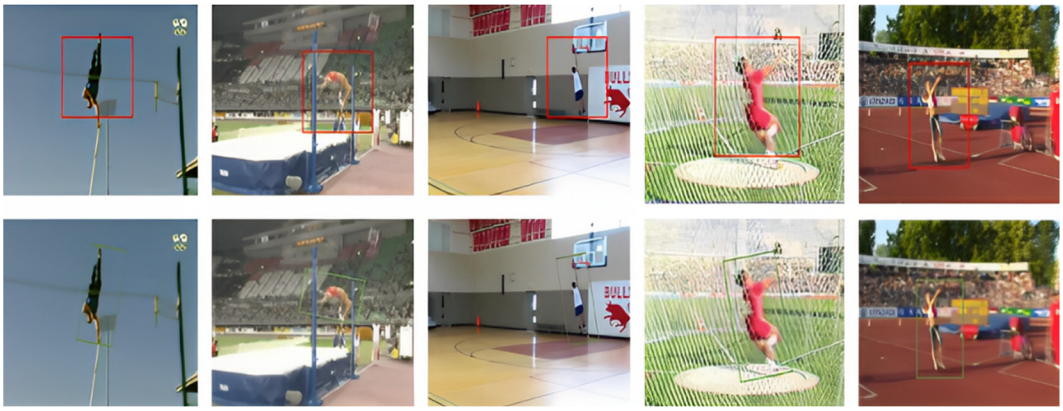
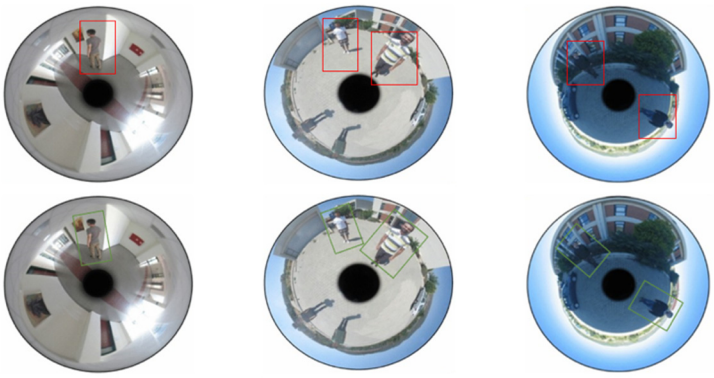
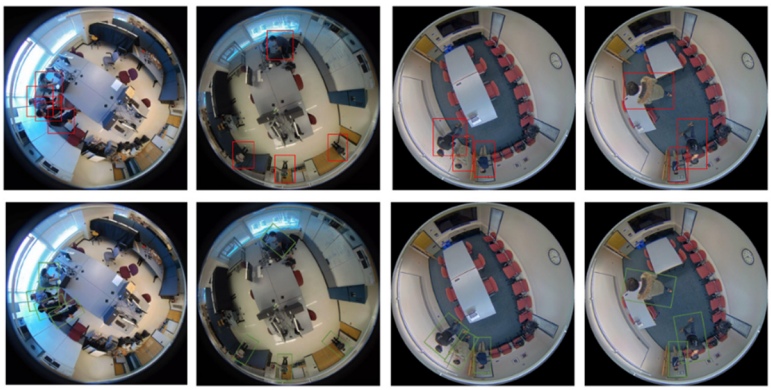


Figure 6. Comparison and qualitative illustration of the proposed detector on Sports Videos in the Wild (SVW). Red colour indicates the original bounding box and green colour indicates the rotated bounding box.



(a)



(b)

Figure 7. (a) Comparison and qualitative illustration of the proposed detector on the HABBOF over-head fish-eye dataset (outdoor setting). Red colour indicates the original bounding box and green colour indicates the rotated bounding box. (b) Comparison and qualitative illustration of the proposed detector on the HABBOF over-head fish-eye dataset (indoor setting). Red colour indicates the original bounding box and green colour indicates the rotated bounding box.

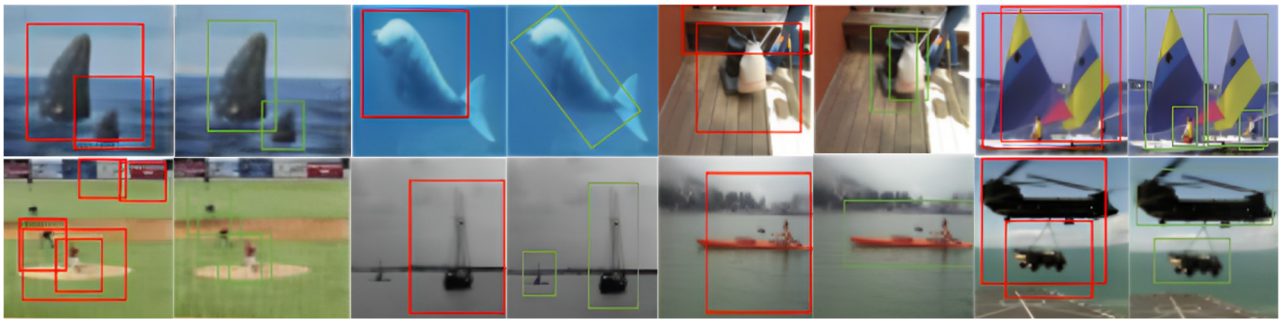


Figure 8. Comparison and qualitative illustration of the proposed detector on ImageNet dataset. Red colour indicates the original bounding box and green colour indicates the rotated bounding box.

Similarly, the reason for the low mAP at the edge is also that the detectors cannot see objects with tiny scales. However, the proposed method shows better performance than other detectors across all datasets, indicating its effectiveness. The performance of detectors degrades at angles greater than 15 degrees, while that of the proposed method is nearly the same at any angle. At greater angles, objects usually appear tilted, making it difficult to tightly fit bounding boxes around them. These are reasons attributed for low performance of the current detectors. The proposed detector differs in that it can closely match boxes to objects oriented at any angle, and therefore exhibits a stable performance at any angle.

Figure 9. Comparison and qualitative illustration of the proposed detector on DOTA.

6.3. Comparison of speed

The benchmark data sets consist of high- and low-resolution test images. These images require additional processing techniques such as cropping, merging, and rotating. Therefore, the speed and accuracy are maintained for each detector and its performance evaluation is conducted under the same test conditions. The effects of different backbone architectures and their frame sizes are also explored, as shown in Table 6. Among the two-stage detectors, Cascade-FF showed the highest speed. For single-stage detector YOLOv2 showed superior speed. These detectors show superior speed when compared to the proposed detector because they only work on a few categories while suppressing others during training. The speed increase of the proposed detector is because datasets with several categories are diverse. As the number of categories increases, so does the speed of the detector. Table 6 presents a comparison of the time required for the proposed detector along with other approaches. These methods have the same parameters, but differ in their post-processing procedures during inference. Despite the

incorporation of multi-scale slowing down the inference speed, the proposed detector demonstrated a performance improvement when compared to the baseline detectors.

6.4. Comparison of loss function

If there exists any angle loss in the model, it leads to stability issues during training. The issue that arises due to regression inconsistency and discontinuity in loss must be eliminated by using an adjustable loss function. Additionally, experimental analysis has shown that the proposed detector is more stable during training and outperforms other state-of-the-art detectors. The training curve loss between the adjustable loss function L_{ar} and the smooth $L1$ loss function is shown in Figure 10. It can be observed that the mean and variance of the both the curves appears stable after using the adjustable loss function.

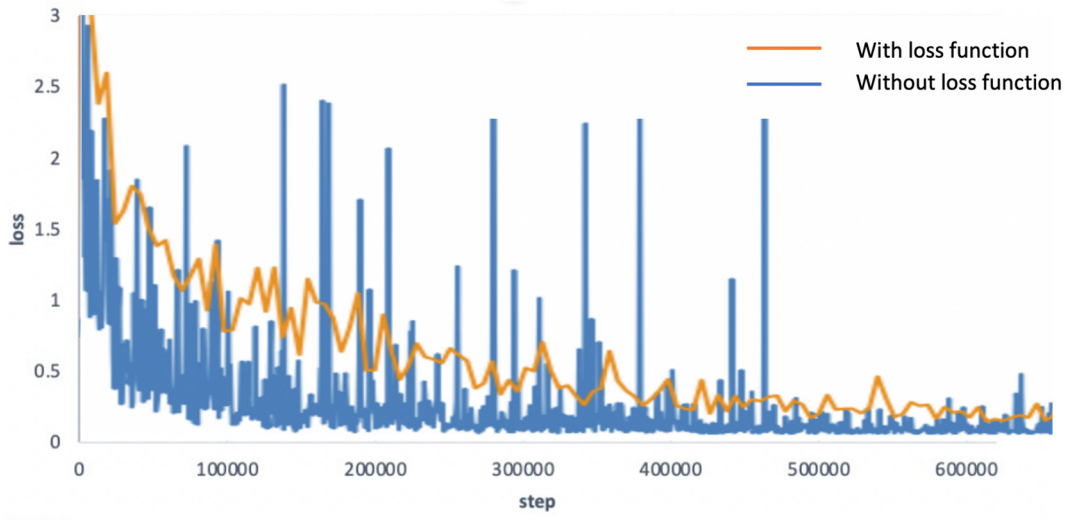


Figure 10. Comparison of training loss curve after using the adjustable loss function L_{ar}.

7. Conclusions

This paper proposes a single-stage rotation detector that can detect oriented, multi-scale objects with high accuracy by introducing an enhanced feature refinement module. This module refines the position of the oriented bounding box to its corresponding feature points by reconstructing the entire feature map. Furthermore, to improve the detection accuracy of oriented objects, the detector introduces an adjustable loss function that solves the problem of loss discontinuity. The effectiveness of the proposed detector is demonstrated by reconstructing the entire feature map multiple times. The proposed detector overcomes the challenges associated with detecting oriented objects, where objects tend to be oriented at various angles, leading to angle alignment issues and discontinuity loss with an adjustable loss function. For future work, the proposed detector can be trained with more diverse datasets to improve its accuracy. Furthermore, it can be optimized for pre-processing in transfer learning and implementation for real-time object detection to improve its performance in challenging scenarios.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. X. Chen, J. Yu, S. Kong, Z. Wu, L. Wen, Dual refinement networks for accurate and fast object detection in real-world scenes, preprint, arXiv: 1807.08638. https://doi.org/10.48550/arXiv.1807.08638
2. G. Zhang, S. Lu, W. Zhang, CAD-Net: A context-aware detection network for objects in remote sensing imagery, *IEEE Trans. Geosci. Remote Sensing*, **57** (2019), 10015–10024. https://doi.org/10.1109/TGRS.2019.2930982
3. H. D. Jang, S. Woo, P. Benz, J. Park, I. S. Kweon, Propose-and-attend single shot detector, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (2020), 815–824.
4. C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, X. Zou, Selective refinement network for high performance face detection, in *Proceedings of the AAAI conference on artificial intelligence*, **33** (2019), 8231–8238. https://doi.org/10.1609/aaai.v33i01.33018231
5. K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, X. Sun, Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images, *ISPRS J. Photogramm. Remote Sensing*, **161** (2020), 294–308. https://doi.org/10.1016/j.isprsjprs.2020.01.025
6. W. Qian, X. Yang, S. Peng, J. Yan, Y. Guo, Learning modulated loss for rotated object detection, in *Proceedings of the AAAI conference on artificial intelligence*, **35** (2021), 2458–2466. https://doi.org/10.1609/aaai.v35i3.16347
7. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2014), 580–587. https://doi.org/10.1109/CVPR.2014.81
8. R. Girshick, Fast R-CNN, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2015), 1440–1448. https://doi.org/10.1109/ICCV.2015.169
9. J. Dai, Y. Li, K. He, J. Sun, R-FCN: Object detection via region-based fully convolutional networks, *Adv. Neural Inf. Process. Syst.*, **2016** (2016), 29.
10. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *Adv. Neural Inf. Process. Syst.*, **39** (2015), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
11. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, preprint, arXiv: 1312.6229. https://doi.org/10.48550/arXiv.1312.6229

12. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, et al., Single shot multibox detector, in *Computer Vision–ECCV 2016: 14th European Conference*, (2016), 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
13. J. Redmon, A. Farhadi, YOLO9000: Better, faster, stronger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 7263–7271. https://doi.org/10.1109/CVPR.2017.690
14. T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 2117–2125.
15. C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, DSSD: Deconvolutional single shot detector, preprint, arXiv: 1701.06659. https://doi.org/10.48550/arXiv.1701.06659
16. Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 6154–6162. https://doi.org/10.1109/CVPR.2018.00644
17. K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, et al., Hybrid task cascade for instance segmentation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 4974–4983. https://doi.org/10.1109/CVPR.2019.00511
18. L. Hou, K. Lu, J. Xue, L. Hao, Cascade detector with feature fusion for arbitrary-oriented objects in remote sensing images, in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, (2020), 1–6. https://doi.org/10.1109/ICME46284.2020.9102807
19. Z. Tian, C. Shen, H. Chen, T. He, FCOS: Fully convolutional one-stage object detection, in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9626–9635. https://doi.org/10.1109/ICCV.2019.00972
20. T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, J. Shi, FoveaBox: Beyond anchor-based object detection, *IEEE Trans. Image Process.*, **29** (2020), 7389–7398. https://doi.org/10.1109/TIP.2020.3002345
21. Z. Yang, S. Liu, H. Hu, L. Wang, S. Lin, Reppoints: Point set representation for object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 9657–9666. https://doi.org/10.1109/ICCV.2019.00975
22. J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, et al., Arbitrary-oriented scene text detection via rotation proposals, *IEEE Trans. Multimedia*, **20** (2017), 3111–3122. https://doi.org/10.1109/TMM.2018.2818020
23. X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, et al., SCRDet: Towards more robust detection for small, cluttered and rotated objects, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 8232–8241.
24. Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, et al., R2CNN: Rotational region CNN for orientation robust scene text detection, preprint, arXiv: 1706.09579. https://doi.org/10.48550/arXiv.1706.09579
25. M. Liao, B. Shi, X. Bai, TextBoxes++: A single-shot oriented scene text detector, *IEEE Trans. Image Process.*, **27** (2018), 3676–3690. https://doi.org/10.1109/TIP.2018.2825107
26. S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, P. Reinartz, Towards multi-class object detection in unconstrained remote sensing imagery, in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision*, (2019), 150–165. https://doi.org/10.1007/978-3-030-20893-6_10

27. H. Rezatofighi, N. Tsoi, J. Y. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 658–666. https://doi.org/10.1109/CVPR.2019.00075
28. H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, X. Sun, Oriented objects as pairs of middle lines, preprint, arXiv: 1912.10694. https://doi.org/10.48550/arXiv.1912.10694
29. S. Zhang, L. Wen, X. Bian, Z. Lei, S. Z. Li, Single-shot refinement neural network for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 4203–4212. https://doi.org/10.1109/CVPR.2018.00442
30. X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, et al., Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks, *Remote Sensing*, **10** (2018),132. https://doi.org/10.3390/rs10010132
31. W. He, X. Y. Zhang, F. Yin, C. L. Liu, Deep direct regression for multi-oriented scene text detection, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2017), 745–753.
32. M. Liao, Z. Zhu, B. Shi, G. Xia, X. Bai, Rotation-sensitive regression for oriented scene text detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 5909–5918. https://doi.org/10.1109/CVPR.2018.00619
33. Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G. S. Xia, et al., Gliding vertex on the horizontal bounding box for multi-oriented object detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, **43** (2019), 1452–1459. https://doi.org/10.1109/TPAMI.2020.2974745
34. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, ImageNet: A large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, (2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848
35. J. C. Niebles, C. W. Chen, F. F. Li, Modeling temporal structure of decomposable motion segments for activity classification, in *Computer Vision–ECCV 2010*, (2010), 392–405. https://doi.org/10.1007/978-3-642-15552-9_29
36. S. M. Safdarnejad, X. Liu, L. Udpa, B. Andrus, J. Wood, D. Craven, Sports videos in the wild (SVW): A video dataset for sports analysis, in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, **1** (2015), 1–7. https://doi.org/10.1109/FG.2015.7163105
37. C. Li, C. Xu, Z. Cui, D. Wang, T. Zhang, J. Yang, Feature-attentioned object detection in remote sensing imagery, in *2019 IEEE International Conference on Image Processing (ICIP)*, (2019), 3886–3890. https://doi.org/10.1109/ICIP.2019.8803521
38. H. Zhang, H. Chang, B. Ma, S. Shan, X. Chen, Cascade RetinaNet: Maintaining consistency for single-stage object detection, preprint, arXiv: 1907.06881. https://doi.org/10.48550/arXiv.1907.06881
39. X. Pan, Y. Ren, K. Sheng, W. Dong, H. Yuan, X. Guo, et al., Dynamic refinement network for oriented and densely packed object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11207–11216. https://doi.org/10.1109/CVPR42600.2020.01122
40. L. Liu, Z. Pan, G. Chen, Y. Gao, Drbox family: A group of object detection techniques for remote sensing images, in *IGARSS 2019 IEEE International Geoscience and Remote Sensing Symposium*, (2019), 1446–1449.

41. Y. Lin, P. Feng, J. Guan, W. Wang, J. Chambers, IENet: Interacting embranchment one stage anchor free detector for orientation aerial object detection, preprint, arXiv: 1912.00969. https://doi.org/10.48550/arXiv.1912.00969
42. L. Zhou, H. Wei, H. Li, W. Zhao, Y. Zhang, Objects detection for remote sensing images based on polar coordinates, preprint, arXiv: 2001.02988.
43. Z. Chen, K. Chen, W. Lin, J. See, H. Yu, Y. Ke, et al., PIoU loss: Towards accurate oriented object detection in complex environments, in *Computer Vision–ECCV 2020: 16th European Conference*, (2020), 195–211. https://doi.org/10.1007/978-3-030-58558-7_12
44. J. Ding, N. Xue, Y. Long, G. S. Xia, Q. Lu, Learning RoI transformer for oriented object detection in aerial images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 2849–2858. https://doi.org/10.1109/CVPR.2019.00296
45. J. Wang, W. Yang, H. C. Li, H. Zhang, G. S. Xia, Learning center probability map for detecting objects in aerial images, *IEEE Trans. Geosci. Remote Sensing*, **59** (2020), 4307–4323. https://doi.org/10.1109/TGRS.2020.3010051
46. H. Liu, L. Jiao, R. Wang, C. Xie, J. Du, H. Chen, et al., WSRD-Net: A convolutional neural network-based arbitrary-oriented wheat stripe rust detection method, *Front. Plant Sci.*, **13** (2022), 876069. https://doi.org/10.3389/fpls.2022.876069
47. T. Zhang, Y. Zhuang, G. Wang, S. Dong, H. Chen, L. Li, Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery, *IEEE Trans. Geosci. Remote Sensing*, **60** (2022), 1–20. https://doi.org/10.1109/TGRS.2021.3108476
48. P. Wu, Z. Wang, B. Zheng, H. Li, F. E. Alsaadi, N. Zeng, AGGN: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion, *Comput. Biol. Med.*, **152** (2023), 106457. https://doi.org/10.1016/j.compbiomed.2022.106457
49. N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, X. Liu, A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection, *IEEE Trans. Instrum. Meas.*, **71** (2022), 1–14. https://doi.org/10.1109/TIM.2022.3153997
50. H. Li, N. Zeng, P. Wu, K. Clawson, Cov-Net: A computer-aided diagnosis method for recognizing COVID-19 from chest X-ray images via machine vision, *Exp. Syst. Appl.*, **207** (2022), 118029. https://doi.org/10.1016/j.eswa.2022.118029
51. D. R. Johnson, V. R. Uthariaraj, A novel parameter initialization technique using RBM-NN for human action recognition, *Comput. Intell. Neurosci.*, **2020** (2020). https://doi.org/10.1155/2020/8852404
52. G. S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, et al., DOTA: A A large-scale dataset for object detection in aerial images, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 3974–3983.
53. W. Yu, B. Lei, M. K. Ng, A. C. Cheung, Y. Shen, S. Wang, Tensorizing GAN with high-order pooling for Alzheimer's disease assessment, *IEEE Trans. Neural Networks Learn. Syst.*, **33** (2020), 4945–4959. https://doi.org/10.1109/TNNLS.2021.3063516
54. R. Yang, Y. Yu, Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis, *Front. Oncol.*, **11** (2021), 638182. https://doi.org/10.3389/fonc.2021.638182

55. S. Inthiyaz, S. K. H. Ahammad, A. S. Krishna, V. Bhargavi, D. Govardhan, V. Rajesh, YOLO (YOU ONLY LOOK ONCE) making object detection work in medical imaging on convolution detection system, *Int. J. Pharm. Res.*, **12** (2020), 312–326. https://doi.org/10.31838/ijpr/2020.12.02.0003
56. A. Kaur, Y. Singh, N. Neeru, L. Kaur, A. Singh, A survey on deep learning approaches to medical images and a systematic look up into real-time object detection, *Arch. Comput. Methods Eng.*, **29** (2021), 2071–2111. https://doi.org/10.1007/s11831-021-09649-9
57. S. Jaiswal, R. Yadav, J. D. Roselind, Emotion detection using natural language process, *Int. J. Sci. Methods Intell. Eng. Networks*, **2023** (2023).

AIMS Press