NOVEL OF ANCHOR ADAPTATION AND STACKED SPARSE AUTOENCODER FOR TINY OBJECT DETECTION

Mrs. G. Rubadevi[1],
Research Scholar,
Department of Computer Science,
PSGR Krishnammal College for Women

Dr. R. Divya[2],
Assistant Professor,
Department of Computer Science,
PSGR Krishnammal College for Women

ABSTRACT: Recent advancements in earth vision object detection highlight challenges in tiny object detection, primarily due to class imbalance between foreground and background, inadequate semantic signals, and limited pixel information. Current object detectors struggle with small objects due to the lack of discriminative feature supervision, leading to suboptimal results. In aerial object detection tasks, research utilizing anchor-based two-stage detectors has significantly improved performance, leading to divergence in object features and impacting network learning. This study presents the Feature Enhanced Attention Module (FEAM), Anchor Adaption Region Proposal Network Head (A2RPH), and Stacked Sparse Autoencoder (SSAE). High-level features are unsupervised learnt by the SSAE from unlabelled aerial images. In order to increase the discriminability of learnt features, supervised learning is also applied to refine the feature representation. To improve the model, a logistic regression classifier is fed these high-level features. In particular, A2RPH enables better positive and negative sample assignments in the Region Proposal Network (RPN) by performing anchor adaptive learning by creating a new anchor bias learning branch from the feature map. In order to achieve better feature representation, FEAM presents Gaussian mask supervision for attention and introduces global features and mask attention based on FPN.

INDEX TERMS: Stacked Sparse Autoencoder, Feature Enhanced Attention Module (FEAM), aerial images, anchor adaption, and tiny object detection.

1. INTRODUCTION

Object detection is a key aspect of computer vision, particularly for identifying tiny objects in aerial images. Applications include vehicle recognition and traffic monitoring. Early research utilized conventional feature extraction methods, focusing on shape, color, and texture. Techniques involved using sliding windows of different scales to identify candidate object regions and classifying them based on stored object features. However, there are a number of drawbacks to traditional object detection methods, including window redundancy and ineffective candidate region selection. These disadvantages make it difficult to identify small targets in remote sensing

images, and conventional object detection methods are limited to images with straightforward backgrounds and important features.

The majority of deep neural networks are focused on recognizing objects of typical size, despite the fact that object detection has made tremendous progress [1]. While tiny objects (less than 16 × 16 pixels) in the Tiny Object Detection in Aerial Images (AI-TOD) dataset [2] in aerial images frequently display incredibly limited appearance information, learning discriminative features presents significant challenges and results in massive failure cases when detecting tiny objects [3–4]. Finding little objects in aerial images is still difficult, though.The reason is that most detectors are designed for regularsized objects. When encountering tiny objects that lack structural details due to their small size, fixed anchors and common feature representations impede network learning. Specifically, the sensitivity of the intersection and union (IoU) to tiny objects [5-7] causes the detection network with fixed anchor cannot perform high-quality assignment of samples in Region Proposal Network (RPN). Moreover, since the tiny object has fewer visual features, it is easy to diverge during the feature representation process.

Faster Region-based Convolutional Neural Network (Faster R-CNN), Feature Pyramid Networks (FPN) [8], and Mask R-CNN [9] are examples of two-stage detectors that achieve state-of-the-art accuracy by first generating region suggestions and then refining classification and localisation. Simultaneously, one-step approaches, such as the Single Shot Detector (SSD) [10] and the You Only Look Once (YOLO) [11], emphasise speed by doing away with the proposal stage. One-stage detectors are excellent at training quickly, but as a trade-off in design, they often have trouble identifying small objects. Super-resolved representations of small objects can be directly generated using Generative Adversarial Networks (GANs) [12]. In an effort to improve the subpar performance of traditional object detectors on micro-objects, Feature Pyramid Networks (FPNs) have also been proposed to learn multi-scale features [13].

The Stacked Sparse Autoencoder (SSAE) uses unsupervised learning to extract high-level features from unlabelled aerial images. Supervised learning is then used to improve discriminability, and a logistic regression classifier is used to fine-tune the results. By incorporating an anchor bias learning branch into the feature map, Anchor Adaption Region Proposal Network Head (A2RPH) facilitates anchor-adaptive learning and enhances positive and negative sample assignment in the RPN. Stronger feature representations are produced by the Feature Enhanced Attention Module (FEAM), which incorporates global features and mask-based attention into the FPN and uses Gaussian mask supervision to direct the attention mechanism.

2. LITERATURE REVIEW

Xu et al., [9] proposed a novel RanKing-based Assigning (RKA) technique and a Normalised Wasserstein Distance (NWD) for detecting small objects. The typical Intersection over Union (IoU) threshold-based detector can be readily replaced by the suggested NWD-RKA technique, which greatly improves label assignment and provides enough supervision information for network training. NWD-RKA may reliably increase microscopic item detection performance by a significant amount, according to tests conducted on four datasets. Additionally, the Tiny Object Detection in Aerial Images (AI-TOD) dataset's noticeable noisy labels inspired careful relabelling, the release of AI-TOD-v2 and its matching benchmark. The location error and missing annotation issues are significantly reduced in AI-TOD-v2, enabling more dependable training and validation procedures.

Kim et al., [14] developed a You Only Look Once (YOLOv8), Convolutional Block Attention Module (CBAM), Squeeze-and-Excitation Block (SE Block), and Mixture of Orthogonal Neural-modules Network (MoonNet) for Tiny Object Detection. Additionally, show how enlarging an image and using augmentation correctly can result in enhancement. created a MoonNet pipeline with attention-augmented CNNs as well. When two popular attention modules—the SE Block and the CBAM—were added to the YOLOv8 backbone with more channels, the MoonNet backbone outperformed the original YOLOv8 in terms of detection accuracy. AP50, AP, recall, and precision were used to assess performance.

Chen et al., [15] introduced a small object detection network for aerial photography built on an enhanced classifier module and a deformable cross-attention module (DCENet). In particular, a self-designed deformable cross-attention module improves object location identification without being impacted by increased background noise interference by adaptively focussing shallow feature maps on Regions of Interest (RoI) in deep feature maps. In the meanwhile, the issue of small objects having lower resolution and appearing more blurry than ground-view objects is resolved by enlarging the cropped images using a Crop-Images Super-Resolution Module (CSRM). Additionally, an improved classifier module is used to improve the model's classification capacity for a few related categories, which raises the network's overall performance. The module combines the ResNet-34 classifier with the CSRM. According to the experimental results, DCENet achieves state-of-the-art performance on the Unmanned Aerial Vehicle Benchmark Object Detection and Tracking (UAVDT) and VisDrone-2019 datasets, respectively, with mean average precision higher values. This implies that aerial image detection is a better fit for the suggested DCENet. Average Precision (AP), Precision, and Recall are the performance evaluation metrics.

Yu et al., [16] formulated a Diagonal Feature Pyramid (DFP) to find small flaws. DFP is suggested to enhance tiny defect detection performance in the backbone network. If additional original characteristics are at the same level, DFP fuses them for greater accuracy. DFP eliminates the bottom-up pathway and several non-original same-level characteristics to minimise the model size at a lower computational cost. To support multi-scale microscopic defect detection, a multi-scale neck network with certain fusion techniques is created. Lastly, to modify the sensitivity of small flaws, an adaptive localisation loss function is created. Comparative experiments using a publicly available Printed Circuit Board (PCB) dataset demonstrate that the suggested model outperforms several popular defect detection techniques in terms of mean Average Precision (mAP) and speed.

Han et al., [17] designed a Sampling-Balance-Based Multistage Network (SB-MSN) for object recognition in aerial images. It adaptively mines high-quality positive and negative occurrences, utilizing a multiscale information retention module, an intersection over union balance sampling approach, a balance L1 loss, and a multistage network to enhance detector training. Evaluated on three High-Resolution Remote Sensing (HRRS) datasets, including the Northwestern Polytechnical University Very High Resolution 10-class remote sensing image dataset (NWPU VHR-10 dataset), the detector effectively addresses low-quality samples and significantly improves mean Average Precision (mAP) in detection performance.

Su et al., [18] suggested a tiny item detector using a Global Context Self-Attention and Dense Nested Connection Feature Extraction Network (GC-DN Network). First, present the GC-DN Network, a light backbone-heavy neck design that effectively extracts and merges the target's multi-scale features. Second, suggest replacing the IoU in label assignment strategies, regression loss functions, and Non-Maximum Suppression (NMS) with a new metric called Mixed Minimum

Point-Wasserstein distance (MMPW). In MMPW, the MMPW Distance is utilized to assess the similarity of boxes by modeling bounding boxes as 2D Gaussian distributions. Testing on the latest aerial image tiny object datasets, AI-TOD and VisDrone-19, indicates that this methodology improves AP50 and AP performance. This confirms the effectiveness of the suggested method for identifying small things in aerial images.

Chirgaiya and Rajavat [19] presented for precise small item detection, a Competitive Multi-Layer Neural Network (CMLNN) architecture consists of three subcomponents: a competitive multi-layer network, a TOD auxiliary, and a multi-level continuous features aggregation. The suggested architecture is based on competitive learning for object detection. Comparison investigation with existing Region-based Convolutional Neural Network (RCNN), Fast RCNN, Faster RCNN, Single Shot MultiBox Detector (SSD) and YOLO reveals considerable improvement in the results. When compared to state-of-the-art detectors, CMLNN achieves remarkable accuracy in terms of mAP.

Guan et al., [20] adopted a Region-Based Efficient Network for precise image object detection. First, a framework was created to provide precise, high-quality suggestions that were independent of class. In order to learn convolutional features, these recommendations were then incorporated into the suggested network together with their input images. A network refinement module decreased the amount of proposals to only a few viable candidate proposals in order to increase detection efficiency. The detection module processed revised candidate proposals for item classification, evaluated using the PASCAL Visual Object Classes Challenge 2007 test set. Results indicate that the model significantly enhances detection efficiency, outperforming existing methods with fewer proposals, as evidenced by improvements in recall, Mean Average Best Overlap (MABO), and mean Average Precision (mAP).

Ming et al., [21] proposed a Sparse Label Assignment Strategy (SLA) to enhance object recognition in aerial images by selecting high-quality sparse anchors through posterior IoU of detections. This approach minimizes the gap between regression and classification, leading to better performance through balanced training. The subsequent phase involves utilizing a Position-Sensitive Feature Pyramid Network (PS-FPN) with a coordinate attention module for accurate localization. Lastly, the introduction of the distance rotated IoU loss aims to align the evaluation measure with training loss for improved bounding box regression. The superiority of the suggested method is demonstrated by extensive testing on the University of Chinese Academy of Sciences high resolution-Aerial Object Detection (UCAS-AOD), High Resolution Ship Collections 2016 (HRSC2016), and Dataset for Object deTection in Aerial images (DOTA).

3. PROPOSED METHODOLOGY

Anchor Adaption Region Proposal Network Head (A2RPH), Feature Enhanced Attention Module (FEAM), and Stacked Sparse Autoencoder (SSAE) have been created for the detection of small objects. High-level features are unsupervised learnt by the SSAE from unlabelled aerial images. In order to increase the discriminability of learnt features, supervised learning is also applied to refine the feature representation. To improve the model, a logistic regression classifier is fed these high-level features. In particular, A2RPH enables higher-quality positive and negative sample assignments in RPN by performing anchor adaptive learning by creating a new anchor bias learning branch from the feature map. In order to achieve better feature representation, FEAM presents Gaussian mask supervision for attention and introduces global features and mask attention based on FPN are shown in figure 1.

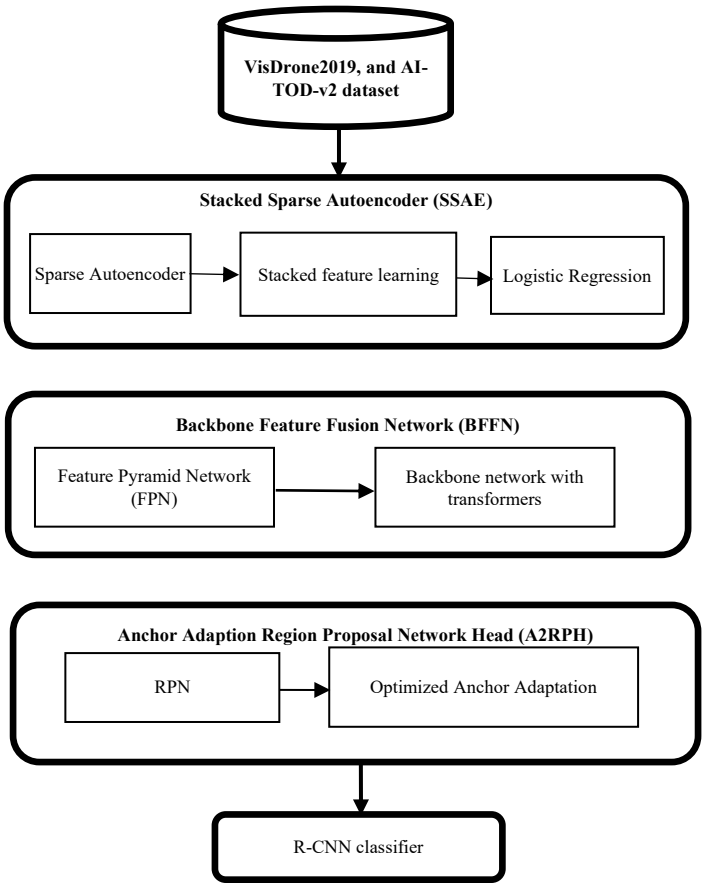

FIGURE 1. OVERALL PROCESS OF PROPOSED MODEL

3.1. Stacked Sparse Autoencoder (SSAE)

SSAE, or Autoencoder (AE), is a feedforward nonlinear neural network comprising three main layers: input, hidden, and output. Each layer consists of nodes that are fully connected to adjacent layers. AE processes input through an encoder-decoder system, where the input vector is encoded to connect the input layer with the hidden layer. During decoding, the model reconstructs the input vector from the learned features in the hidden layers. AE is employed to establish input aerial image representations that facilitate optimal reconstruction. For effective feature extraction, the hidden layer's dimension must be smaller than that of the input layer to avoid trivial error minimization solutions. An alternative method known as Sparse Autoencoder (SAE) imposed sparsity regularisation on the AE hidden layers instead of limiting the hidden layer dimension. The regularisation of the hidden layer's replies is implemented by SAE in order to prevent the basic AE's simplistic solutions. Sparsity regularisation is applied to the AE, and those fundamental AEs required that the dimension of the hidden layer be smaller than the dimension of the input layer. For each input node, only the most appropriate hidden node responses are used to drive the SAE to represent the training aerial image in sparse features in order to strike a balance between the sparsity of the hidden layer and reconstruction error. It can be stated by equation (1),

$$\underset{w,b,\hat{w},\hat{b}}{\operatorname{argmin}} \sum_{i=1}^{N} \left| x_i - \left(\hat{w}\left(\sigma(wx_i - b)\right) + \hat{b}\right|_2^2 + \delta \sum_{j=1}^{M} KL(\rho|\rho^j) \right. \tag{1}$$

$$KL(\rho|\rho^j) = \rho \log \frac{\rho}{\rho^j} + (1-\rho) \log \frac{1-\rho}{1-\rho^j} \quad (2)$$

In this context, δ refers to the balancing parameter that influences the trade-off between sparsity and reconstruction, with M defining the hidden layer's dimensions. The Kullback-Leibler [22] divergence, denoted as $KL(\rho|\rho^j)$, quantifies the difference between two Bernoulli distributions characterized by probabilities ρ and ρ^j. Sparsity is minimized when ρ^j approaches ρ for each hidden neuron j. Additionally, the Stacked Autoencoder (SAE) is capable of identifying low-level features from aerial image patches of various objects. Low-level feature learning is insufficient due to the variations in small object appearances. The Stacked Sparse Autoencoder (SSAE) combines several Sparse Autoencoders (SAEs) to extract high-level features from input aerial image patches. The model utilizes an unsupervised technique for pretraining, leveraging overlapping patches without requiring labeled aerial images.

In the described process, a Stacked Autoencoder (SAE) is employed for feature learning from training aerial images. Initially, the overlapped patches are adjusted with a weight $W^{(1)}$ to train the first SAE for obtaining representation activations $h_1^{(1)}(x)$. The secondary representations $h_2^{(2)}(x)$ are then derived by utilizing these primary representations as input for another SAE, which is adjusted with weight $W^{(2)}$. These secondary representations serve as input for a sigmoid layer, enabling the mapping to labels through further adjustment with weight $W^{(3)}$. Ultimately, the architecture comprises one input layer, two hidden layers for the SAE, and a final sigmoid output layer specifically designed to distinguish vertebrae from the background[23]. SSAE model employs a bottom-up unsupervised training approach, followed by a supervised sigmoid classifier for top layer refinement, with the number of nodes in the sigmoid layer corresponding to the number of labels. In this method, the sigmoid layer consists of two nodes—one for vertebra and another for the background. It predicts the likelihood of the input data label based on learned features and the second hidden layer representation. Despite the introduction of a Multilayer Perceptron (MLP) with many layers and nodes, challenges such as overfitting and local minima remain. The sigmoid logistic regression aids in jointly optimizing the entire deep framework through fine-tuning shown in equation (3),

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (3)$$

The sigmoid output function σ, used for classification, allows the jointly optimized weights and biases of SSAEs and the sigmoid layer to fine-tune the model. A gradient descent-based model minimizes the cost function to compute two output values that represent the categorization probabilities of input aerial images x_i. After high-level feature learning, the sigmoid layer receives the learned representations and labels. Test patches are inputted into the trained model, which outputs a binary classification (0 or 1) indicating the presence of small objects in the image patch[23].

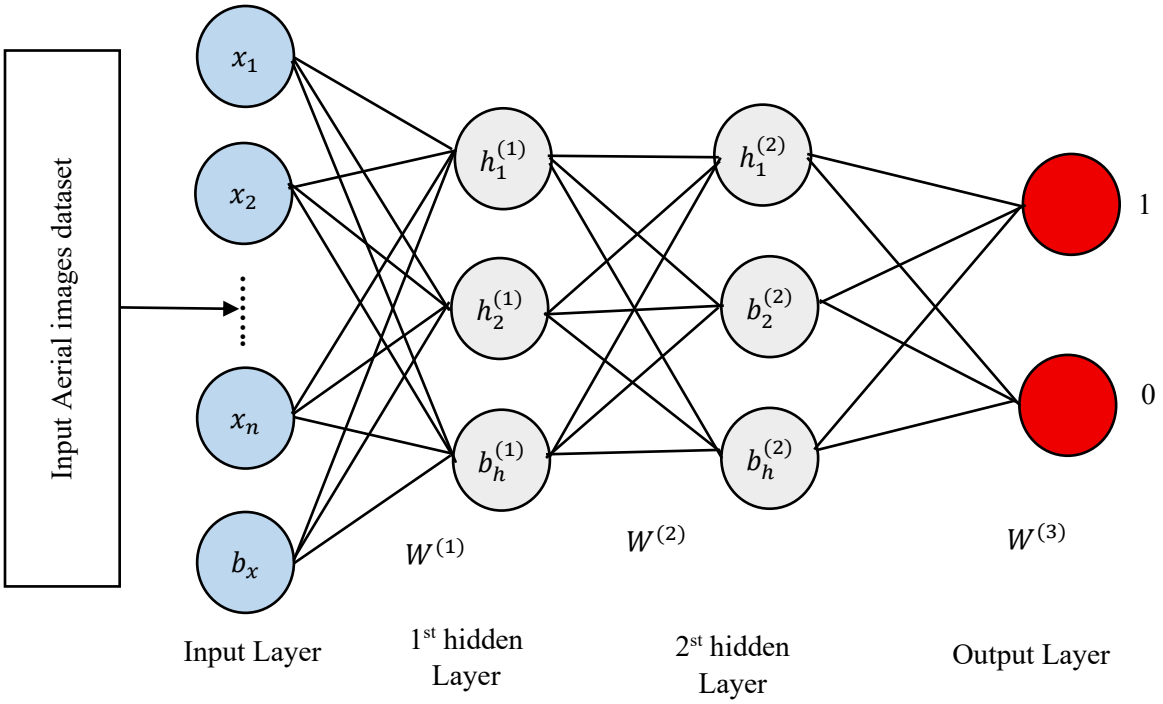

FIGURE 2. STACKED SPARSE AUTOENCODER (SSAE)

3.2. Anchor Adaption Region Proposal Network Head (A2RPH)

Creating prospective object areas to offer useful recommendations in the second stage of object classification and bounding box regression is the function of the Region Proposal Network (RPN), a crucial part of the two-stage object detector [24]. RPN uses the downsampling rate of each layer's feature map to create a collection of anchor boxes A in the image dimension. The 4-tuple form $a = (a_x, a_y, a_w, a_h)$, represents each anchor box a. The center of the box is denoted by x, y, while the length and width of the box are denoted by w, and h, respectively. A2RPH is RPN, which anchor adaption technique is used to address the issue of fixed anchors with reduced detection performance for tiny objects. By learning a bias toward the ground truth from the initial anchor, it is able to create learnt anchors that can assign higher quality positive and negative samples to tiny objects. As a formal anchor, the learnt anchor contributes in the regression branch as region proposal and loss computation. In particular, a, l, and t is denoted as the original anchor, learned anchor, and ground truth, correspondingly. The transformation ϑ from a to l is predicted by the anchor bias learning branch by equation (4),

$$\vartheta_x = (l_x - a_x)/a_w, \vartheta_y = (l_y - a_y)/a_{h,} \vartheta_w = \log\left(\frac{l_w}{a_w}\right), \vartheta_h = \log\left(\frac{l_h}{a_h}\right) \quad (4)$$

Anchor bias learning branch f' takes the image feature x as input which returns the prediction $\hat{\vartheta} = \mathbf{f}'(\mathbf{x})$for bias optimization. Because bias learning has the development to the ground truth, the target $\theta$for supervised anchor bias stable optimization learning is approximated using εtimes δ by equation (5),

$$\mathcal{L}_{ab}(\hat{\vartheta}, \vartheta) = \mathcal{L}(\hat{\vartheta}, \varepsilon\delta) = \sum_{k\epsilon\{x,y,w,h\}} L_1(\hat{\vartheta}_k - \varepsilon\delta_k) \quad (5)$$

where δ is the transformation from the original anchor to the ground truth, $\mathcal{L}_{ab}$ is the loss function for anchor bias, and $\varepsilon \in [0,1]$ is the anchor bias rate. A2RPH degenerates into an RPN head when ε is 0, and ε is 1, the anchor bias learning branch contributes to all of proposal boxes. The transformation δ^* to the ground truth is then calculated by substituting the learned anchor for the original anchor by equation (6),

$$\delta_x^* = \frac{t_x - l_x}{l_w}, \delta_y^* = \frac{t_y - l_y}{l_h}, \delta_w^* = \log(t_w/l_w), \delta_h^* = \log(t_h/l_h) \tag{6}$$

Similarly, smallest bounding box loss for regression branch prediction δ^* by equation (7),

$$\mathcal{L}_{reg}^*(\hat{\delta}, \delta^*) = \sum_{k\epsilon\{x,y,w,h\}} L_1(\hat{\delta}_k - \delta_k^*) \tag{7}$$

where the new bounding box regression loss is represented by $\mathcal{L}_{reg}^*$, remaining components of A2RPH are identical to those of RPN.

3.3. Feature Enhanced Attention Module (FEAM)

FEAM, a feature representation capability augmentation module based on FPN was developed to address the issue of feature divergence to tiny objects [25]. It consists of two sub-modules: masks attention and context enhanced. Globally coordinating features introduce greater contextual information in the context enhanced module. In the context enhanced, the features of each layer output by the backbone are globally coordinated using global average pooling and point convolution. The feature is added after the first projection layer of FPN. To achieve a more robust feature representation, the mask attention learns paying attention effectively through mask supervision. The features produced by FPN are dot-multiplied by the mask attention, which is created via deformable convolution and point convolution. In order to effectively eliminate background noise interference and focus on feature learning for the objects, mask supervision was added to the mask attention. The binary mask is a typical mask kind of supervisory technique. It uses either 0 or 1 to indicate the spatial distribution information of the item and background in the image. Consequently, improve the supervised mask and propose the probability-based 2-D Gaussian mask. It converts each ground truth's binary mask into a Gaussian mask using equation (8).

$$f(X_j|\mu, \Sigma) = \frac{\exp\left[-\frac{1}{2}(X_j - \mu)^T \Sigma^{-1}(X_j - \mu)\right]}{2\pi|\Sigma|^{1/2}} \tag{8}$$

where $f(X_j|\mu, \Sigma) \in [0, 1]$. A vector (x, y) denoting the j^{th} 2-D coordinate in the ground truth is called X_j. Gaussian distribution with mean vector and covariance matrix are denoted by μ and Σ. Consequently, the following expression is used for the ground truth mask MA by equation (9),

$$MA_i' = \sum_{j=1}^{M} (I_i^0 \Leftarrow f(X_j|\mu, \Sigma)), MA \leftarrow MA'[MA' > 1] = 1 \tag{9}$$

MA and MA′have the same size and dimensions, where MA_i is the i^{th} ground truth mask. For every image, M is the number of ground truths. I_i^0 is a template of the same size as MA_i that has been initialized to 0. The expression ($I_i^0 \Leftarrow f(X_j \mid \mu, \Sigma)$) indicates that the conversion result of

each ground truth should be placed at the location that corresponds toI_i^0. Consequently, the following equation (10) is used for the loss function to maximize mask attention,

$$\mathcal{L}_{ma} = \frac{1}{N}\sum_{i=1}^{N} CEL(\hat{m}_i, m_i) \tag{10}$$

where N is the total number of mask samples, $\mathcal{L}_{ma}$ is is the loss of mask supervision, $CEL(\cdot)$ is denoted as cross entropy loss, and $\hat{m}_i$and m_i is denoted as the probability of the ground truth mask and the predicted attention mask, respectively. Gaussian mask reduces the noise feature contribution of the edge portion of object while more clearly differentiating instances than the binary mask. For the fixed anchor limitations and poor representation ability of the network for tiny objects, optimize the Faster R-CNN network by AFS, called AFSNet, It eliminates the fixity of reset anchors, making anchors more adaptive for each scale object. FEAM based on FPN, which context enhanced module and mask attention module is introduced to get stronger feature representation ability for the network.

4. RESULTS AND DISCUSSION

The PyTorch deep learning framework and the MMDetection module were used to implement every experiment in this study [26]. The OpenMMLab community created MMDetection, an open-source object detection toolset. A workstation with an Intel® multi-core CPU, 64 GB RAM, and a single NVIDIA RTX 3090 GPU with 24 GB VRAM was used for all of the studies. A workstation with a single NVIDIA RTX 3090 GPU was used for model training and inference. Stochastic gradient descent (SGD) is used as the optimiser in all trials, with a batch size of two, a momentum parameter of 0.9, and a weight decay parameter of 0.0001. Apply non-maximum suppression (NMS) with a threshold of 0.5 to produce the top 3000 bounding boxes sorted by confidence during inference, then use a predetermined threshold of 0.05 to filter out background boxes. An improved version of the AI-TOD dataset, AI-TOD-v2 [9], keeps the same average instance size and image count as AI-TOD. By resolving problems like positional mistakes and missing annotations in the dataset, it encourages more dependable network training and produces mAP on the test set. The AISKYEYE team of Tianjin University's Machine Learning and Data Mining Laboratory gathered the VisDrone2019 [27] dataset.

Precision: Precision is to assess of correctness for assessing the performance of a classifier. A high precision indicates a lower number of false positives by Equation (11),

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

Recall: Recall is a measure to establish the totality of the classifier. Equation (12) can be used to compute it,

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

F-Measure: F-measure is calculated as the product of recall and precision is divided by the sum of recall and precision by Equation (13),

$$F - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

Accuracy: Accuracy quantifies the percentage of accurate model predictions among all forecasts by equation (14),

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{14}$$

TABLE 1. RESULTS ANALYSIS OF TINY OBJECT DETECTION METHODS

AI-TOD-v2 (%)				
Methods	**Precision**	**Recall**	**F-Measure**	**Accuracy**
CMLNN	85.16	87.63	86.38	86.24
DCENet	86.43	88.95	87.67	87.57
SB-MSN	88.18	90.87	89.49	89.39
SSAE+R-CNN	89.62	92.47	91.02	90.85
VisDrone2019 (%)				
Methods	**Precision**	**Recall**	**F-Measure**	**Accuracy**
CMLNN	84.29	87.26	85.75	85.17
DCENet	85.45	88.38	86.87	86.69
SB-MSN	86.73	90.24	88.45	88.11
SSAE+R-CNN	89.18	92.85	90.99	90.75

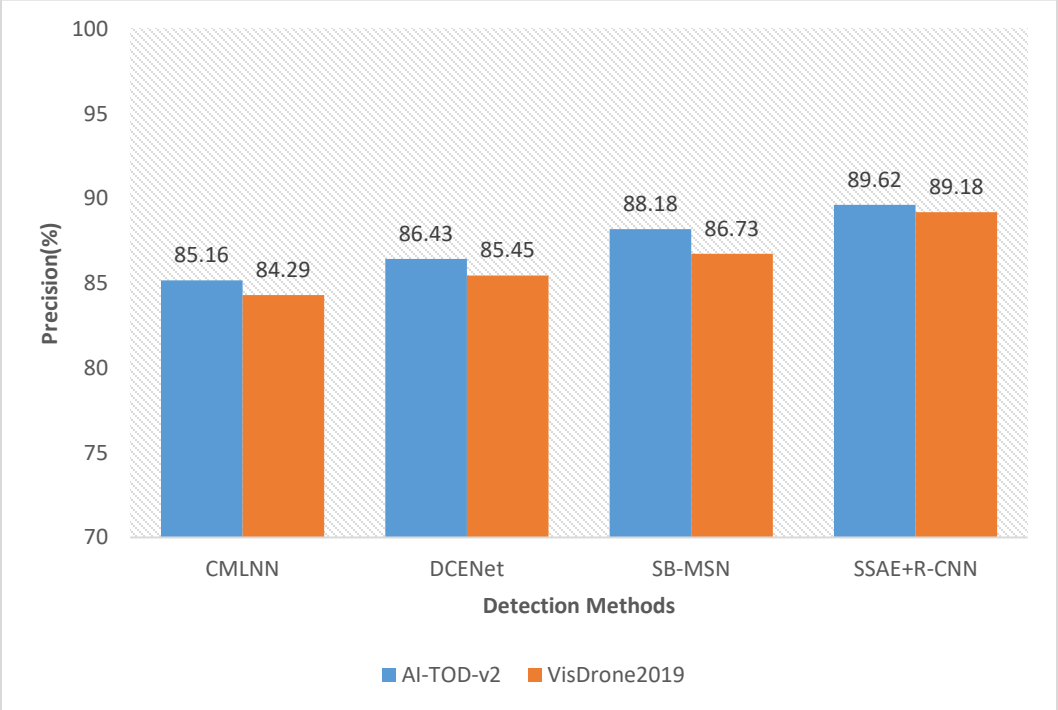

FIGURE 3. PRECISION COMPARISON OF DETECTION METHODS

Figure 3 shows the precision comparison of detection methods of CMLNN, DCENet, SB-MSN, and SSAE+R-CNN with respect to AI-TOD-v2, and VisDrone2019. The dataset 1, CMLNN, DCENet, SB-MSN, and SSAE+R-CNN gives the results of 85.16%, 86.43%, 88.18%, and 89.62%. The dataset 2, CMLNN, DCENet, SB-MSN, and SSAE+R-CNN gives the results of 84.29%, 85.45%, 86.73%, and 89.18%. The dataset 1, CMLNN, DCENet, and SB-MSN has increased results of 4.46%, 3.19%, and 1.44% when compared to proposed classifier. The dataset 2, CMLNN, DCENet, and SB-MSN has increased results of 4.89%, 3.73%, and 2.45% when compared to proposed classifier. Figure 4 shows the recall comparison of detection methods of CMLNN, DCENet, SB-MSN, and SSAE+R-CNN with respect to AI-TOD-v2, and VisDrone2019. The dataset 1, CMLNN, DCENet, SB-MSN, and SSAE+R-CNN gives the results of 87.63%, 88.95%, 90.87%, and 92.47%. The dataset 2, CMLNN, DCENet, SB-MSN, and SSAE+R-CNN gives the results of 87.26%, 88.38%, 90.24%, and 92.85%. The dataset 1, CMLNN, DCENet, and SB-MSN has increased results of 4.84%, 3.52%, and 1.60% when compared to proposed classifier. The dataset 2, CMLNN, DCENet, and SB-MSN has increased results of 5.59%, 4.47%, and 2.61% when compared to proposed classifier.

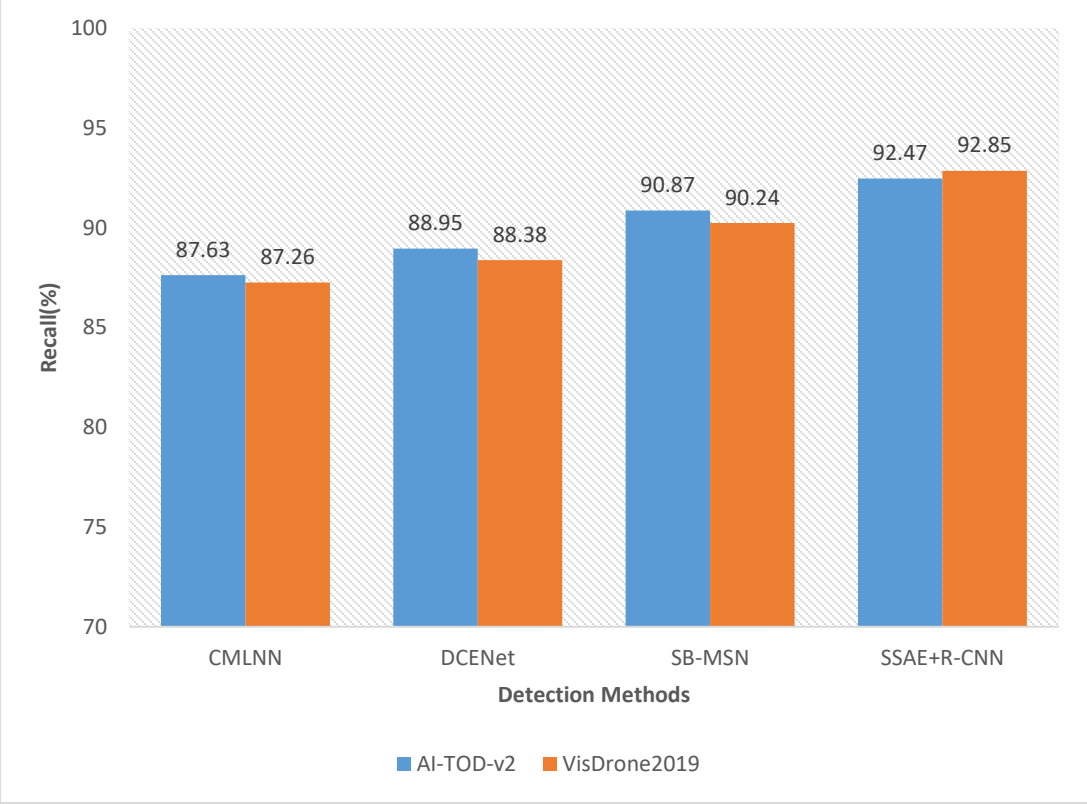

FIGURE 4. RECALL COMPARISON OF DETECTION METHODS

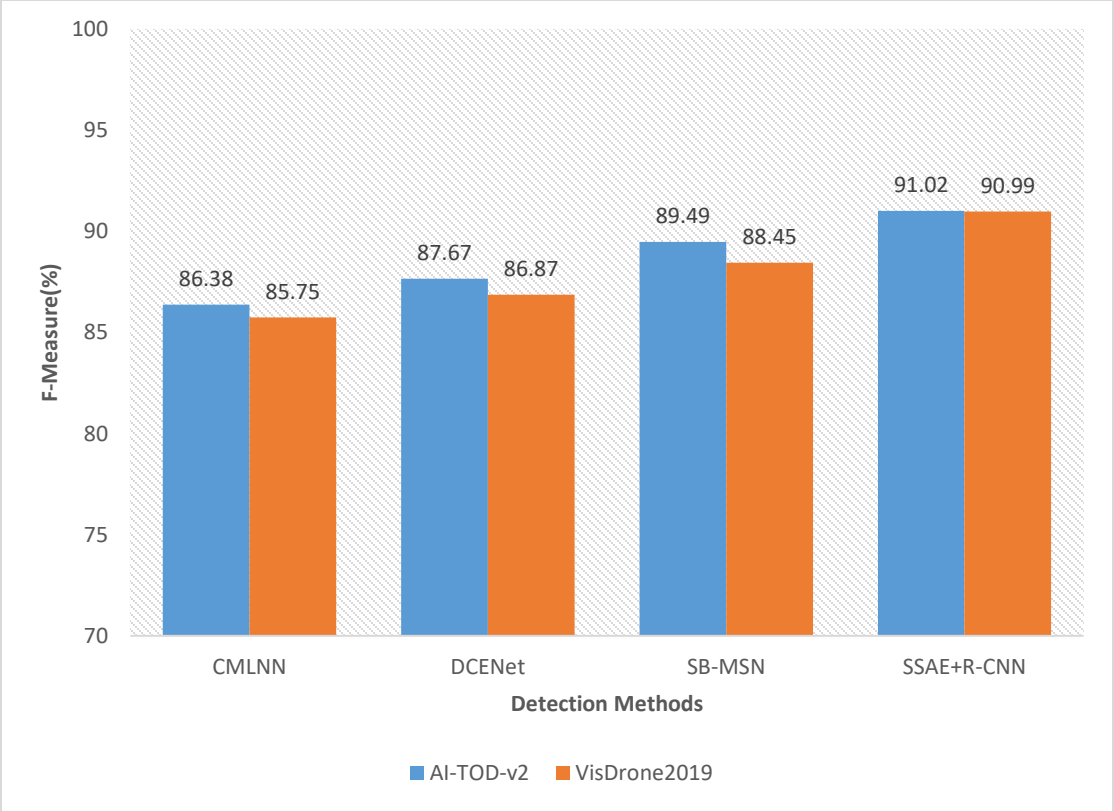

FIGURE 5. F-MEASURE COMPARISON OF DETECTION METHODS

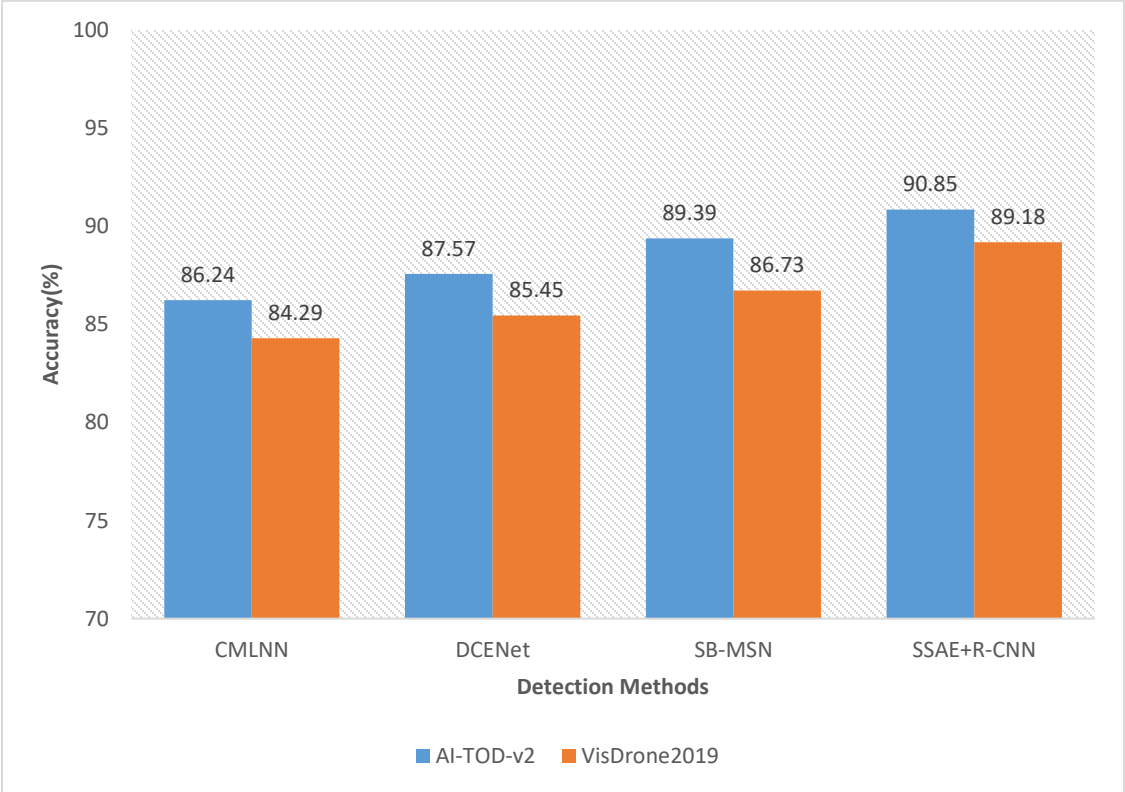

FIGURE 6. ACCURACY COMPARISON OF DETECTION METHODS

F-measure comparison of detection methods of CMLNN, DCENet, SB-MSN, and SSAE+R-CNN with respect to AI-TOD-v2, and VisDrone2019 are illustrated in figure 5. The dataset 1, CMLNN, DCENet, SB-MSN, and SSAE+R-CNN gives the results of 86.38%, 87.67%, 89.49%, and 91.02%. The dataset 2, CMLNN, DCENet, SB-MSN, and SSAE+R-CNN gives the results of 85.75%, 86.87%, 88.45%, and 90.99%. The dataset 1, CMLNN, DCENet, and SB-MSN has increased results of 4.64%, 3.35%, and 1.53% when compared to proposed classifier. The dataset 2, CMLNN, DCENet, and SB-MSN has increased results of 5.24%, 4.12%, and 2.54% when compared to proposed classifier.

Accuracy comparison of detection methods of CMLNN, DCENet, SB-MSN, and SSAE+R-CNN with respect to AI-TOD-v2, and VisDrone2019 are shown in figure 6. The dataset 1, CMLNN, DCENet, SB-MSN, and SSAE+R-CNN gives the results of 86.24%, 87.57%, 89.39%, and 90.85%. The dataset 2, CMLNN, DCENet, SB-MSN, and SSAE+R-CNN gives the results of 84.29%, 85.45%, 86.73%, and 89.18%. The dataset 1, CMLNN, DCENet, and SB-MSN has increased results of 4.61%, 3.28%, and 1.46% when compared to proposed classifier. The dataset 2, CMLNN, DCENet, and SB-MSN has increased results of 4.89%, 3.73%, and 2.45% when compared to proposed classifier.

5. CONCLUSION AND FUTURE WORK

In this paper, Stacked Sparse Autoencoders (SSAE), and Anchor Adaptation is able to be jointly employed to improve localization accuracy and feature representation for tiny objects. Anchor Adaptation focuses on redesigning or dynamically adjusting anchor boxes to better match the scale, aspect ratio, and distribution of tiny objects in the dataset. Anchor adaptation techniques analyze ground-truth bounding box statistics and generate smaller, scale-aware anchors, or adapt anchors during training using learned offsets. This leads to improved IoU matching, higher positive sample generation, and enhanced sensitivity to small-scale objects. SSAE are deep unsupervised feature learning models that learn compact and discriminative representations by enforcing sparsity constraints on hidden layers. In tiny object detection, SSAEs help extract subtle and high-level semantic features from low-resolution object regions that are otherwise difficult to distinguish from background noise. By stacking multiple sparse autoencoders, hierarchical feature representations are learned, capturing fine-grained textures and structural patterns essential for recognizing tiny objects. When combined, anchor adaptation enhances precise localization, while SSAE strengthens feature representation. The adapted anchors ensure that tiny objects are effectively proposed during the detection stage, and the SSAE-refined features improve classification confidence and robustness. This integrated framework significantly increases detection results for tiny objects, making it suitable for applications such as aerial imagery analysis, remote sensing, surveillance, and traffic monitoring. Future will focus on effective solution to overcome the inherent limitations of tiny object detection in complex visual scenes.

REFERENCES

1. Tian, Z., Shen, C., Chen, H. and He, T., 2019. FCOS: Fully convolutional onestage object detection. In: IEEE International Conference on Computer Vision, pp. 9627–9636.
2. Wang, J., Yang, W., Guo, H., Zhang, R. and Xia, G.-S., 2021. Tiny object detection in aerial images. In: International Conference on Pattern Recognition, pp. 3791–3798.
3. Singh, B., Najibi, M. and Davis, L. S., 2018. Sniper: Efficient multiscale training. In: Advances in Neural Information Processing Systems, pp. 9310–9320.
4. Yu, X., Gong, Y., Jiang, N., Ye, Q. and Han, Z., 2020. Scale match for tiny person detection. In: IEEE Workshops on Applications of Computer Vision, pp. 1257–1265.

5. Xu, C., Wang, J., Yang, W. and Yu, L., 2021. Dot distance for tiny object detection in aerial images. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1192-1201.
6. Liu, H., Tong, Q., Qi, L., Hao, Y. and Liu, X., 2023, A double diagonal ratio metric for tiny object detection in aerial images. In 2023 International Conference on Computer Engineering and Distance Learning (CEDL), pp. 11-19.
7. Xu, C., Wang, J., Yang, W., Yu, H., Yu, L. and Xia, G.S., 2022. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 190, pp.79-93.
8. Wang, C. and Zhong, C., 2021. Adaptive feature pyramid networks for object detection. IEEE access, 9, pp.107024-107032.
9. Hassan, E. and El-Rashidy, N., 2022. Mask R-CNN models. Nile Journal of Communication and Computer Science, 3(1), pp.17-27.
10. Leng, J. and Liu, Y., 2019. An enhanced SSD with feature fusion and visual reasoning for object detection. Neural Computing and Applications, 31(10), pp.6549-6558.
11. Ma, M. and Pang, H., 2023. SP-YOLOv8s: An improved YOLOv8s model for remote sensing image tiny object detection. Applied sciences, 13(14), pp.1-17.
12. Zhao, B., Wang, C., Fu, Q. and Han, Z., 2020. A novel pattern for infrared small target detection with generative adversarial network. IEEE Transactions on Geoscience and Remote Sensing, 59(5), pp.4481-4492.
13. Zhou, S. and Zhou, H., 2024. Detection based on semantics and a detail infusion feature pyramid network and a coordinate adaptive spatial feature fusion mechanism remote sensing small object detector. Remote Sensing, 16(13), pp.1-23.
14. Kim, K., Lazarou, M. and Stathaki, T., 2025. Enhanced Detection of Tiny Objects in Aerial Images. Computer Vision and Pattern Recognition, pp.1-5.
15. Chen, S., Wen, M., Tian, Y., Xue, Y. and Wang, H., 2024. DCENet: a tiny object detection network for aerial images based on deformable cross-attention and enhanced classifier. Journal of Electronic Imaging, 33(6), pp.063032-063032.
16. Yu, Z., Wu, Y., Wei, B., Ding, Z. and Luo, F., 2023. A lightweight and efficient model for surface tiny defect detection. Applied Intelligence, 53(6), pp.6344-6353.
17. Han, W., Fan, R., Wang, L., Feng, R., Li, F., Deng, Z. and Chen, X., 2020. Improving training instance quality in aerial image object detection with a sampling-balance-based multistage network. IEEE Transactions on Geoscience and Remote Sensing, 59(12), pp.10575-10589.
18. Su, N., Zhao, Z., Yan, Y., Wang, J., Lu, W., Cui, H., Qu, Y., Feng, S. and Zhao, C., 2024. MMPW-Net: Detection of Tiny Objects in Aerial Imagery Using Mixed Minimum Point-Wasserstein Distance. Remote Sensing, 16(23), pp.1-22.
19. Chirgaiya, S. and Rajavat, A., 2023. Tiny object detection model based on competitive multi-layer neural network (TOD-CMLNN). Intelligent Systems with Applications, 18, pp.1-9.
20. Guan, Y., Aamir, M., Hu, Z., Abro, W.A., Rahman, Z., Dayo, Z.A. and Akram, S., 2021. A Region-Based Efficient Network for Accurate Object Detection. Traitement du Signal, 38(2), pp.481-494.
21. Ming, Q., Miao, L., Zhou, Z., Song, J. and Yang, X., 2021. Sparse label assignment for oriented object detection in aerial images. Remote Sensing, 13(14), pp.1-21.
22. Spineli, L.M., 2024. Local inconsistency detection using the Kullback–Leibler divergence measure. Systematic Reviews, 13(1), pp.1-11.
23. Aouedi, O., Piamrat, K. and Bagadthey, D., 2022. Handling partially labeled network data: A semi-supervised approach using stacked sparse autoencoder. Computer Networks, 207, pp.1-16.
24. Avola, D., Cinque, L., Diko, A., Fagioli, A., Foresti, G.L., Mecca, A., Pannone, D. and Piciarelli, C., 2021. MS-Faster R-CNN: Multi-stream backbone for improved Faster R-CNN object detection and aerial tracking from UAV images. Remote Sensing, 13(9), pp.1-18.
25. Cao, C., Wang, B., Zhang, W., Zeng, X., Yan, X., Feng, Z., Liu, Y. and Wu, Z., 2019. An improved faster R-CNN for small object detection. IEEE Access, 7, pp.106838-106846.

26. Cao, Y., He, Z., Wang, L., Wang, W., Yuan, Y., Zhang, D., Zhang, J., Zhu, P., Van Gool, L., Han, J. and Hoi, S., 2021. VisDrone-DET2021: The vision meets drone object detection challenge results. In Proceedings of the IEEE/CVF International conference on computer vision, pp. 2847-2854.