OBJECT DETECTION IN DRONE IMAGE

Pushkar Bhavsar[1], Sonal Fatangare[2], Darshan Chavan[3], Priya Bhalearo[4], Deven Govilkar[5]

[1]Department of Computer Engineering, SPPU University, Pune City, India.

[2]Department of Computer Engineering, SPPU University, Pune City, India.

[3]Department of Computer Engineering, SPPU University, Pune City, India.

[4]Department of Computer Engineering, SPPU University, Pune City, India.

[5]Department of Computer Engineering, SPPU University, Pune City, India.

ABSTRACT

Amidst the dynamic evolution of object detection technology tailored for unmanned aerial vehicles (UAVs), harnessing data from UAV aerial photographs has become remarkably convenient. With diverse applications spanning monitoring, geological exploration, precision agriculture, and disaster early warning, UAV-based object detection stands at the forefront of innovation. In recent strides, artificial intelligence, particularly deep learning, has emerged as the cornerstone of advancement in this domain. This paper embarks on a comprehensive review of recent breakthroughs in deep-learning-based UAV object detection. Offering a panoramic view of UAV development, it meticulously delineates the trajectory of deep-learning methodologies employed in object detection for UAVs. Moreover, it dissects pivotal challenges endemic to UAV object detection, including but not limited to, the nuances of detecting small objects, grappling with objects amidst complex backgrounds, addressing issues of object rotation and scale variance, and mitigating category imbalance quandaries. Within this discourse, the paper meticulously encapsulates a spectrum of innovative solutions rooted in deep learning, poised to surmount the challenges. Whether through novel architectures, data augmentation techniques, or tailored loss functions, these solutions represent a concerted effort to push the boundaries of UAV object detection efficacy. In conclusion, the paper deliberates on prospective avenues for research in the realm of UAV object detection. From refining existing methodologies to exploring interdisciplinary synergies with fields such as sensor fusion and reinforcement learning, the discourse offers a compass for navigating the uncharted territories of UAV object detection, illuminating pathways for future innovation and discovery.

KEYWORDS

Object detection, unmanned aerial vehicles, Deep learning, Computer vision.

1. INTRODUCTION

Object detection has long been a focal point in surveillance applications, primarily cantered around ground-based cameras. However, the emergence of camera-equipped drones has revolutionized the landscape, offering unparalleled flexibility, affordability, and compactness. Drones have swiftly overtaken satellites and conventional cameras across various domains like agriculture, aerial photography, delivery services, and surveillance. At the heart of drone intelligence lies object detection, a fundamental technology underpinning numerous intelligent algorithms such as segmentation, object tracking, and crowd estimation. The evolution of object detection algorithms spans several decades, marked by significant advancements in accuracy, speed, and efficiency. Initially, traditional methods dominated the field, relying on handcrafted features and classical machine learning

DOI: 10.5121/ijcsit.2024.02102

algorithms such as Haar cascades, Histogram of Oriented Gradients (HOG), and Deformable Part Models (DPM). While effective in certain scenarios, these techniques often struggled with complex scenes and varied object appearances. However, the landscape dramatically shifted with the advent of deep learning, particularly Convolutional Neural Networks (CNNs). This ushered in the era of the R-CNN family of algorithms, including R-CNN, Fast R-CNN, and Faster R-CNN, which integrated CNNs with region proposal techniques to significantly enhance both accuracy and speed.

Following this, Single Shot Detectors (SSDs) emerged as a class of methods aiming for real-time performance by directly predicting object bounding boxes and class probabilities in a single pass through the network. Notably, SSDs, along with variants like YOLO (You Only Look Once), became popular choices for their speed and decent accuracy. The field then witnessed a dichotomy between two-stage detectors, typified by Faster R-CNN, and one-stage detectors like SSD and YOLO, with the former prioritizing accuracy and the latter prioritizing speed. Recent trends have focused on developing more efficient architectures, such as EfficientDet and MobileNet, which strike a balance between model size, speed, and performance. Furthermore, the integration of transformer architectures, as seen in models like DETR (DEtection TRansformer), has shown promise in revolutionizing object detection tasks.

Despite these advancements, research in object detection remains vibrant, with ongoing efforts to enhance accuracy, efficiency, and robustness through techniques such as attention mechanisms, feature pyramid networks, and domain adaptation. Despite the soaring demand for drone-based object detection, progress has been hindered by formidable algorithmic challenges, posing a bottleneck in drone technology's advancement. The accuracy and real-time performance of object detection algorithms profoundly impact mission outcomes, influencing whether drones succeed or face destruction. Efficiency in object detection algorithms encompasses a diverse array of factors including speed, accuracy, resource utilization, scalability, robustness, ease of implementation, energy efficiency, and adaptability. Striking a delicate balance among these components is essential to ensure optimal performance across a multitude of applications and environmental conditions. Each facet plays a crucial role in shaping the algorithm's effectiveness, enabling it to detect objects reliably and efficiently amidst varying challenges and complexities.

By meticulously addressing these aspects, developers can craft algorithms that not only excel in their primary task but also exhibit versatility and resilience in real-world scenarios. Drone-based object detection confronts unique challenges, including the instability of fast-moving UAVs, the prevalence of small targets in images, continuous UAV motion, environmental fluctuations, and stringent real-time computing requirements.

These factors compound the difficulty of feature extraction, leading to blurred or false detections, especially when capturing fast-moving targets or tiny objects from high altitudes. Furthermore, the dynamic nature of the drone's surroundings, coupled with the presence of excessive background in images, exacerbates the challenge by introducing variability and noise into the detection process.

This necessitates robust algorithms capable of discerning relevant objects amidst changing environmental conditions while minimizing false positives and negatives. In response to these challenges, this survey paper aims to comprehensively explore recent advancements, hurdles, and ethical considerations in object detection and recognition, particularly in the context of drone imagery. By delving into methodological evolutions, benchmark datasets, evaluation metrics, and emerging trends, the paper strives to offer valuable insights into the

current landscape and future trajectories of object detection technologies. Ultimately, our goal is to foster a deeper understanding of the complexities surrounding the development of robust and ethical object detection systems for drone imagery, thereby advancing responsible and inclusive applications of AI in computer vision.

2. LITERATURE SURVEY

In this segment, we will talk about the literature survey which provides a concise exploration of object detection algorithms. Covering academic research and technological advancements, it offers insights into the evolution, trends, and challenges of this dynamic field. From classical image processing methods to the latest in deep learning, the survey functions as a concise reference for researchers and practitioners navigating the diverse terrain of computer vision research cantered around faces.

Han, J.; Zhang, D.; Cheng, G.; Liu, N.; Xu, D. describes the recent progress in this research field, including 1) definitions, motivations, and tasks of each subdirection;2) modern techniques and essential research trends; 3) bench-mark data sets and evaluation metrics; and 4) comparisons and analysis of the experimental results. More importantly, they reveal the underlying relationship among OD, SOD, and COD and discuss in detail some open questions as well as point out several unsolved challenges and promising future works.

Scale variation is one of the key challenges in object detection. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. first present a controlled experiment to investigate the effect of receptive fields for scale variation in object detection. Based on the findings from the exploration experiments, they propose a novel Trident Network (Trident Net) aiming to generate scale-specific feature maps with a uniform representational power. they construct a parallel multi-branch architecture in which each branch shares the same transformation parameters but with different receptive fields. Then, they adopt a scale-aware training scheme to specialize each branch by sampling object instances of proper scales for training. As a bonus, a fast approximation version of Trident Net could achieve significant improvements without any additional parameters and computational cost compared with the vanilla detector. On the COCO dataset, our Trident Net with ResNet-101 backbone achieves state-of-the-art single-model results of 48.4 mAP.

Angelova, A.; Zhu, S. propose an algorithm which combines region-based detection of the object of interest and full-object segmentation through propagation. The segmentation is applied attest time and is shown to be very useful for improving the classification performance on four challenging datasets. They tested their approach on the most contemporary and challenging datasets for fine-grained recognition improved the performances on all of them. They further tested with 578-category flower dataset which is the largest collection of flower species. The improvements in performance over the baseline are about 3-4%, which is consistent across all the experiments.

Ma, Y.; Wu, X.; Yu, G.; Xu, Y.; Wang, Y. proposed a pedestrian detection and tracking system. A two-stage blob-based approach is first developed for pedestrian detection. This approach first extracts pedestrian blobs using the regional gradient feature and geometric constraints filtering and then classifies the detected blobs by using a linear Support Vector Machine (SVM) with a hybrid descriptor, which sophisticatedly combines Histogram of Oriented Gradient (HOG) and Discrete Cosine Transform (DCT) features in order to achieve accurate detection.

Ren, S.; He, K.; Girshick, R.; Sun, J. introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for

detection. They further merge RPN and Fast R-CNN into a single network by sharing their convolutional features—using the recently popular terminology of neural networks with 'attention' mechanisms, the RPN component tells the unified network where to look. For the very deep VGG-16 model, there detection system has a frame rate of 5fps (including all steps) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007, 2012, and MS COCO datasets with only 300 proposals per image.

3. ALGORITHMS

The evolution of object detection algorithms unfolds through two distinct phases: traditional methodologies and deep-learning-based approaches. Within the realm of deep learning, these methods further branch into one-stage and two-stage algorithms, delineating different technical trajectories. Figure 1 offers a visual depiction of this developmental journey spanning from 2001 to 2023.Traditional object detection algorithms hinge upon sliding window techniques and manual feature extraction mechanisms. Typically, they entail three sequential steps: region proposal, feature extraction, and classification regression. Region proposal involves identifying potential regions of interest harboring objects. Subsequently, artificial feature extraction methods are applied to translate images within candidate regions into feature vectors. Finally, classification and regression techniques are employed to categorize objects based on the extracted features. However, these conventional algorithms suffer from several drawbacks, including high computational complexity, limited feature representation capabilities, and challenges in optimization. Representative examples include the Viola–Jones detector and the HOG pedestrian detector, which have played seminal roles in shaping the landscape of object detection. Despite their contributions, traditional algorithms face inherent limitations that impede their efficacy in handling complex real-world scenarios. The emergence of deep learning heralds a paradigm shift in object detection, offering unprecedented opportunities for advancement. Through the utilization of neural networks, deep-learning-based algorithms can autonomously learn hierarchical representations of features from raw data, circumventing the need for manual feature engineering. Within the realm of deep learning, one-stage and two-stage algorithms represent divergent methodologies in approaching object detection tasks. One-stage algorithms aim for simplicity and efficiency by directly predicting object bounding boxes and class labels in a single step. In contrast, two-stage algorithms employ a two-step process involving region proposal and refinement, thereby achieving higher accuracy at the cost of increased computational overhead. As we navigate through this evolutionary trajectory of object detection algorithms, it becomes evident that deep learning has revolutionized the field, offering unprecedented capabilities in handling complex real-world scenarios with enhanced efficiency and accuracy.

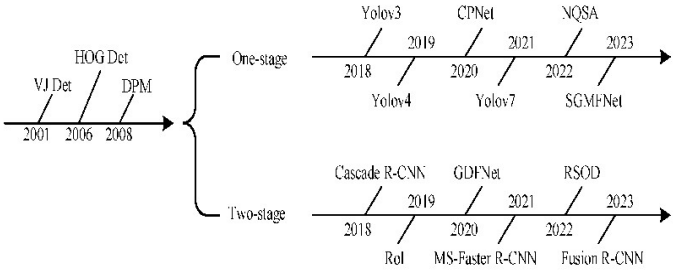

Figure 1. The development of object detection from 2001 to 2023.

In 2012, the advent of convolutional neural networks (CNNs) accelerated object detection, leveraging deep learning to automatically extract high-level features from images. Two-

stage detectors, like R-CNN and Faster R-CNN, excel in accuracy but suffer from computational inefficiency. Conversely, one-stage detectors such as SSD and Yolo prioritize speed but may struggle with small object localization and detailed feature capture.

3.1 One-Stage UAV Object Detection Algorithm

In the realm of UAV object detection, one-stage algorithms like Yolo and SSD revolutionize the landscape with their swift processing and high accuracy. Yolo, introduced by Redmon et al. in 2015, divides images into fixed-size grids, predicting bounding boxes and object probabilities directly. Similarly, SSD, proposed by Liu et al. in 2016, generates default bounding boxes across different scales, predicting object categories and positions. Their rapid execution and precision make Yolo and SSD highly sought after. Hossain et al. harnessed these algorithms on GPU JetsonTX2 for UAV ground object detection and tracking. Lu et al. fused Yolov5 with shallow features, enhancing efficiency in UAV marine fishery law enforcement. Marta utilized Yolo with dense point clouds to identify atypical aviation obstacles effectively. Further innovations abound, with Li et al. enhancing SSD with a convolutional block attention mechanism (SSD-CBAM) for earthquake disaster building detection. Scholars continually refine one-stage detectors through network optimization, multi-task learning, contextual information inclusion, and network fusion, amplifying their capabilities for multi-object detection under the UAV perspective.

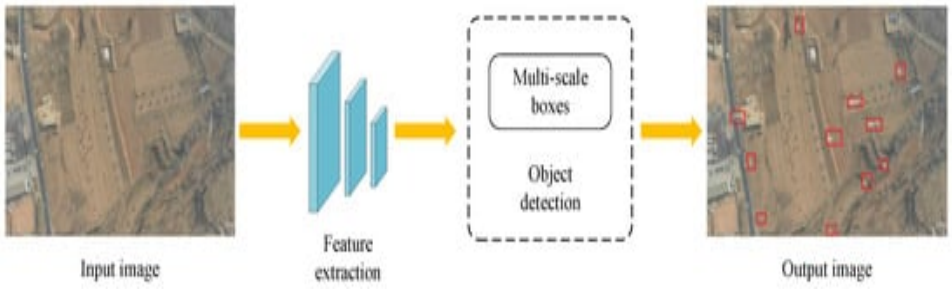

Figure 2. One-stage object detection framework.

3.2 The Two-Stage Object Detection Algorithm

In the realm of UAV object detection, the two-stage approach excels in accuracy by first proposing regions of interest (ROI) and then classifying them. However, this method tends to be slower due to additional processing stages and region proposals. In 2014, Girshick et al. pioneered the fusion of Region Proposal and CNN with their R-CNN algorithm, showcasing significant performance enhancements. He et al. innovated further by integrating the Spatial Pyramid Pooling (SPP) module into CNN, overcoming limitations of fixed-size images and redundant feature extraction.

To address the inefficiencies of R-CNN, Girshick introduced Fast R-CNN, leveraging ROI pooling for end-to-end detection. Ren et al. elevated the game with Faster R-CNN, replacing selective search with the region proposal network (RPN) for more efficient candidate region generation. By sharing convolutional features, this network enhances detection speed significantly Each advancement in the two-stage approach marks a stride

toward precision and efficiency in UAV object detection, laying the groundwork for increasingly sophisticated applications in the field.

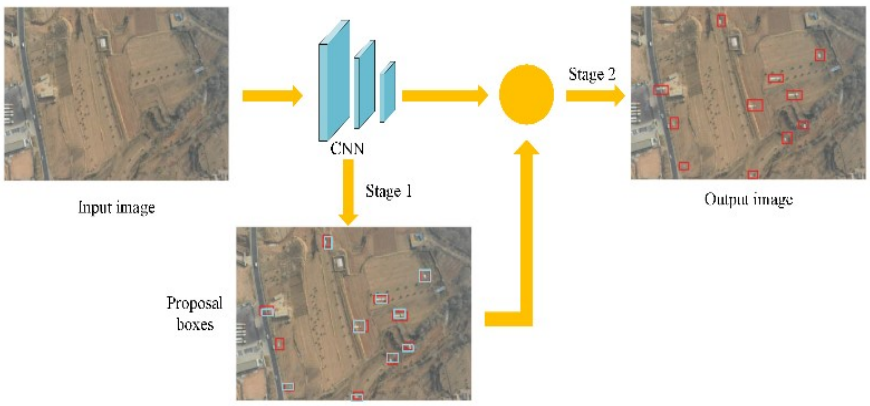

Figure 3. Two-stage object detection framework

4. ALGORITHMIC SURVEY

In the past two decades, it is widely accepted that the progress of object detection has generally gone through two historical periods: "traditional object detection period (before 2014)" and "deep learning-based detection period (after 2014)". In the following, we will summarize the milestone detectors of this period, with the emergence time and performance serving as the main clue to highlight the behind driving technology.

4.1. Viola Jones Detectors:

In 2001, Viola and Jones revolutionized face detection with a real-time algorithm that surpassed prior methods by leveraging integral images, Haar-like features, and detection cascades on a 700MHz Pentium III CPU, setting new standards in efficiency and accuracy without relying on skin colour segmentation. This seminal milestone in computer vision marked a transformative shift, paving the way for rapid advancements in the field and underlining the enduring impact of their innovative approach.

4.2.HOG Detector:

In 2005, Dalal and Triggs introduced Histogram of Oriented Gradients (HOG) feature descriptor, a pivotal advancement akin to the transformative Scale-Invariant Feature Transform and Shape Contexts of its era. Balancing feature invariance and nonlinearity, HOG computes on a dense grid with overlapping local contrast normalization, revolutionizing object detection, particularly in pedestrian detection. Its versatility extends to various object classes, with the HOG detector seamlessly adapting to different sizes through image rescaling while maintaining fixed window dimensions. This foundational technique has underpinned myriad object detectors and diverse computer vision applications, solidifying its enduring legacy in the field.

4.3. Deformable Part-based Model (DPM):

DPM, hailed as the champion of the VOC-07, -08, and -09 detection challenges, stands as a quintessential representation of traditional object detection methods. Conceived by P.

Felzenszwalb in 2008 as an extension of the HOG detector, DPM embodies a "divide and conquer" approach, wherein training involves learning the optimal decomposition of an object, while inference entails an ensemble of detections across its constituent parts. This philosophy, epitomized by the "star-model" introduced by Felzenszwalb et al., dissects objects like "cars" into windows, bodies, and wheels for detection. R. Girshick later expanded this concept into "mixture models," enhancing adaptability to real-world object variations and introducing a suite of improvements. While contemporary object detectors have surpassed DPM in accuracy, its enduring influence persists in modern techniques such as mixture models, hard negative mining, bounding box regression, and context priming, underscoring its invaluable contributions to the evolution of object detection methodologies.

4.4. RCNN:

The genesis of RCNN lies in its simplicity: initiating with the extraction of object proposals through selective search, generating a set of candidate boxes. Each proposal undergoes rescaling to a standardized size before being inputted into a pretrained CNN model like Alex Net, facilitating feature extraction. Subsequently, linear SVM classifiers come into play, discerning object presence within each region, and identifying object categories. RCNN heralded a substantial performance leap on VOC07, elevating mean Average Precision (mAP) from 33.7% (DPM-v5) to an impressive 58.5%. However, its efficacy is counterbalanced by glaring drawbacks: the computational redundancy arising from feature computation across numerous overlapped proposals (exceeding 2000 boxes per image) severely throttles detection speed (14s per image with GPU). Addressing this bottleneck, SPPNet emerged later in the same year, effectively resolving the issue and marking a pivotal advancement in object detection efficiency.

4.5. SPPNet:

In 2014, K. He et al. introduced Spatial Pyramid Pooling Networks (SPPNet), marking a significant advancement in convolutional neural network (CNN) architecture. Unlike previous models requiring fixed-size inputs, SPPNet's key innovation lies in the Spatial Pyramid Pooling (SPP) layer, enabling CNNs to generate fixed-length representations irrespective of image or region size, eliminating the need for rescaling. This breakthrough allows for feature map computation from the entire image just once, streamlining the process of generating fixed-length representations for training detectors and obviating the repetitive computation of convolutional features. Notably, SPPNet achieves over a 20-fold increase in speed compared to R-CNN without compromising detection accuracy (VOC07 mAP=59.2%). Despite its strides in enhancing detection speed, SPPNet still grapples with certain limitations: it retains a multi-stage training approach, and only fine-tunes its fully connected layers, disregarding earlier layers. Subsequently, in the following year, Fast R-CNN emerged, addressing these issues, and further refining the landscape of object detection.

4.6. Fast RCNN:

In 2015, R. Girshick introduced the Fast R-CNN detector, representing a notable evolution beyond both R-CNN and SPPNet. Noteworthy for its simultaneous training of a detector and bounding box regressor within unified network configurations, Fast R-CNN marks a significant leap forward in object detection. On the VOC07 dataset, it elevated the mean Average Precision (mAP) from 58.5% (R-CNN) to an impressive 70.0%, all while achieving detection speeds over 200 times faster than its predecessor. Despite amalgamating the strengths of R-CNN and SPPNet, Fast R-CNN still grapples with the bottleneck of proposal detection, thus prompting the question: "Can object proposals be

generated using a CNN model?" This query found its answer in the subsequent development of Faster R-CNN, further pushing the boundaries of object detection capabilities.

4.7. Faster RCNN:

In 2015, S. Ren et al. introduced the Faster R-CNN detector shortly after the Fast R-CNN, marking a watershed moment as the first near-real time deep learning detector. With impressive metrics such as COCO mAP@.5=42.7% and VOC07 mAP=73.2%, along with a remarkable speed of 17fps with ZF-Net, Faster R-CNN revolutionized the landscape of object detection. Its groundbreaking contribution lies in the introduction of the Region Proposal Network (RPN), enabling nearly cost-free region proposals and significantly enhancing efficiency. Transitioning from R-CNN to Faster R-CNN, numerous components of the object detection pipeline, including proposal detection, feature extraction, and bounding box regression, were seamlessly integrated into a unified, end-to-end learning framework. Despite breaking through the speed bottleneck of its predecessor, Fast R-CNN, Faster R-CNN still grapples with computation redundancy at the subsequent detection stage. Subsequent innovations, such as RFCN and Light head R-CNN, have sought to address these challenges, underscoring the relentless pursuit of efficiency and accuracy in object detection methodologies.

4.8. Feature Pyramid Networks (FPN):

In 2017, T.-Y. Lin et al. introduced the Feature Pyramid Network (FPN), revolutionizing object detection. Unlike previous methods, FPN utilizes a top-down architecture with lateral connections to effectively integrate features from all levels of a convolutional neural network (CNN). This approach improves object localization across scales, leading to significant advancements in detection accuracy. FPN has since become a cornerstone technology in modern detectors, setting new benchmarks without requiring additional complexity. Its versatility and effectiveness have reshaped the field of computer vision, inspiring numerous innovations in object detection.

4.9. You Only Look Once (YOLO):

In 2015, R. Joseph et al. introduced You Only Look Once (YOLO), marking a revolutionary shift in object detection within the deep learning era. YOLO distinguished itself as the pioneering one-stage detector, boasting unparalleled speed and efficiency. Unlike its predecessors, YOLO took a radically different approach by employing a single neural network to process the entire image in one pass. The hallmark of YOLO's speed lies in its remarkable performance: the fast variant operates at an impressive 155 frames per second (fps) with a VOC07 mean Average Precision (mAP) of 52.7%, while its enhanced iteration maintains a swift pace at 45fps while achieving a significantly improved mAP of 63.4%. This exceptional speed made YOLO a frontrunner in real-time object detection applications. However, despite its breakthrough in speed, YOLO faced challenges in localization accuracy, particularly with smaller objects, when compared to traditional two-stage detectors. Subsequent iterations of YOLO and the introduction of Single Shot Multibox Detector (SSD) sought to address this limitation, focusing on enhancing localization precision. Notably, YOLOv7, an evolution from the YOLOv4 team, represents a significant advancement in both speed and accuracy. By introducing innovative optimizations such as dynamic label assignment and model structure reparameterization, YOLOv7 achieves unparalleled performance, boasting speeds ranging from 5 to 160 fps while surpassing most existing object detectors in accuracy. In essence, YOLO's pioneering approach to single-stage detection, coupled with its relentless pursuit of speed and accuracy enhancements through successive iterations like YOLOv7, continues to redefine the

landscape of object detection, making it a cornerstone in real-time visual recognition systems.

4.10. Single Shot MultiBox Detector (SSD):

In 2015, W. Liu et al. introduced the Single Shot Multibox Detector (SSD), pioneering multi-reference and multi-resolution techniques for superior detection accuracy, especially with small objects. Unlike previous detectors, SSD operates across network layers, boosting both speed (59fps) and accuracy (COCO mAP@.5=46.5%). This unique approach redefines object detection, making SSD a cornerstone technology in computer vision.

4.11. Retina Net:

In 2017, T.-Y. Lin et al. addressed the longstanding accuracy gap between one-stage and two-stage detectors with the introduction of Retina Net. They identified the predominant issue of extreme foreground-background class imbalance in dense detectors as the primary obstacle. To mitigate this challenge, Retina Net introduced the innovative "focal loss" function, reshaping standard cross entropy loss to prioritize hard, misclassified examples during training. This unique approach empowers one-stage detectors to attain comparable accuracy to their two-stage counterparts while preserving exceptional detection speed, marking a transformative milestone in the field of object detection.

4.12. Corner Net:

In a departure from conventional methods relying on anchor boxes, H. Law et al. introduced Corner Net, which reimagines object detection as a key point prediction task. By predicting key points and then leveraging additional embedding information, Corner Net dynamically assembles bounding boxes, eliminating the need for extensive anchor boxes. This novel approach circumvents issues like category imbalance, hyper-parameter tuning, and prolonged convergence times associated with traditional methods. Corner Net's innovative paradigm shift yields superior performance compared to prevailing one-stage detectors, marking a significant advancement in the field of object detection.

4.13. CenterNet:

In 2019, X. Zhou et al. introduced CenterNet, a groundbreaking object detection framework. CenterNet is an anchorless object detection architecture. This structure has an important advantage in that it replace the classical NMS at the post process, with a much more elegant algorithm, that is natural to the CNN flow. By treating objects as single points and regressing all attributes directly from their centres, CenterNet eliminates complex post-processing steps like group-based key point assignment and NMS. Its simplicity and elegance enable integration of multiple tasks like 3D object detection and human pose estimation, all while achieving competitive detection results.

4.14. DETR:

In recent years, the transformative impact of Transformers has reshaped deep learning, especially in computer vision. Departing from traditional convolution operators, Transformers rely solely on attention mechanisms to overcome CNN limitations and achieve a global-scale receptive field. In 2020, N. Carion et al. introduced DETR, revolutionizing object detection by framing it as a set prediction task and leveraging Transformers for end-to-end detection. This marked a paradigm shift, eliminating the need for anchor boxes or points. Subsequently, X. Zhu et al. proposed Deformable DETR to address DETR's challenges, including long convergence times and limited performance on small objects. Deformable DETR achieves state-of-the-art performance on the MSCOCO

dataset, boasting a remarkable COCO mAP@.5 of 71.9%. This underscores its unique ability to enhance detection accuracy while maintaining efficiency, propelling object detection into a new era of effectiveness and scalability.

Table -1: Algorithmic Survey of Research Studies

Algorithm Name	**Accuracy** AP_{50}
Faster R-CNN, VGG-16	42.7
Fast R-CNN, VGG-16	35.9
R-FCN, ResNet-101	51.9
Couple Net, ResNet-101	54.8
Faster R-CNN G-RMI, Inception-ResNet-v2	55.5
Faster R-CNN+++, ResNet-101-C4	55.7
Faster R-CNN w FPN, ResNet-101-FPN	59.1
Faster R-CNN w TDM, Inception-ResNet-v2-TDM	57.7
Deformable R-FCN, Aligned-Inception-ResNet	58.0
Cascade R-CNN, ResNet-101-FPN	62.1
Mask R-CNN, ResNeXt-101	62.3
YOLOv2, DarkNet-53	57.9
YOLOv3, DarkNet-19	44

SSD300∗, VGG-16	43.1
SSD321, ResNet-101	45.4
RetinaNet500, ResNet-101	53.1
CornerNet512, Hourglass	57.8

5.Complexity Analysis

In the realm of deep learning, time complexity transcends traditional algorithmic analysis, focusing instead on the total training time and inference speed of models like SSD. While deep learning entails millions of computations, the parallel execution across thousands of neurons per layer optimizes computational efficiency. Notably, leveraging hardware like Nvidia GeForce GTX 1070i GPU can accelerate SSD training by a factor of ten.

Matrix multiplication in the base CNN's forward pass predominates in time consumption. Its complexity is contingent upon various factors including layer count, neuron quantity, filter sizes, feature map dimensions, and image resolution. Moreover, the ReLU activation function, operating quadratically for each neuron, further influences time complexity.

Considering these factors holistically, we can gauge the time complexity of the forward pass in the base CNN. This unique approach to evaluating time complexity reflects the intricacies of deep learning models, emphasizing the parallel nature of computation and the interplay of various architectural components.

$$\text{Timeforward}=\text{timeconvolution}+\text{timeactivation}$$
$$=O(\sum b=1Bxl-1.(h.h).xb.(sb.sb))+O(B.xc)=O(\text{weights})$$

Here, b denotes the index of the CNN layer, B is the total amount of CNN layers,x_b is the number of filters in the b_{th} layer, h is the filter width and height, x_c is the number of neurons, x_{b-1} is the total number of input channels of the b_{th} layer, s_b is the size of the output feature map.
It should be noted that five to ten percent of the training time is taken up by things like dropout,regression,batch normalisation,classification as well As for SSD's accuracy, it is determined by Mean Average Precision or mAP which is simply the average of APs over all classes from the area under the precision-recall curve. A higher mAP is an indication of a more accurate model.

6.CONCLUSION

As advancements in computing power continue to surge forward, driving the evolution of object detection technology based on deep learning, the pace of progress accelerates with remarkable momentum. This rapid advancement is fueled by an escalating demand for high-precision real-time systems, prompting researchers to explore an expansive array of avenues aimed at achieving both unparalleled accuracy and unparalleled efficiency in object detection. Novel architectures, feature extraction methods, and representations are

not only being continuously developed but also meticulously refined to ensure optimal performance across a broad spectrum of applications. Efforts to enhance processing speed, including strategies such as training from scratch and the adoption of anchor-free methods, are actively underway to meet the increasingly stringent demands of real-time applications. Moreover, addressing intricate challenges such as detecting small or occluded objects requires a multifaceted approach, wherein researchers amalgamate techniques from both one-stage and two-stage detectors to attain optimal results. Refinements in post-processing techniques, such as the fine-tuning of non-maximum suppression methods and the mitigation of negative-positive imbalance, further contribute to the elevation of object detection accuracy. Beyond the fundamental task of detection, there is a burgeoning emphasis on precise localization and classification confidence, propelling researchers to innovate fervently in these pivotal areas. The application landscape of object detection spans an eclectic array of fields, encompassing everything from security and military applications to transportation, medicine, and beyond.

The wide-ranging applicability of object detection has engendered the emergence of various specialized branches within the detection domain, each uniquely tailored to address specific challenges and requirements inherent to their respective domains. While recent advancements in object detection have undeniably been significant, the field continues to beckon with boundless opportunities for further development and refinement. The relentless pursuit of ongoing innovation ensures that object detection technology remains at the vanguard of diverse and evolving real-world applications, steadfastly pushing the boundaries of what's achievable and redefining the realms of possibility with each stride forward.

7.Future Scope

In In the realm of drone-based object detection, the future holds exciting prospects, particularly in optimizing real-time implementation for swift decision-making in dynamic scenarios, while considering the limitations of drone hardware. The focus will shift towards enhancing the framework's robustness against environmental variabilities, ensuring consistent performance across diverse lighting, weather, and landscape conditions. Researchers will explore transfer learning techniques across different drone scenarios to overcome the challenge of limited datasets, while integrating multi-sensor data like LiDAR or thermal imaging to enhance detection accuracy across varied environments. Additionally, investigating semi-supervised or unsupervised learning methods could reduce reliance on labelled data, ensuring adaptability to evolving scenarios with minimal manual annotation. As the field advances, considerations around human-drone interaction, ethical implications, and scalability to larger datasets and varied domains will become increasingly important for the framework's success and applicability. It's crucial for researchers to remain vigilant to evolving benchmarks and standards, continuously evaluating, and improving the framework in comparison to the latest metrics and challenges in the field. The optimization of real-time implementation for drone-based object detection systems will involve further exploration into lightweight algorithms that can efficiently utilize the limited computational resources available on drones. Techniques such as model pruning, quantization, and efficient network architectures will be crucial for achieving real-time performance without compromising on detection accuracy. Moreover, advances in hardware, such as specialized chips designed for deep learning inference on drones, will also play a significant role in enabling faster and more efficient object detection. Enhancing the robustness of object detection algorithms against environmental variabilities will require the development of novel data augmentation techniques that can simulate a wide range of lighting, weather, and landscape conditions. Generative adversarial networks (GANs) and domain adaptation methods will be valuable tools for generating synthetic data

that can improve the generalization capabilities of detection models. Additionally, techniques for online adaptation, where the model continuously updates its parameters based on incoming data during deployment, will be crucial for maintaining high detection performance in changing environments. The integration of multi-sensor data, such as LiDAR and thermal imaging, will provide complementary information that can improve the accuracy and reliability of object detection systems, especially in challenging scenarios such as low-visibility conditions or cluttered environments. Fusion techniques, such as sensor fusion networks and probabilistic fusion methods, will be essential for effectively combining information from different sensors while accounting for their respective uncertainties.

In addition to technical challenges, ethical considerations will play a significant role in the development and deployment of drone-based object detection systems. Privacy concerns, potential misuse of surveillance capabilities, and the impact on civil liberties will need to be carefully addressed through transparent governance frameworks and stakeholder engagement. Moreover, ensuring fairness and preventing bias in detection algorithms will be critical for avoiding unintended consequences, such as discriminatory or unjust outcomes.

As drone technology continues to advance and become more pervasive, the scalability of object detection systems to larger datasets and varied domains will be essential for their widespread adoption and practical utility. Scalable training algorithms, distributed computing frameworks, and cloud-based deployment architectures will enable efficient processing of large volumes of data and support the deployment of detection systems in diverse applications, from wildlife monitoring to infrastructure inspection.

8.REFERENCE

[1] Han, J.; Zhang, D.; Cheng, G.; Liu, N.; Xu, D. Advanced deep-learning techniques for salient and category-specific object detection: A survey. IEEE Signal Process. Mag. 2018, 35, 84–100.

[2] Dev, S.; Wen, B.; Lee, Y.H.; Winkler, S. Ground-based image analysis: A tutorial on machine-learning techniques and applications. IEEE Geosci. Remote Sens. Mag. 2016, 4, 79–93.

[3] Voigtlaender, P.; Krause, M.; Osep, A.; Luiten, J.; Sekar, B.B.G.; Geiger, A.; Leibe, B. Mots: Multi-object tracking and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7942–7951.

[4] Angelova, A.; Zhu, S. Efficient object detection and segmentation for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 811–818.

[5] Ma, Y.; Wu, X.; Yu, G.; Xu, Y.; Wang, Y. Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery. Sensors 2016, 16, 446.

[6] Liang, X.; Zhang, J.; Zhuo, L.; Li, Y.; Tian, Q. Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis. IEEE Trans. Circuits Syst. Video Technol. 2019, 30, 1758–1770.

[7] Wu, Z.; Suresh, K.; Narayanan, P.; Xu, H.; Kwon, H.; Wang, Z. Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 1201–1210.

[8] Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.

[9] Zhu, P.; Wen, L.; Du, D.; Bian, X.; Hu, Q.; Ling, H. Vision meets drones: Past, present and future. arXiv 2020, arXiv:2001.06303

[10] Bazi, Y.; Melgani, F. Convolutional SVM networks for object detection in UAV imagery. IEEE Trans. Geosci. Remote Sens. 2018, 56, 3107–3118.

[11] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 1137–1149.

[12] Lu, Q.; Liu, C.; Jiang, Z.; Men, A.; Yang, B. G-CNN: Object detection via grid convolutional neural network. IEEE Access 2017, 5, 24023–24031.

[13] Chu, J.; Guo, Z.; Leng, L. Object detection based on multi-layer convolution feature fusion and online hard example mining. IEEE Access 2018, 6, 19959–19967

[14] Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks [EB/OL]. arXiv 2018, arXiv:1605.06409.

[15] Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks [EB/OL]. arXiv 2018, arXiv:1605.06409.

[16] Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-aware trident networks for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6054–6063.

[17] Kong, T.; Yao, A.; Chen, Y.; Sun, F. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.

[18] Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

[19] Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.

[20] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.

[21] Chen, K.; Li, J.; Lin, W.; See, J.; Zou, J. Towards Accurate One-Stage Object Detection with AP-Loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.

[22] Li, S.; Yang, L.; Huang, J.; Hua, X.S.; Zhang, L. Dynamic Anchor Feature Selection for Single-Shot Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019.

[23] Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.

[24] Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.

[25] Law, H.; Teng, Y.; Russakovsky, O.; Deng, J. Cornernet-lite: Efficient keypoint based object detection. arXiv 2019, arXiv:1904.08900.

[26] Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6569–6578.

[27] Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 850–859.

[28] Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. arXiv 2019, arXiv:1904.07850.

[29] Yang, Z.; Liu, S.; Hu, H.; Wang, L.; Lin, S. Reppoints: Point set representation for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9657–9666.

[30] Shafiee, M.J.; Chywl, B.; Li, F.; Wong, A. Fast YOLO: A fast you only look once system for real-time embedded object detection in video. arXiv 2017, arXiv:1709.05943.

[31] Huang, L.; Yang, Y.; Deng, Y.; Yu, Y. Densebox: Unifying landmark localization with end to end object detection. arXiv 2015, arXiv:1509.04874.

[32] Zhu, C.; He, Y.; Savvides, M. Feature selective anchor-free module for single-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 840–849.

[33] ong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; Shi, J. Foveabox: Beyound anchor-based object detection. IEEE Trans. Image Process. 2020, 29, 7389–7398.

[34] Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9627–9636.

[35] Shrivastava, A.; Gupta, A.; Girshick, R. Training region-based object detectors with online hard example mining. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769.

[36] Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.

[37] Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.

[38] Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. Adv. Neural Inf. Process. Syst. 2014, 3, 2672–2680.

[39] *D*enton, E.; Chintala, S.; Szlam, A.; Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. arXiv 2015, arXiv:1506.05751.

[40] Li, C.; Wand, M. Combining markov random fields and convolutional neural networks for image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2479–2486

Authors

Prof. Sonal Fatangare
Project Guide and Prof. of Computer Engineering Department at RMD Sinhgad School of Engineering, SPPU, Pune.

Mr. Pushkar M. Bhavsar
Project Team Lead and B.E. Student at Department of Computer Engineering, RMD Sinhgad School of Engineering, SPPU, Pune.

Mr. Darshan S. Chavan
Project Research Fellow and B.E. Student at Department of Computer Engineering, RMD Sinhgad School of Engineering, SPPU, Pune.

Ms. Priya D. Bhalerao
Project Research Fellow and B.E. Student at Department of Computer Engineering, RMD Sinhgad School of Engineering, SPPU, Pune.

Mr. Deven M. Govilkar
Project Research Fellow and B.E. Student at Department of Computer Engineering, RMD Sinhgad School of Engineering, SPPU, Pune.