



# An Advanced Object Detection Framework for UAV Imagery Utilizing Transformer-Based Architecture and Split Attention Module: PvSAMNet



Museboyina Sirisha<sup>\*</sup>, Sadasivam Vijayakumar Sudha<sup>†</sup>

School of Computer Science and Engineering, VIT-AP University (Beside AP Secretariat), Near Vijayawada 522237, Andhra Pradesh, India

Corresponding Author Email: [sirisha.20phd7010@vitap.ac.in](mailto:sirisha.20phd7010@vitap.ac.in)

<https://doi.org/10.18280/ts.400434>

## ABSTRACT

**Received:** 16 November 2022

**Revised:** 15 March 2023

**Accepted:** 6 May 2023

**Available online:** 31 August 2023

### Keywords:

*object detection, transformer, split-attention module, cardinal groups, VisDrone-DET, IoU balanced loss*

In recent years, advancements in deep learning have fostered the development of sophisticated object detectors, specifically in the realm of computer vision. The inherent complexity of images captured by unmanned aerial vehicles (UAVs) presents a multitude of challenges for object detection. These include, but are not limited to, the detection of small and densely clustered objects, scale variance, occluded objects, and intricate backgrounds, which are particularly prevalent in drone-captured imagery when compared to natural scenes with larger and more distinct objects. The current landscape of object detection research has seen a surge in interest surrounding advanced, anchor-free object detectors, attention mechanisms, and the use of transformers as an alternative to convolutional neural networks. In light of these developments, this study introduces a novel object detection framework that eschews anchor utilization and leverages a transformer backbone for feature extraction. A cardinal grouping-based split attention module is integrated into this network to selectively extract the most pertinent features. The object detection head, termed the Pyramid Vision Split Attention Module Network (PvSAMNet), comprises three branches: classification, confidence, and regression, which collaboratively facilitate the final object detection from drone images. Additionally, an Intersection over Union (IoU) balanced loss function is employed to effectively equilibrate the classification and localization steps. The performance of the proposed detector is evaluated using the Visdrone-DET dataset, with the efficacy gauged by the average precision (AP) and average recall (AR) metrics. The results demonstrate that the proposed model outperforms other detector models with an average precision of 38.74. This study contributes to the ongoing discourse in the field of object detection, providing a novel framework that addresses the unique complexities of UAV imagery and demonstrates promising results in comparative evaluations.

## 1. INTRODUCTION

The field of computer vision has been marked by significant leaps forward, with object detection becoming an indispensable method. This technique, which involves generating bounding boxes and assigning categories to objects in images, forms the bedrock for further downstream tasks such as segmentation, image captioning, and object tracking. Object detection goes beyond mere object classification; it not only classifies objects in images but also pinpoints their exact locations by generating bounding boxes around them. As a cornerstone of computer vision, it finds applicability in an array of fields, including autonomous driving, image categorization, and face recognition, with specific tasks involving weed detection, face detection, license plate detection, and pedestrian detection. Given its fundamental role in video analysis and visual comprehension, this area has seen a surge of research interest.

The advancements in neural networks have played a crucial role in the progress of object detection. Notably, deep learning, an evolution of conventional neural network structures and methods, has significantly enhanced object detection capabilities. One particular area where object detection has found substantial application is in the analysis of images

captured by unmanned aerial vehicles (UAVs) or drones [1]. These images often contain dense, small-scale object information, presenting a unique challenge for object detection.

The widespread use of drones in fields like agriculture, security, aerial photography and videography, and hazard monitoring necessitates effective and automatic object detection for scene parsing on UAV platforms. However, drone images present a host of challenges, including small objects, dramatic scale variances, complicated backgrounds, occluded objects, and flexible viewpoints (Figure 1). These factors pose substantial challenges for convolutional neural networks (CNNs) used for general object detection.

Historically, numerous object detection strategies have been proposed, such as R-CNN [2], Faster R-CNN [3], Mask R-CNN [4] under two-stage detectors category, and YOLO [5], RetinaNet [6], SSD [7] under one-stage detectors category. While these have achieved remarkable performance for ground images, aerial images present a more formidable challenge. The common backbones used by various object detectors for image classification are VGG [8], AlexNet [9], ResNet [10]. Notably, dense residual networks, incorporated into the detectors of the YOLO series, known as Darknet, have proven to be highly effective in feature extraction.



**Figure 1.** Sample aerial view images from Visdrone dataset

The detection of small objects has garnered considerable attention, considering the ever-expanding range of UAV applications. These small objects are inherently unstructured, making their detection a persistent challenge in aerial images. Given the fixed receptive fields of convolutional kernels, they adversely affect the detection of dense and small targets in aerial images.

Significant efforts have been made to enhance the accuracy and performance of object detection. An approach to this problem involves using a density generation network to create density maps, which are then cropped to match the density maps. A convolutional neural network system that integrates SpotNet with SNIPER (Scale normalization in image pyramids) has been proposed to enhance the detection of small objects [11]. Energy consumption is another important consideration for networks, with solutions like ABCSA [12] applying clustering techniques to effectively manage energy consumption using a cluster-based head selection technique in the health domain.

Attention mechanisms, such as SENet [13], RAN [14], CBAM [15], and others, have emerged to enhance detection performance by exploiting position information and reducing the channel dimensions of input tensors with large-sized convolutional kernels. However, these attention mechanisms are typically integrated into deep convolutional networks, which, despite their significant contribution to strengthening contextual information, fail to capture long-range dependencies. These dependencies are crucial for detecting dense objects in aerial photography.

To improve semantic discriminability and eliminate category confusion in large and complex scenes captured by drones, the collection and association of scene information from huge neighborhoods may be beneficial for discovering object associations. In contrast, convolutional networks cannot capture contextual global information due to the locality of their convolution operation. As opposed to transformers, which can maintain sufficient spatial information to detect objects through multi-head self-attention, while focusing globally on dependencies between image feature patches. Furthermore, the object detector must be capable of adjusting to changing viewpoints in aerial images as well as possess

dynamic receptive fields. Many advanced convolutional networks have been proposed for detecting these targets and have achieved remarkable results but due to the receptive fields generated by convolutional networks, they do not have a positive impact on small and dense targets detection especially for aerial view or drone images. As a result, there may be some uncertainty regarding detection accuracy. Studies have demonstrated that vision transformers [16] have resilience to extreme occlusions, domain shifts, ETC when compared to convolutional neural networks (CNNs). One of the most impressive structures of transformer-based models is the Swin Transformer [17]. Pyramid Vision Transformers (PvTs) [18] are pure transformer backbones that can function as an alternative to CNN backbones for a wide variety of downstream tasks including dense predictions at pixel-level as well as image-level predictions. Though PvT achieved significant performance results, due to its single pooling operation it seems to be less powerful for learning dominant contextual representations from input images. Recent studies have demonstrated that transformer-based approaches are effective at detecting objects. In natural image datasets such as ImageNet [19] and MSCOCO [20], these methods have performed exceptionally well. In addition, transformer-based models have been used for the detection of targets in remotely sensed images and aerial images. The gathering and association of scene data from vast neighborhoods is a prominent feature of transformers helping in discovering object associations, which in turn acquires more contextual information and learn noticeable feature representations from the complex scenes taken by drones. These formidable features of transformers have paved new pathways to research replacing convolutions that fail in detecting long range dependencies. In challenging conditions, however, transformer-driven object detection methods are still insufficiently accurate in learning distinguishable features. To tackle these problems and improve the detection accuracy of transformers, we introduce the split attention module into the proposed PvT transformer network that can learn noticeable feature representations and acquire more contextual information through cardinal grouping helping in detecting the small and dense objects category in aerial view images. As the computation of the dataset will perform poorly if the localization and classification functions are not connected. An IoU balanced classification loss function is used to improve it, and to adaptively change the weights of the samples using loss functions, IoU balanced localization loss function is adapted thus improving the small and dense object categories competently.

Contributions of this work include the following:

- Proposing an anchor-free transformer based network that uses Pyramid vision transformer (PvT) as its backbone for efficient feature extraction.
- Incorporating a split attention module into the transformer network enabling the network to learn distinguishable features effectively and enhance contextual information through cardinal grouping.
- PvSAMNet anchor-free detection head with three branches classification, centerness and regression.
- Introducing IoU balanced loss functions for improving accuracy in classification and localization of detecting targets.

## 2. RELATED WORKS

### 2.1 Object detection

With the advent of deep learning, the performance of object detection has improved substantially. Existing object detectors are generally classified based on generating region of interest proposals. Two-stage detectors mainly rely on generating region proposals. R-CNN [2], Fast R-CNN [3], Mask R-CNN [4] fall under the examples of two-stage detectors. Though these detectors show significant accuracy improvements in detection they are slow at detecting the targets. Conversely, detectors with one stage require only a single pass down a neural network for predicting bounding boxes simultaneously. Examples falling under one-stage category are YOLO [5] series, SSD [7], RetinaNet [6] to mention a few. These detectors show significant performance in terms of speed but achieve less accuracy compared to two-stage models. In recent years, anchor-free detectors have been introduced which have been advancement to one-stage detectors category. Unlike anchor-based models which produce a number of preset anchors and require huge number of hyper parameters for fine-tuning, the anchor-free models are free from anchor generation and post processing step NMS(Non-maximum-suppression). CornerNet [21], CenterNet [22], FCOS [23], are few examples that come under the anchor-free category of one-stage detection that have received extensive attention as the need for setting pre-defined anchors and post-processing is eliminated which are crucial steps in other existing detectors. From conventional object detection models, the advancements in methodologies gave rise to these deep learning model algorithms using CNNs capable of extracting spatial information from images related to depth and edges by accomplishing them as matrices by applying several pooling and convolutional layers in the deep networks.

### 2.2 Transformers in vision

In recent years, the Transformer model that purely relies on attention mechanisms has developed as the standard solution for many natural language processing (NLP) problems, demonstrating remarkable accomplishments in the fields of text classification, machine translation, query answering, etc. The Transformer achieves this success as it uses self-attention that enables the acquisition of intricate interdependencies between input successions. There has been recent research into the application of transformers to computer vision, and numerous studies have demonstrated impressive performance when compared to convolutional neural network-based architectures.

Vision transformer (ViT) [16] is the first of its kind in Transformer models that can be employed as a reliable backbone for numerous computer vision related tasks. Input image is initially broken into numerous well-separated sections to adapt it for visual tasks and entrenched via linear layer. By using the Transformer, features suitable for downstream tasks are generated based on the dependency between patches. Although ViT has made significant progress, it is limited by its inability to integrate multiscale features and high computational overhead on processing high-resolution images. Utilizing the advantages of the CNN backbone, some works apply hierarchical transformer network structures that may effectively exploit multiscale characteristics to reduce computing complexity while continually reducing the amount

of patches for every layer. Another transformer model, PvT or Pyramid vision transformer [18] comes with a pyramid structure proposing a spatial-reduction attention (SRA) mechanism making the architecture capable of learning features over a wide range of scales and resolutions. DETR (Detection Transformer) [24] model stands as the first approach that successfully uses transformers in order to detect objects. It uses set-matching loss functions and encoder-decoder modules layered on top of typical CNN models such as ResNet [25]. By limiting the calculation of self-attention to uncorrelated local windows, Swin Transformer [17] comes up with a hierarchical transformer structure that optimizes efficiency. To reduce computational cost and improve modeling capabilities, CSWin Transformer [26] cultivates a cruciate window with self-attention method that it computes simultaneously in straight and upright bands. CrossFormer [27] is a vision transformer that incorporates multi-scale embedding and distance-based attentions among layers that help building attentions among objects with varying scales. Initial traditional detection models have been replaced with numerous advancements and improvements that introduced the deep learning architectures capable of improving detection accuracy. The widely used categories of deep learning object detection models using convolutionals including two stage, one stage under and anchor-free models have shown momentous improvements in the object detection field. Attention mechanisms and transformers models are other noteworthy models giving progressive results in the detection of objects.

## 3. METHODOLOGY

### 3.1 Transformers in object detection

There are typically three components in mainstream detectors: 1) a backbone that has strong capability in extracting relevant features that can help in detecting objects in images, large datasets can also be pre-trained on well-known image databases like ImageNet using the backbone and fine-tuned to many specific tasks, common examples of backbones include VGG, AlexNet, ResNet, ResNext when processing on GPU (Graphical processing unit) platform and squeezeNet, MobileNet, ShuffleNet when processing on a CPU(central processing unit) platform; 2) a neck that embeds few layers between the backbone and the head to exploit the features obtained from the backbone of the first step and strengthen the information by fusing and refining useful features that can be essential for the final detection step, some typical examples of networks used as a part of the neck layer include feature pyramid network (FPN), path aggregation network (PAN), Bi-directional feature pyramid network (Bi-FPN); 3) the final step of object detection task is performed by the detection head of the network that classifies and localizes the predictions based on the refined features from the previous step. Many recent works have been using attention mechanisms and transformer based models that rely purely on attentions and achieved remarkable results in detection based tasks. Transformers are redesigned encoder-decoder models that were introduced with attention mechanisms to boost machine translation performance. The most pertinent vectors are given the highest weighting by the attention mechanism in order to efficiently decode the entire encoded input pattern. This mechanism aims to maximize the utility of the decoder by utilizing the sequence

input parts most relevant to the problem as flexible as possible. In order to reason more effectively, transformers rely on a mechanism called attention, which allows them to select and focus on specific parts of their input. They take the lead of shape bias and also use an encoding-decoding architecture to focus on key parts of the image. Comparatively to superfluous networks like long short-term memory (LSTM), transformers can reproduce long-term colonies between sequence elements and enable parallel processing of sequences. A transformer is naturally suited for use as a set-function, unlike CNNs that require substantial inductive bias. Transformers provide excellent scalability to networks with large capacity and large datasets, thanks to their simple design. This allows them to process data inputs such as images, videos, text, and speech. Transformers incorporate extensive pre-training, mutual feature encoding and self-attention into the networks. The use of attention-based transformer modules is a feasible alternative to convolution operations. This work proposes a network that incorporates split attention module into transformer network to generate cardinal groups that can retain essential contextual information for detecting small and serried objects taken from aerial point of view from drones.

The schematic diagram of the proposed architecture is shown in Figure 2. The proposed detector is primarily made of three modules which are: 1) Pyramid vision transformer (PvT) as the backbone of the network due to its ability to make fine-grained patches from input images; 2) Split attention module ResNeSt [28] that produces cardinal groups for feature maps; 3) Detection head of PvSAMNet that is responsible for classification and localization in the final step.

### 3.2 PvT backbone

Pyramid vision transformer (PvT) [18] is the first convolutional free network that can be used as an alternate to convolutional backbone structures and is helpful for many down-stream tasks which include image-level and pixel-level dense predictions. PvT overcomes the limitations of conventional transformers. It can generate fine-grained patches from input images that are crucial to learn

representations in high-resolution images for dense prediction. It introduces progressive shrinking pyramid structure that minimizes the sequence length as the transformer network deepens thereby reducing computational cost involved. It incorporates a spatial reduction attention (SRA) module that reduces resource consumption while learning representations in high-resolution images. Another important merit of PvT is that they generate global receptive fields that are more suitable for detection when compared to CNN that can produce local receptive field that increases with the depth of the network. The pyramid structure of PvT makes it more suitable to be embedded into many dense prediction models such as Mask R-CNN, RetinaNet, etc. Dense prediction tasks at pixel-level are well handled by PvT by combining with other task specific decoders for the detection step. The need for dense anchors and NMS (non-maximum suppression) post processing step is eliminated thereby increasing speed in detection. As a result of these astounding benefits, we have chosen PvT as the backbone of our proposed model. The features extracted are fed as input to the neck where a spatial attention module with ResNest block is used for cardinality grouping of the feature maps generated. The architecture diagram of PvT is shown in Figure 3.

### 3.3 Split attention network module using ResNeSt block

Figure 4 shows the split attention network architecture. The split-attention block consists of a computational unit that combines feature map grouping and splitting attention operations. As in the ResNext [29] networks, the features can be grouped into several blocks and numbering is assigned by a hyper-parameter with cardinality  $K$  to the feature map group blocks. The resulting grouped blocks are named as cardinal groups. A hyper-parameter denoted by  $R$  indicates number of splits identified in cardinal groups and is given by Eq. (1):

$$G = K * R \quad (1)$$

with the overall feature groups.

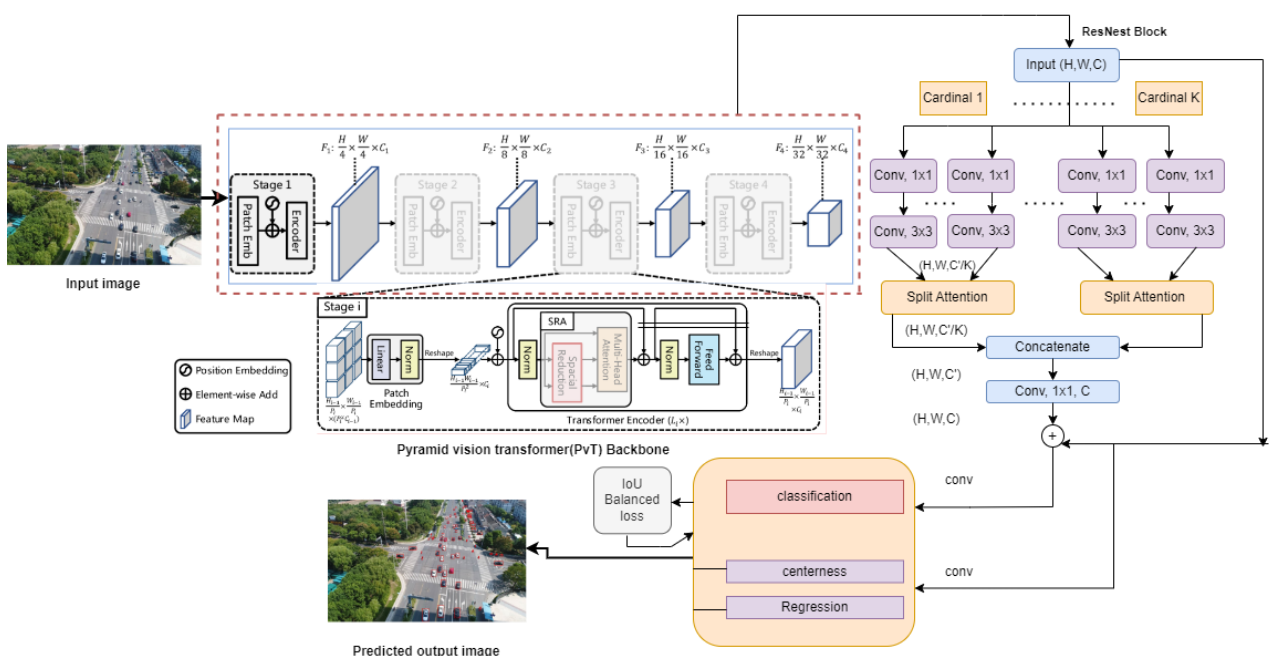


Figure 2. Proposed architecture of PvSAMNet

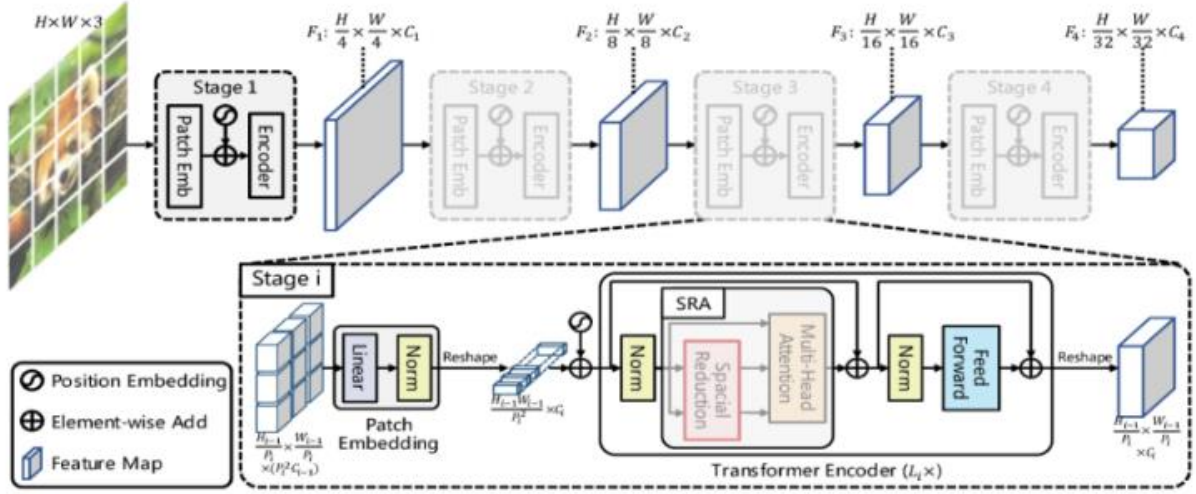


Figure 3. Pyramid vision transformer (PvT) architecture

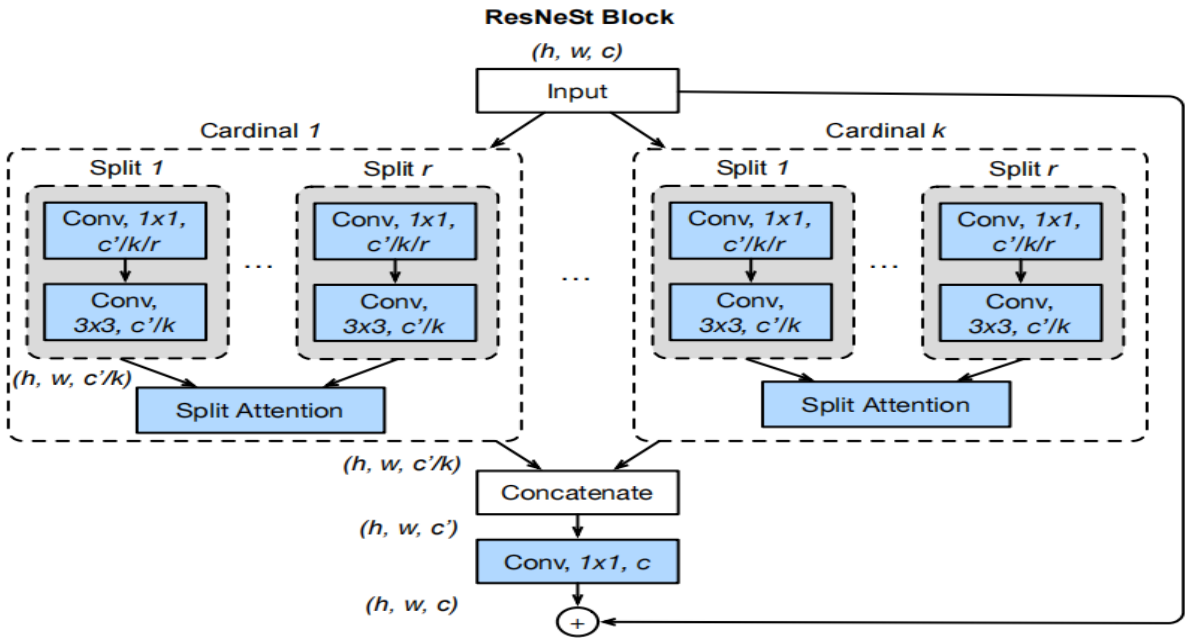


Figure 4. Split attention module

A sequence of amendments  $\{F_1, F_2, F_3, \dots, F_G\}$  is applied on each group and the intermediate representation per group is given as Eq. (2):

$$u_i = f_i(x) \quad (2)$$

where,  $i \in \{1, 2, 3, \dots, G\}$ . The modules using split attention are assigned for fusing feature maps among split groups. By combining multiple splits through an element-wise summation, each cardinal group can be represented in a combined manner. Additionally, each cardinal group representation is acquired by fusion of various splits on the totality of all components.  $K^{\text{th}}$  cardinal group is represented by Eq. (3) as follows:

$$\hat{U}^k = \sum_{j=R(k-1)+1}^{Rk} U_j \quad (3)$$

where,  $\hat{U}^k \in R^{H \times W \times c/K}$  for  $k \in 1, 2, 3, \dots, K$ .

The dimensions of the output feature maps are measured by H, W, and C. The overall average pooling of spatial

dimensions allows the collection of contextual information along with embedded channel-wise stats on a global scale given as Eq. (4):

$$S^k \in R^{c/k} \quad (4)$$

The component at  $c^{\text{th}}$  location is calculated using Eq. (5) as shown below.

$$S_c^k = \frac{1}{HXW} \sum_{i=1}^H \sum_{j=1}^W \hat{U}_c^k(i, j) \quad (5)$$

Cardinal representation of groups after fusion of weights represented by Eq. (6)

$$V^k \in R^{H \times W \times c/K} \quad (6)$$

Eq. (6) is combined with a weighted fusion that can be

added stream-wise using soft attention for each stream, where a feature map is produced by combining weights across splits. The  $C^{th}$  channel calculation is done using Eq. (7):

$$V_c^k = \sum_{i=1}^R a_i^k(c) * u_{R(k-1)+i} \quad (7)$$

in which,  $a_i^k(c)$  represents a weighted position as shown in Eqs. (8) and (9):

$$a_i^k(c) = \frac{\exp(G_i^c(s^k))}{\sum_{j=1}^R \exp(G_j^c(s^k))} \text{ for } R>1, \text{ and} \quad (8)$$

$$= \frac{1}{1+\exp(-G_i^c(s^k))} \text{ for } R=1 \quad (9)$$

According to the global context representation of  $s^k$ ,  $G_i^c$  determines the weight for each split depending on the global context representations of the  $c$ -th channel.

### 3.3.1 ResNeSt unit

Representations of the feature map groups are inserted together with the channel proportions, and is given in Eq. (10) as follows:

$$v = \text{concat} \{ v_1, v_2, \dots, v^k \} \quad (10)$$

By using shortcut connection, the split-attention block yields the final output given by  $y$  as in other residual nets and is represented by Eq. (11):

$$y = v + x \quad (11)$$

with similar shapes shared both by input and output feature maps. Suitable shortcut connections are transformed so that they line up with output shapes using transformation  $T$  for blocks with a stride. It is represented by Eq. (12):

$$y = v + T * x \quad (12)$$

In this case,  $T$  possibly is the result of combining convolution with pooling, striding convolution.

The split attention module with the ResNeSt unit and cardinal grouping of feature maps generates more robust features under complex scenes.

### 3.4 PvSAMNet detection head

The PvSAMNet detection head detects and generates results taking the dominant positive features produced by split attention module. Most widely used detectors use anchor-based methods for detecting objects which require generating many preset anchors based on the feature maps. These anchor based methods lack in generalization and are confined to specific tasks. The anchor-free methods on the other hand are simple in design requiring a smaller number of parameters for fine-tuning. We consider an anchor-free approach for our detection head unit which is composed of three branches for final step detection, a classification branch, a regression branch and a centerness branch parallel to regression. The classification branch produces heatmap at the center, given by Eq. (13):

$$\hat{y} \in [0, 1]^{C \times H \times W} \quad (13)$$

where,  $C$  represents number of categories corresponding to Eq. (14):

$$y \in [0, 1]^{C \times H \times W} \quad (14)$$

Parameters  $H$  and  $W$  representing the height and width respectively. Dynamic adjustment of positive samples is performed by the centerness branch to represent object regions. The centerness branch abolishes the low confidence bounding boxes that helps eliminate conflicts in classifying and localizing the targets in the final step. Using regression branch, a tensor with dimension  $4 \times H \times W$  is produced representing each location with a bounding box associated with a particular object. Further we adopt IoU balanced classification and localization loss functions [30] to reduce conflicts between classification and localization and improve the association among them for an optimal detection.

### 3.5 IoU balanced loss functions

#### 3.5.1 IoU-balanced classification

The performance of object detection models is heavily reliant on loss functions. The development of object detection techniques has led to the proposal of numerous distinct types of loss functions. Cross-entropy loss, SSD and RetinaNet is widely used as the classification loss in most prominent object detectors. Regardless of the localization accuracy, it will motivate the models to learn as many positive samples with high categorization rates as they can. The gradient dominates the training process of the localization branch for the localization loss, which affects the accuracy of localization. Therefore, IoU-balanced classification and localization is performed in the proposed method. Both of these losses have the potential to improve the object detection accuracy for precise localization.

The lack of connection among localization and classification function will negatively impact the performance while computing the dataset. This leads to a loss-balanced classification model that improves the correlation between classification and localization as in Eq. (15) below:

$$\begin{aligned} \text{Class} = & \sum_{i \in po}^N \omega_i(iou_i) * CE(p_i, \hat{p}_i) \\ & + \sum_{i \in Ne}^M CE(p_i, \hat{p}_i) \end{aligned} \quad (15)$$

where,  $po$  and  $Ne$  is denotes the sets of +ive and -ive training samples, respectively.  $iou_i$  indicates the regressed IoU for each regressed +ive sample. The IoU among the regressed bounding boxes and its matching ground truth boxes is positively connected with the weights  $\omega_i(iou_i)$ . The higher IoU will provide greater gradients during training, making it easier for the model to gain higher classification scores for the dataset.

#### 3.5.2 IoU-balanced localization loss

Using the loss functions adaptively alters the weight of positive samples based on their localization accuracy. The localization accuracy of detectors will suffer gradients driven by outliers dominating the training progression. In this way, examples with a high IoU are given more weight, and examples with low IoU are given less weight and is

represented by Eqs. (16) and (17):

$$Loc = \sum_{i \in Pos}^N \sum_{m \in \{lx, ly, \omega, h\}} \omega_i(iou_i) * L1(v_i^m - \hat{d}_i^m) \quad (16)$$

$$\omega_i(iou_i) = \omega_{loc} * iou_i^\lambda \quad (17)$$

With parameters  $lx, ly, \omega, h$  representing the predicted box parameterized coordinates and  $\hat{d}_i^m$  parameter representing the coordinates of the ground truth box respectively. Parameter  $\lambda$  controls how much localization loss concentrates on inliers while suppressing outliers. For the first iteration of the training method, the localization loss weight  $w_{loc}$  is manually modified to maintain the total of localization loss constant relative to the original smooth L1 loss.

#### 4. RESULT AND DISCUSSION

Aerial view dataset VisDrone-DET 2021 is used to assess the proposed framework used for object detection [31]. Using Python programming with i5 processor and 4GB-RAM system, testing of the proposed method's efficacy has been done through experiments using anchor-free transformer-based object detection model, PvSAMNet. The proposed object detector's performance is compared to the performance of a variety of existing detectors in order to assess its effectiveness.

##### 4.1 VisDrone-DET

Among the datasets used under aerial images, VisDrone-DET is one of the most widely used datasets. There are 6471 images in the Visdrone-DET dataset, 548 images in the validation package, and 1580 images in the test-challenge package, respectively. A total of 10 categories of objects are assigned to all the data. The Visdrone-DET dataset presents an extensive challenge in the object detection task. There are ten classes of data included in the VisDrone-DET dataset, including persons, pedestrians, cars, bicycles, vans, tricycles, trucks, awning-tricycles, motors, buses, and others. Figure 5 shows the category distribution of visdrone-DET dataset. Degrees of occlusion among different categories on training, validation, test-challenge, test-dev in visdrone dataset is shown in Figure 6 and category-wise degrees of occlusion in training, validation, test-challenge and test-dev on visdrone dataset is shown in Figure 7.

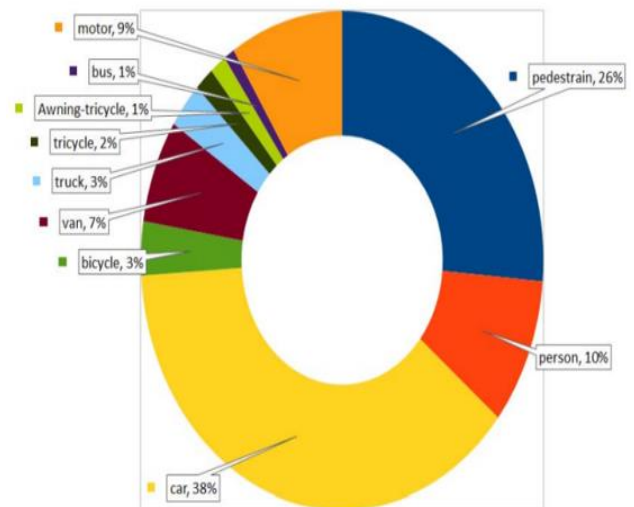
##### 4.2 Performance evaluation of VisDrone-DET dataset

The effectiveness of proposed object detection approach with VisDrone-DET dataset is shown in Table 1. The proposed object detector obtains detection accuracy of 38.74 average precision value. Accordingly, the accuracy on the overall dataset is improved. The VisDrone-DET dataset performance is characterized by the measurement of Average Precision (AP) and Average Recall (AR) and Mean Average Precision (mAP) metrics. AP, AP50, AP75, AR1, AR10, AR100, and AR500 indicators are used for assessment and ranking using the assessment of MS COCO dataset. The analysis focuses on the AP indicator, which is calculated by averaging the total step size of 0.05 for all the 10 object categories to the intersection over union (IoU) threshold at values between 0.50 and 0.95.

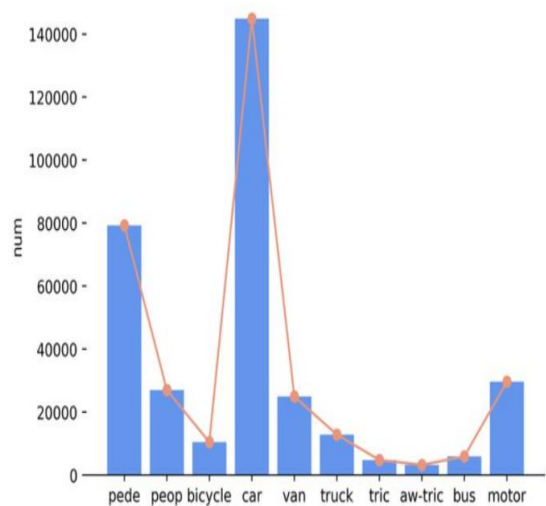
When the threshold for the IoU is 0.50 and 0.75, the accuracy is represented as AP50 and AP75 respectively. Furthermore, the sum of 1,10,100 is calculated based on the average recall. The proposed detector model reports AP of 38.74, AP50 of 62.98, AP75 of 40.48 values. The overall average precision is enhanced eliminating false bounding boxes. The average recall values obtained are AR of 1.01, AR10 of 6.02, AR100 of 43.03, AR500 of 45.14. The AP and AR values are shown in Table 1 below and the values are plotted in graph as shown in Figure 8.

**Table 1.** Object detection results on Visdrone-DET dataset

Method	AP (%)	AP50 (%)	AP75 (%)	AR1 (%)	AR10 (%)	AR100 (%)	AR500 (%)
CascadeR-CNN	16.08	31.92	14.98	0.27	2.76	20.98	28.41
Droneeye2020	34.56	58.24	35.79	0.26	1.91	7.01	52.35
DPNet-Ensemble	37.37	62.04	38.98	0.84	7.95	41.97	53.75
EfficientDet	38.51	63.24	39.56	1.81	10.99	44.01	55.13
DNEFS	38.52	62.86	39.98	1.41	9.62	43.02	55.02
Cascade++	38.71	62.94	41.07	1.08	7.01	42.97	43.32
Proposed	38.74	62.98	40.48	1.01	6.02	43.03	45.14



**Figure 5.** Category distribution in Visdrone-DET dataset



**Figure 6.** Analysis of category-wise statistics on Visdrone-DET training set

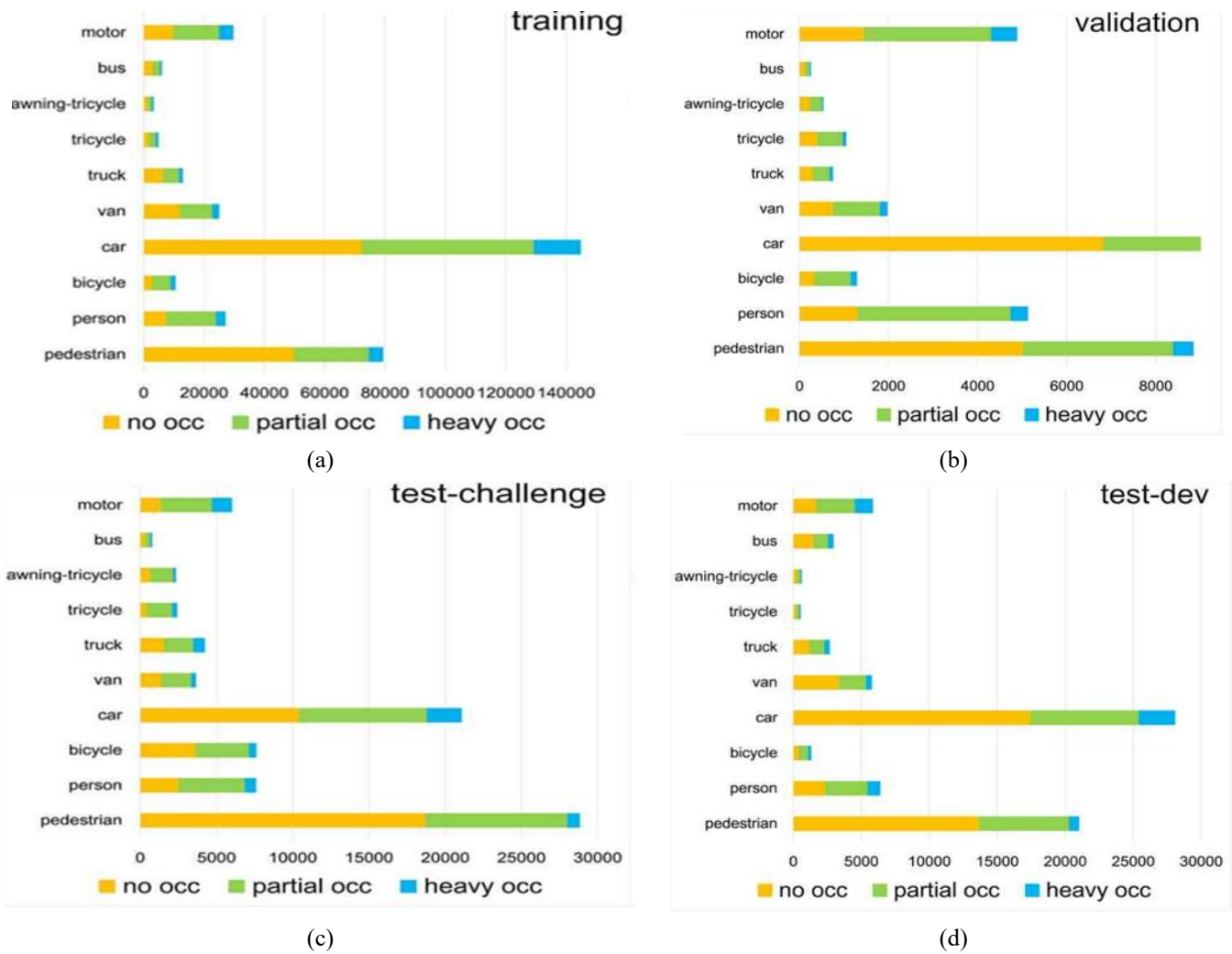


Figure 7. Occlusion categories among different classes in (a) training, (b) validation, (c) test-challenge, (d) test-dev of Visdrone-DET dataset

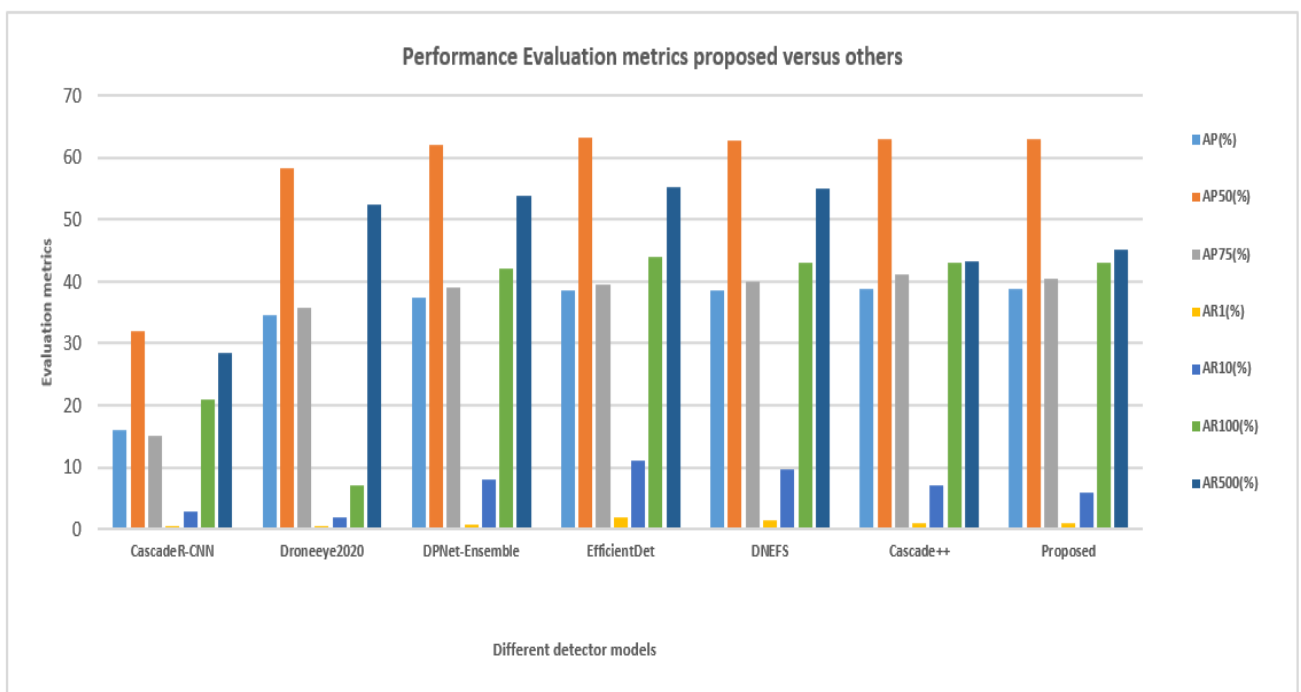
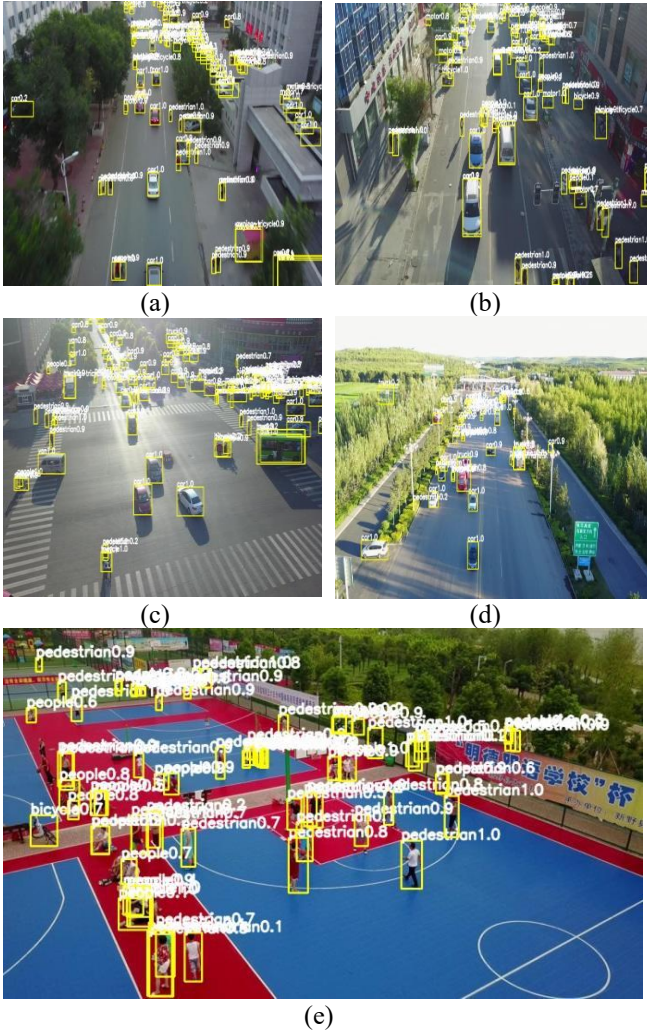


Figure 8. Average precision (AP) and average recall (AR) metrics for various detectors versus proposed



**Table 2.** The results of each class on Visdrone-DET

Method	Pedestrian (%)	Person (%)	Bicycle (%)	Car (%)	Van (%)	Truck (%)	Tricycle (%)	Awning (%)	Bus (%)	Motor (%)
CornerNet	20.43	6.55	4.56	40.94	20.23	20.54	14.03	9.25	24.39	12.10
Light-RCNN	17.02	4.83	5.73	32.39	22.12	18.39	16.63	11.91	29.02	11.93
FPN	15.69	5.02	4.93	38.47	20.82	18.82	15.03	10.84	26.72	12.83
Cascade	16.28	6.16	4.18	37.29	20.38	17.11	14.48	12.37	24.31	14.85
RRNet	27.34	20.13	21.45	32.56	29.35	25.74	20.46	18.58	35.71	26.17
Cascade++	36.41	34.56	29.61	45.31	39.18	35.32	31.53	25.38	46.67	36.56
Proposed	36.79	35.31	29.15	45.37	38.98	34.97	31.25	24.98	46.69	37.13



**Figure 9.** Object detection on VisDrone-DET dataset for five images (a, b, c, d, e)

**Table 3.** Comparison analysis of VisDrone-DET dataset

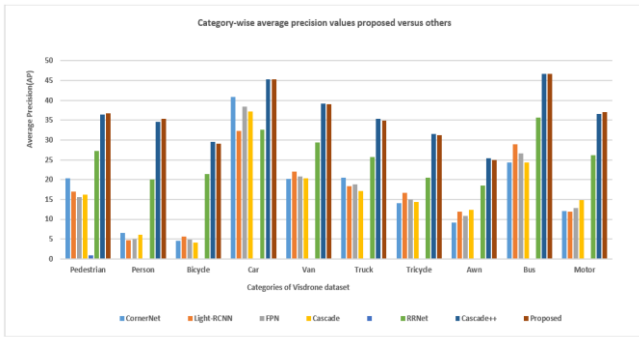
Methods	MAP (%)	AP50 (%)
RetinaNet	11.81	21.37
Cascade R-CNN	16.08	31.93
FPN	16.48	32.23
Light R-CNN	16.51	33.01
CornerNet	17.42	34.21
RRNet	29.23	56.01
DPNet Ensemble	29.61	53.98
DPNetv3	37.37	62.04
Cascade++	38.71	62.94
Proposed	38.74	62.98

Figure 9 shows qualitative object detection on the VisDrone-DET dataset for five images (a, b, c, d, e)

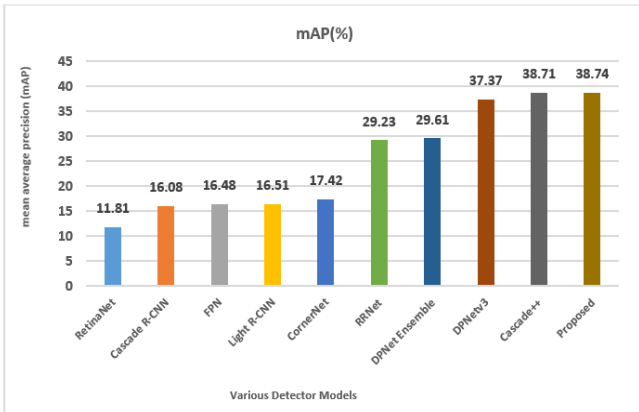
respectively. The proposed method is compared with CornerNet, Light-RCNN [32], FPN [33] and Cascade R-CNN [34], EfficientDet [35], Cascade++ [36]. Table 2 shows results of each class in the dataset comparing with other detector models results. Category-wise average precision of existing detectors and the proposed is plotted as shown in Figure 10. The overall AP of all classes is significantly improved using the anchor-free transformer detector helping to detect small and dense objects more efficiently. The proposed network obtains class-wise average precision values of pedestrian (36.79%), person (35.31%), bicycle (29.15), car (45.37), van (38.98%), truck (34.97%), tricycle (31.25%), awning-tricycle (24.98%), bus (46.69%), and motor (37.13%).

Table 3 shows the VisDrone-DET comparison evaluation of mean average precision (MAP) and average precision (AP) metrics obtained. The proposed model is compared with other detectors FPN, CornerNet, DPNetv3 [36], Cascade++ on MAP and AP metrics. The graphical representation of these values is shown in Figure 11. The method proposed achieves mean average precision of 38.74 and shows a significant improvement in comparison to other state-of-the-art works such as DPNetv3 and Cascade++ which could achieve 37.37 and 38.71 respectively. As the UAV images pose challenges with occlusion, scale variance and other factors, several works and methods have still achieved a percentage below 40 for the mean precision value. This shows the exceptional challenge for detection in these UAV images and the current work has improved the precision value to 38.74, which shows significant improvement in spite of the value still being below 40. This is due to the various factors that impact the detection in drone images unlike other natural scene images datasets where the mean precision values can go up to 70 with different methodologies. Comparison thus shows vast differentiation in natural scenes images and the aerial view or drone images. The proposed method obtains a MAP of 38.74 showing significant improvement. Though the obtained value is better than the other detectors, because of the various challenging characteristics posed by drone images, the MAP value is still below 40 and needs to be further improved.

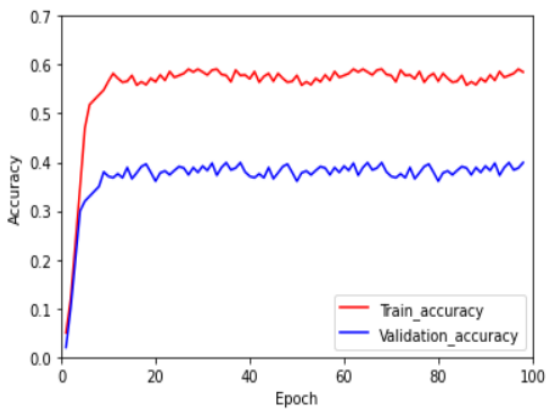
Figure 12 shows an illustration of the training versus validation accuracy graph. The suggested network has stable and quick convergent training processes, according to observations. The training and validation datasets are used to compare accuracy. The accuracy analysis shows that the suggested network generates better outcomes and fosters a more stable training procedure. The training and validation loss graph is shown in Figure 13. The initial loss value in the suggested strategy is minimal and effectively lowers as the number of epochs rises. The loss value is remarkably little after 100 training epochs have been completed on the data. As a result, there is a higher accuracy rate and less loss for the proposed model.



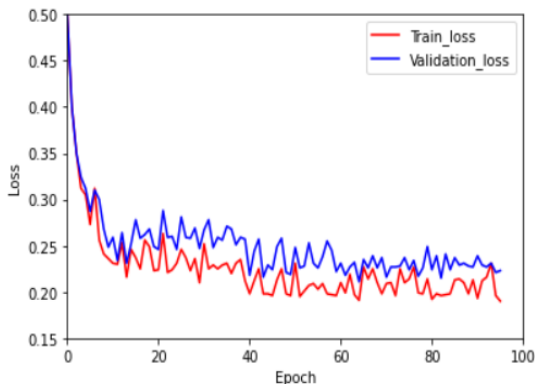
**Figure 10.** Category-wise average precision of various detectors to the proposed



**Figure 11.** Comparison of mean average precision values of various detectors to the proposed



**Figure 12.** Training and validation accuracy curve



**Figure 13.** Training and validation loss curve

## 5. CONCLUSION

Detecting objects in drone images or UAV images is challenging due to their annihilating characteristics such as scale variances, occlusion, small and dense objects. Detectors with convolutional units as backbones have limited receptive fields and require more hyper parameter tuning. To solve these limitations an anchor-free transformer based network is proposed to detect the objects in aerial images. Pyramid vision transformer is used as backbone to extract features and a split attention module using ResNeSt block for cardinal grouping is embedded into the transformer network to learn distinguishable feature representations and acquire adequate contextual information from the preprocessing step producing the most dominant features from the backbone. The acquired features are fed to an anchor-free detection head with three branches classification, centerness and regression. To improve the accuracy and connectivity between classification and localization, two IoU balanced loss functions are used for prediction. This work introduces a transformer network PvSAMNet helping to increase the detection accuracy by 38.74 MAP where the other state of the art deep learning model Cascade++ produces 38.71 MAP. The proposed transformer model fares better in improving the MAP (mean average precision) value compared to the other state of the art detectors that resulted in MAP values less than 37. The obtained MAP is still below 40 representing that detection in aerial view images remains a ceaseless challenge in the detection field. Using other attention modules in the transformer networks can enhance the results capable of connecting even the long range dependencies to a far extent. Even though small and dense objects are detected with better accuracy, transformers are yet to be explored with other attention modules to handle scale variances of these special categories of images. Transformer based network models show a promising direction towards research in object detection. Adopting these anchor-free transformers based models to object detection for optimal results in videos can be future research.

## AUTHOR CONTRIBUTIONS

This study was conceptualized and designed by all authors. Mrs. Museboyina Sirisha defined the problem, which was verified and refined by the other coauthor Dr. S.V. Sudha. In addition to organizing chapters and designing the proposed work, implementation sections along with results, Mrs. Museboyina Sirisha and Dr. S.V. Sudha approved the final work. It is confirmed that the following contributions to the paper have been made: study conceptualization and design: Mrs. Museboyina Sirisha, Dr. S.V. Sudha Data collection: Mrs. Museboyina Sirisha Analysis of Results: Mrs. Museboyina Sirisha Draft Manuscript Preparation: Mrs. Museboyina Sirisha, Dr. S.V. Sudha. All authors reviewed and approved the final manuscript after reviewing the results.

## REFERENCES

- [1] Bazi, Y., Melgani, F. (2018). Convolutional SVM networks for object detection in UAV imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6): 3107-3118.

- <https://doi.org/10.1109/TGRS.2018.2790926>
- [2] Girshick, R. (2015). Fast R-CNN. In 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 1440-1448. <https://doi.org/10.1109/ICCV.2015.169>
  - [3] Ren, S.Q., He, K.M., Girshick, R., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
  - [4] He, K.M., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. In 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2961-2969. <https://doi.org/10.1109/ICCV.2017.322>
  - [5] Fang, W., Wang, L., Ren, P.M. (2019). Tinier-YOLO: A real-time object detection method for constrained environments. IEEE Access, 8: 1935-1944. <https://doi.org/10.1109/ACCESS.2019.2961959>
  - [6] Lin, T.Y., Goyal, P., Girshick, R., He, K.M., Dollár, P. (2017). Focal loss for dense object detection. In 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2980-2988. <https://doi.org/10.1109/ICCV.2017.324>
  - [7] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). SSD: Single shot multibox detector. In Computer Vision-ECCV 2016: 14th European Conference, Springer International Publishing, pp. 21-37. [https://doi.org/10.1007/978-3-319-46448-0\\_2](https://doi.org/10.1007/978-3-319-46448-0_2)
  - [8] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25(2): 1097-1105. <https://doi.org/10.1145/3065386>
  - [9] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv Preprint arXiv: 1409.1556. <https://doi.org/10.48550/arXiv.1409.1556>
  - [10] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. (2016). Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vegas, NV, USA, pp. 770-778. <https://doi.org/10.1109/CVPR.2016.90>
  - [11] Sirisha, M., Sudha, S.V. (2022). A novel deep learning-based object detector using SPOTNET-SNIPER network. In Mobile Computing and Sustainable Informatics, Springer, Singapore, pp. 627-639. [https://doi.org/10.1007/978-981-19-2069-1\\_43](https://doi.org/10.1007/978-981-19-2069-1_43)
  - [12] Anguraj, D.K., Thirugnanasambandam, K., Raghav, R.S., Sudha, S.V., Saravanan, D. (2021). Enriched cluster head selection using augmented bifold cuckoo search algorithm for edge-based internet of medical things. International Journal of Communication Systems, 34(9): e4817. <https://doi.org/10.1002/dac.4817>
  - [13] Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Lake City, UT, USA, pp. 7132-7141. <https://doi.org/10.1109/CVPR.2018.00745>
  - [14] Behera, A., Wharton, Z., Liu, Y.H., Ghahremani, M., Kumar, S., Bessis, N. (2020). Regional attention network (RAN) for head pose and fine-grained gesture recognition. IEEE Transactions on Affective Computing, 14(1): 549-562. <https://doi.org/10.1109/TAFFC.2020.3031841>
  - [15] Woo, S., Park, J., Lee, J.Y., Kweon, I.S. (2018). Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19. [https://doi.org/10.1007/978-3-030-01234-2\\_1](https://doi.org/10.1007/978-3-030-01234-2_1)
  - [16] Ranftl, R., Bochkovskiy, A., Koltun, V. (2021). Vision transformers for dense prediction. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 12179-12188. <https://doi.org/10.1109/ICCV48922.2021.01196>
  - [17] Liu, Z., Lin, Y.T., Cao, Y., Hu, H., Wei, Y.X., Zhang, Z., Lin, S., Guo, B.N. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 10012-10022. <https://doi.org/10.1109/ICCV48922.2021.00986>
  - [18] Wang, W.H., Xie, E., Li, X., Fan, D.P., Song, K.T., Liang, D., Lu, T., Luo, P., Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 568-578. <https://doi.org/10.1109/ICCV48922.2021.00061>
  - [19] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90. <https://doi.org/10.1145/3065386>
  - [20] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. (2014). Microsoft coco: Common objects in context. In European Conference on Computer Vision, Springer, Cham, pp. 740-755. [https://doi.org/10.1007/978-3-319-10602-1\\_48](https://doi.org/10.1007/978-3-319-10602-1_48)
  - [21] Law, H., Deng, J. (2018). CornerNet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 734-750. [https://doi.org/10.1007/978-3-030-01264-9\\_45](https://doi.org/10.1007/978-3-030-01264-9_45)
  - [22] Zhou, X.Y., Wang, D.Q., Krähenbühl, P. (2019). Objects as points. arXiv Preprint arXiv: 1904.07850. <https://doi.org/10.48550/arXiv.1904.07850>
  - [23] Tian, Z., Shen, C.H., Chen, H., He, T. (2019). FCOS: Fully convolutional one-stage object detection. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 9627-9636. <https://doi.org/10.1109/ICCV.2019.00972>
  - [24] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. (2020). End-to-end object detection with transformers. In European Conference on Computer Vision, Springer, Cham, pp. 213-229. [https://doi.org/10.1007/978-3-030-58452-8\\_13](https://doi.org/10.1007/978-3-030-58452-8_13)
  - [25] Zhang, K., Sun, M., Han, T.X., Yuan, X.F., Guo, L., Liu, T. (2017). Residual networks of residual networks: Multilevel residual networks. In IEEE Transactions on Circuits and Systems for Video Technology, 28(6): 1303-1314. <https://doi.org/10.1109/TCSVT.2017.2654543>
  - [26] Dong, X.Y., Bao, J.M., Chen, D.D., Zhang, W.M., Yu, N.H., Yuan, L., Chen, D., Guo, B.N. (2022). CSWin transformer: a general vision transformer backbone with cross-shaped windows. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12124-12134. <https://doi.org/10.1109/CVPR52688.2022.01181>
  - [27] Wang, W.X., Yao, L., Chen, L., Lin, B.B., Cai, D., He, X.F., Liu, W. (2021). CrossFormer: A versatile vision transformer based on cross-scale attention. arXiv Preprint arXiv: 2108.00154.

- <https://doi.org/10.48550/arXiv.2108.00154>
- [28] Zhang, H., Wu, C.R., Zhang, Z.Y., Zhu, Y., Lin, H.B., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., Smola, A. (2022). ResNeSt: Split-attention networks. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Orleans, LA, USA, pp. 2736-2746. <https://doi.org/10.1109/CVPRW56347.2022.00309>
- [29] Xie, S.N., Girshick, R., Dollár, P., Tu, Z.W., He, K.M. (2017). Aggregated residual transformations for deep neural networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1492-1500. <https://doi.org/10.1109/CVPR.2017.634>
- [30] Wu, S.K., Yang, J.R., Wang, X.G., Li, X.P. (2022). Iou-balanced loss functions for single-stage object detection. *Pattern Recognition Letters*, 156: 96-103. <https://doi.org/10.1016/j.patrec.2022.01.021>
- [31] Du, D.W., Zhu, P.F., Wen, L.Y., Bian, X., Lin, H.B., Hu, Q.H., Peng, T., Zheng, J.Y., Wang, X.Y., Zhang, Y., Bo, L.F., Shi, H.L., Zhu, R., Kumar, A., Li, A.J., Zinollayev, A., Askergaliyev, A., Schumann, A., Mao, B.J., Lee, B., Liu, C., Chen, C.R., Pan, C.H., Huo, C.L., Yu, D., Cong, D.C., Zeng, D.N., Pailla, D.R., Li, D., Wang, D., Cho, D., Zhang, D.Y., Bai, F.R., Jose, G., Gao, G.Y., Liu, G.Z., Xiong, H.T., Qi, H., Wang, H.R., Qiu, H.Q., Li, H.L., Lu, H.C., Kim, I, Kim, J., Shen, J., Lee, J., Ge, J., Xu, J.J., Zhou, J.K., Meier, J., Choi, J.W., Hu, J.H., Liu, Z.M. (2019). VisDrone-DET2019: the vision meets drone object detection in image challenge results. In 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), pp. 213-226. <https://doi.org/10.1109/ICCVW.2019.00030>
- [32] Li, Z.M., Peng, C., Yu, G., Zhang, X.Y., Deng, Y.D., Sun, J. (2017). Light-Head R-CNN: In defense of two-stage object detector. arXiv Preprint arXiv: 1711.07264. <https://doi.org/10.48550/arXiv.1711.07264>
- [33] Lin, T.Y., Dollár, P., Girshick, R., He, K.M., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 2117-2125. <https://doi.org/10.1109/CVPR.2017.106>
- [34] Cai, Z.W., Vasconcelos, N. (2018). Cascade R-CNN: Delving into high quality object detection. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Lake City, UT, USA, pp. 6154-6162. <https://doi.org/10.1109/CVPR.2018.00644>
- [35] Tan, M.X., Pang, R.M., Le, Q.V. (2020). Efficientdet: Scalable and efficient object detection. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 10781-10790. <https://doi.org/10.1109/CVPR42600.2020.01079>
- [36] Cao, Y.R., He, Z.J., Wang, L.J., Wang, W.G., Yuan, Y.X., Zhang, D.W., Zhang, J.L., Zhu, P.F., Gool, L.V., Han, J.W., Hoi, S., Hu, Q.H., Liu, M., Cheng, C., Liu, F.F., Cao, G.J., Li, G.Z., Wang, H.K., He, J.Y., Wan, J.F., Wan, Q., Zhao, Q., Lyu, S.C., Zhao, W.Z., Lu, X.Q., Zhu, X.K., Liu, Y.J., Lv, Y.X., Ma, Y.J., Yang, Y.T., Wang, Z., Xu, Z.Y., Luo, Z.P., Zhang, Z.M., Zhang, Z.G., Li, Z.H., Zhang, Z.X. (2021). VisDrone-DET2021: The vision meets drone object detection challenge results. In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, pp. 2847-2854. <https://doi.org/10.1109/ICCVW54120.2021.00319>