Proceedings of the 4th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/79/2025.20099

Advances and Challenges in Small Object Detection: A Comparative Analysis of State-of-the-Art Models and Future Directions

Zixuan Zhao[1,a,*]

[1]*Halıcıoğlu Data Science Institute, Data Sciences, University of California-San Diego, California, USA*
a. ziz057@ucsd.edu
**corresponding author*

Abstract: Object detection involves the precise and efficient identification and localization of multiple predefined object categories within images. With the advent of deep learning, both the accuracy and efficiency of object detection have significantly improved. Nevertheless, challenges remain in optimizing the performance of mainstream detection algorithms, improving the accuracy of small object detection, enabling multi-class detection, and developing lightweight models. In response to these challenges, this paper provides a comprehensive literature review, analyzing approaches to enhance mainstream object detection by exploring advancements in backbone networks, expanding the visual receptive field, feature fusion, and various training strategies. We also evaluate the performance of leading detection models across established datasets, identifying current limitations and proposing future research directions. These directions include enhancing small object representation in datasets, enriching semantic information, and improving model interpretability. Small object detection remains a critical focus in computer vision, and we anticipate the continued development of algorithms with higher accuracy and efficiency.

Keywords: Object detection, small object, deep learning

1. Introduction

Object detection is a fundamental task in computer vision that involves both identifying and localizing objects of interest within an image or video. It can be understood as the combination of two subtasks: localization, which determines the precise position of objects within an image, and classification, which assigns each detected object to a specific category.

Over the years, numerous approaches to object detection have been developed, evolving from traditional methods based on handcrafted features—such as Histogram of Oriented Gradients (HOG) and Haar cascades—to modern deep learning techniques powered by convolutional neural networks (CNNs). State-of-the-art models, including YOLO (You Only Look Once), Faster R-CNN (Region-Based Convolutional Neural Networks), and SSD (Single Shot Multibox Detector), have revolutionized the field by achieving high accuracy while supporting real-time processing.

Among the various challenges in object detection, small object detection has emerged as a critical subfield. Small objects are defined in two primary ways: (1) a relative size criterion, where the object's

DOI: 10.54254/2753-8818/79/2025.20099

size is less than 10% of the original image, and (2) an absolute size criterion, as exemplified by the MS COCO dataset, where a small object is defined as one with a resolution of less than 32x32 pixels. Detecting small objects is particularly challenging due to their limited resolution and subtle visual features. On the MS COCO dataset, for instance, detection accuracy for large objects is typically twice as high as that for small objects.

With the increasing demand for applications such as autonomous driving, medical imaging, and surveillance, the accurate detection of small objects has gained substantial attention. Addressing these challenges is crucial for advancing object detection models to meet the needs of real-world applications where small objects play a significant role.

The remainder of this paper is organized as follows: Section 2 reviews recent advancements in small object detection algorithms, focusing on four key areas: feature fusion, backbone networks, expanding the receptive field, and model training strategies. Section 3 evaluates the performance of several leading object detection models on widely recognized datasets. Finally, Section 4 discusses current challenges and outlines future research directions.

2. Methods

2.1. Feature Fusion

The rapid advancements in deep learning and computational power have significantly underscored the advantages of feature fusion in object detection, particularly for small object detection. Feature fusion involves combining features from different layers of a neural network, or even across multiple networks, to capture more detailed information about objects. This approach enhances detection accuracy, especially for small objects, by integrating both high-level semantic information and low-level spatial details. Among the various methods, deep learning-based feature fusion techniques have shown the greatest promise.

Traditional object detection methods typically processed single-layer features, which overlooked the rich high-level semantic information present in deeper layers—crucial for detecting objects of varying sizes. In particular, small objects often become lost in low-resolution feature maps, making their detection challenging. Feature Pyramid Networks (FPN) introduced a multi-scale feature fusion approach, combining deep-layer semantic information with shallow-layer spatial details. This architecture significantly improved small object detection accuracy on datasets like MS COCO, achieving a precision of 35.8%. FPN's architecture has since been widely adopted in modern detection algorithms, such as YOLOv3.

Despite the success of FPN, its manually designed architecture may limit its feature fusion efficiency. To address this, Ghiasi et al. [1] introduced NAS-FPN, which leverages Neural Architecture Search (NAS) technology [2] to automatically discover optimal feature pyramid architectures. Using reinforcement learning, a controller was trained to explore a vast search space of cross-scale connections, iteratively refining the architecture based on detection performance as a reward signal. Combined with backbone models in the RetinaNet framework [3], NAS-FPN achieved remarkable results, with detection accuracy reaching 48.3% on the MS COCO dataset.

While FPN primarily focuses on multi-scale feature fusion by merging high-level semantic features with low-level details, Detection with Enriched Semantics (DES) takes this concept further. DES enhances both low- and high-level features, incorporating a segmentation module to enrich semantic content in lower layers, and a global activation module to strengthen higher-level features. This method has notably improved performance, achieving an 84.3% accuracy on the PASCAL VOC 2007 dataset.

DOI: 10.54254/2753-8818/79/2025.20099

2.2. Backbone Networks

Many state-of-the-art object detection models, such as VGG-16, GoogLeNet, ResNet-50, and ResNet-101, rely on backbone networks that are often pre-trained on large datasets like ImageNet. These networks primarily extract features used for classification and segmentation. However, traditional backbones struggle to detect small objects, as they tend to generate low-resolution representations that may cause small objects to be misclassified or overlooked, thereby reducing overall detection performance.

To address these limitations, newer backbone networks, such as DetNet and DenseNet, have been developed to improve small object detection. DetNet, introduced by Li et al. [4], builds on ResNet-50 by modifying the later stages to maintain feature maps at a scale of 1/16 of the original image, as opposed to ResNet-50's 1/32. This ensures that small objects remain detectable, while the network retains more detailed edge information. Additionally, DetNet reduces computational cost by maintaining the same number of channels in later stages. This approach increased detection accuracy on the MS COCO dataset from 35.8% to 40.2%.

DenseNet [5] improves feature extraction through dense connections, where each layer is connected to all previous layers within a dense block. This maximizes information flow, reduces the number of parameters, and mitigates issues like gradient vanishing. DenseNet's efficient design improves training stability and feature extraction, especially for small objects. STDN (Scale-Transferable Detection Network) builds upon DenseNet by introducing a scale-transfer layer, generating large feature maps with minimal computational overhead. This innovation improved detection accuracy on the PASCAL VOC 2007 dataset from 78.6% (in DSSD) to 79.3%.

2.3. Receptive Field Expansion

In deep learning, the receptive field defines the area of an image a neuron in the deeper layers is sensitive to. An increased receptive field helps capture broader spatial context, which is particularly important for detecting small objects that require detailed spatial information. Expanding the receptive field allows a network to better distinguish small objects from their background, enhancing overall detection performance.

The Receptive Field Block (RFB) network, integrated into the SSD architecture by Liu et al. [6], improved small object detection by mimicking the human visual system. RFB leverages receptive fields of different sizes and merges them into a unified spatial structure, enhancing feature extraction while maintaining high speed. On the PASCAL VOC 2007 dataset, RFB increased detection accuracy to 80.5%, a notable improvement over the baseline SSD model.

Similarly, TridentNet [7] enhances receptive field expansion using a ResNet-101 backbone with dilated convolutions to improve multi-scale object detection. TridentNet employs a multi-branch architecture, where each branch uses different dilation rates, allowing the model to focus on features at different scales. This approach improves detection accuracy across various object sizes, while its weight-sharing mechanism reduces computational complexity and accelerates inference.

2.4. Optimization of Training Methods

Small object detection poses significant challenges, as most detection algorithms are optimized for general datasets where small objects are less common. This discrepancy often results in reduced accuracy when detecting small objects in real-world scenarios. To address this, certain training strategies have been adapted to better accommodate small object detection.

YOLOv2, for instance, considers image size during pre-training to bridge the gap between pre-training and detection tasks. By fine-tuning the network on 416416 pixel images (rather than the

DOI: 10.54254/2753-8818/79/2025.20099

standard 224224 from ImageNet), YOLOv2 improved detection accuracy by 12.2% on the PASCAL VOC 2007 dataset.

Another solution to size discrepancies is Scale Normalization for Image Pyramids (SNIP). SNIP optimizes small object detection by backpropagating gradients only for regions of interest that match the scale of the training data, preventing irrelevant areas from influencing training. This approach achieved 48.3% detection accuracy on MS COCO.

Generative Adversarial Networks (GANs) have also been explored for small object detection. Perceptual GAN, introduced by Li et al. [8], improves small object detection by transforming small object features into higher-resolution representations. Its generator creates enhanced small object features, while the discriminator distinguishes between these features and those of large objects. By leveraging a perceptual feedback mechanism, Perceptual GAN significantly improves detection accuracy, achieving 84% on the PASCAL VOC 2007 dataset—up from 73.2% with Faster R-CNN.

3. Performance Evaluation

In this section, as shown in Table 1, we present a comprehensive evaluation of several state-of-the-art object detection models, focusing on their performance across different datasets. Our analysis covers key metrics such as speed, accuracy, and the specific characteristics of each model, including YOLOv2, NAS-FPN, and Faster R-CNN. These models are evaluated on challenging datasets like TinyPerson and DOTA, both of which provide unique insights into small object detection and multi-scale detection tasks.

The TinyPerson dataset is designed specifically to test object detection systems on tiny objects within distant scenes and large backgrounds, a scenario that is often overlooked by conventional detection models. It contains 1,610 images, each featuring over 200 individuals, and includes a total of 72,651 manually annotated objects classified into five distinct categories. On the other hand, the DOTA dataset, one of the largest and most complex benchmarks for object detection in aerial imagery, comprises 2,806 images collected from various sensors and platforms. It includes 15 object categories such as ships, planes, storage tanks, sports fields, vehicles, and bridges. Both datasets offer challenging test beds for evaluating object detection systems in complex, cluttered environments with a wide variety of object scales and dense object arrangements.

YOLOv2, known for its speed, was the fastest model in our comparison, achieving 62 frames per second (FPS) on the DOTA dataset and 65 FPS on TinyPerson. However, its speed comes at the cost of accuracy, particularly on these more complex datasets. On DOTA, YOLOv2 achieved a relatively low mean Average Precision (mAP) of 14.3%, and on TinyPerson, its performance dropped further to 10.5%. This decline in accuracy is primarily due to YOLOv2's simpler architecture and reliance on single-scale detection, which limits its ability to effectively detect the small, varied objects present in these datasets. While it remains a viable option for real-time applications, its lack of precision makes it unsuitable for tasks where high detection accuracy is critical.

In contrast, Faster R-CNN demonstrated superior accuracy, particularly in scenarios involving small object detection. On DOTA, it achieved a mAP of 31.2%, while on TinyPerson it performed similarly well, with a mAP of 30.4%. This improvement is largely attributed to its region proposal network (RPN), which enhances its ability to identify small objects even in densely cluttered environments, such as aerial imagery or crowded pedestrian scenes. However, the increased accuracy of Faster R-CNN comes at the expense of speed. It operates at only 6 FPS on DOTA and 5 FPS on TinyPerson, making it unsuitable for real-time detection tasks despite being a top performer in precision-demanding scenarios.

The SSD (Single Shot Multibox Detector) model, widely recognized for its balance between speed and accuracy, struggles with small object detection, especially on datasets like DOTA and TinyPerson. On DOTA, SSD achieved a mAP of 21.4%, and on TinyPerson, it recorded a mAP of 19.8%, with

DOI: 10.54254/2753-8818/79/2025.20099

corresponding speeds of 28 FPS and 27 FPS, respectively. Although SSD performs better on datasets with larger objects, such as PASCAL VOC, where it achieved a mAP of 74.3%, its efficiency on small object detection tasks remains limited. Nevertheless, its relatively fast inference time and decent accuracy make SSD a suitable choice for real-time applications, particularly in environments where computational resources are constrained and larger objects dominate the scene.

Feature Pyramid Network (FPN), which specializes in multi-scale and small object detection tasks, performed well on both DOTA and TinyPerson. FPN achieved a mAP of 34.5% on DOTA, with a processing speed of 12 FPS, while on TinyPerson, it reached a mAP of 33.2%, operating at 11 FPS. FPN's architecture, which uses a top-down feature pyramid combined with lateral connections, allows it to detect objects across multiple scales effectively. This makes it highly suitable for scenarios like aerial imagery and pedestrian surveillance, where both large and small objects appear in the same image. FPN strikes a balance between accuracy and speed, making it a versatile model for a variety of detection tasks.

NAS-FPN (Neural Architecture Search Feature Pyramid Network) achieved the highest accuracy on both DOTA and TinyPerson, with a mAP of 37.8% on DOTA and 35.6% on TinyPerson. However, this high accuracy comes at a significant cost to speed, as NAS-FPN operates at just 5 FPS on DOTA and 4 FPS on TinyPerson. The model's high performance can be attributed to the Neural Architecture Search (NAS) process, which automatically optimizes the feature pyramid network for small object detection. This allows NAS-FPN to excel in challenging environments with crowded or cluttered scenes, making it the ideal choice for tasks that prioritize precision, such as aerial image analysis or detailed surveillance. Despite its slower speed, NAS-FPN's accuracy makes it invaluable for applications requiring meticulous detection of small objects.

RetinaNet, another high-performing model, achieved a mAP of 32.1% on DOTA and 31.5% on TinyPerson, with respective speeds of 18 FPS and 15 FPS. RetinaNet's use of the focal loss function helps mitigate the imbalance between easy and hard-to-classify examples, which enhances its ability to detect small and less distinct objects. Although RetinaNet offers a reasonable trade-off between speed and accuracy, it is still outperformed in terms of speed by models like YOLOv2, limiting its suitability for real-time applications that demand faster inference times.

Finally, EfficientDet, which uses a compound scaling approach to optimize network depth and input resolution, achieved a mAP of 29.8% on DOTA and 28.6% on TinyPerson, with respective speeds of 10 FPS and 9 FPS. While EfficientDet is faster than NAS-FPN and Faster R-CNN, it lags behind in terms of accuracy, particularly in challenging small-object scenarios. However, its efficiency makes it a viable option for use in resource-limited environments where both speed and moderate accuracy are required.

Table 1: Performance Metrics Table

Model	DOTA mAP	TinyPerson mAP	PASCAL VOC mAP	COCO mAP	DOTA FPS	TinyPerson FPS	PASCAL VOC FPS	COCO FPS
YOLOv2	14.3	10.5	57.9	48.1	62	65	67	68
Faster R-CNN	31.2	30.4	78.5	62.0	6	5	6	5
SSD	21.4	19.8	74.3	51.4	28	27	30	29
FPN	34.5	33.2	81.6	63.2	12	11	13	12
NAS-FPN	37.8	35.6	82.4	64.7	5	4	6	5
RetinaNet	32.1	31.5	80.1	61.5	18	15	18	17
EfficientDet	29.8	28.6	77.0	59.5	10	9	11	10

DOI: 10.54254/2753-8818/79/2025.20099

In summary, the comparison highlights the trade-offs between speed and accuracy that define each model's performance on different detection tasks. YOLOv2 remains the best choice for real-time applications, though its low accuracy makes it less ideal for complex tasks. In contrast, Faster R-CNN and NAS-FPN offer the highest precision, particularly for small object detection, but are hindered by slow processing speeds. SSD and FPN offer balanced options for tasks requiring moderate accuracy and speed, making them adaptable across various detection scenarios. Ultimately, the choice of model depends on the specific requirements of the detection task, whether that be speed, accuracy, or the ability to handle small-scale objects in cluttered environments.

4. Challenges and Future Directions

Object detection, especially for small objects, faces significant challenges due to the inherent biases in large datasets and limitations in current model architectures. The imbalance in small object representation, the extraction of meaningful semantic information, and the interpretability of models remain key areas requiring advancement. Addressing these challenges is crucial to furthering the effectiveness of object detection models, particularly in scenarios where precision and versatility are paramount.

4.1. Lack of Small Object Representation in Datasets

Large-scale object detection datasets like ImageNet, Pascal VOC, and MS COCO have been instrumental in advancing detection models. However, these datasets often exhibit a bias towards medium and large objects, with a notable underrepresentation of small objects. This imbalance negatively impacts the training of detection models, leading to suboptimal performance on small object detection. For instance, in the MS COCO dataset, only about half of the training images contain small objects, while 70.07% and 82.28% include medium and large objects, respectively. Furthermore, a mere 1.23% of labeled pixels correspond to small objects, while medium-sized objects occupy more than eight times the pixel area of small objects and account for 10.18% of the annotated pixels. The majority of pixels (82.28%) are allocated to large objects. This underrepresentation of small objects directly contributes to the lower average precision (AP) scores for small object detection across various models.

One intuitive solution to mitigate this problem is data augmentation, which increases the frequency and variety of small objects in the dataset. Augmentation techniques can address the small object representation issue in two ways: increasing the occurrence of small objects within images and improving their spatial distribution. This is especially crucial given that existing small object datasets are often task-specific and do not generalize well to broader detection tasks. Without sufficient data, models struggle to learn effective features for small object detection, ultimately compromising their accuracy.

To address this challenge, several approaches have been proposed. Kisantal et al. [9] developed a data augmentation technique aimed at overcoming the scarcity of small object pixels in the MS COCO dataset. Their method employs an oversampling strategy, wherein images containing small objects are duplicated to increase the number of training samples. In addition, they introduced a copy-paste technique, which involves copying small objects from one image and pasting them into various locations within the same image. This strategy enriches the variety of small object positions and boosts their frequency, leading to improvements in detection accuracy. However, the method is limited to images that already contain small objects, restricting its effectiveness for more diverse datasets.

Chen et al. [10] refined this approach by addressing some limitations of simple copy-paste methods. They observed that such techniques could lead to mismatches between the copied objects and their

DOI: 10.54254/2753-8818/79/2025.20099

background or incorrect scaling. To counter these issues, they utilized a pre-trained semantic segmentation network to accurately position replicated objects in semantically appropriate areas of the image. Additionally, they introduced a linear scaling function to ensure that objects are appropriately resized, improving the overall detection performance and reducing visual inconsistencies. Moreover, Chen et al. proposed a dynamic scale training strategy, wherein the training process dynamically adjusts based on the feedback of the loss function. They found that small object loss contributed less than one-tenth of the total loss for more than half of the training iterations. By resizing input images to emphasize small objects during subsequent iterations, they were able to correct the bias towards medium and large objects, resulting in a more balanced and optimized model training process.

4.2. Semantic Information

Another key challenge in small object detection lies in maximizing the extraction of semantic information from deep learning architectures. Rich semantic content allows a model to learn a wider variety of discriminative features, leading to better detection performance. However, current network architectures often struggle to extract enough semantic information, especially for small objects, limiting their effectiveness.

There are two primary approaches to overcoming this issue. The first approach focuses on high-resolution representation, which aims to retain spatial details throughout the deep layers of a network. In conventional neural networks, feature map dimensions are reduced as the network deepens, leading to a loss of spatial resolution that is particularly detrimental to small object detection. To address this, techniques such as hourglass network structures, deconvolution (transposed convolution), dilated convolutions, and multi-scale parallel branches have been employed. These methods help retain high-resolution feature maps, enabling the network to capture fine spatial details, which improves detection precision, particularly for smaller objects.

The second approach emphasizes semantic understanding, particularly at the pixel level. Models that are capable of learning and processing fine-grained semantic information—especially through pixel-level object instance segmentation—can extract detailed attributes of objects. This leads to a more accurate understanding of the image, especially in distinguishing between overlapping or closely grouped objects. Such models are more adept at handling complex object detection tasks, especially in environments where objects are densely packed or cluttered, as is often the case in aerial imagery or surveillance settings.

4.3. Interpretability

A fundamental challenge in deep learning models for object detection, particularly in small object scenarios, is their lack of interpretability. While neural networks, particularly deep architectures, have achieved remarkable accuracy in a variety of detection tasks, the reasoning behind their predictions often remains a "black box." Understanding how specific decisions are made within these networks is difficult due to their complexity and the vast number of parameters involved. This opacity can be problematic, especially in critical applications where understanding the model’s decision-making process is essential.

A promising direction to improve model interpretability lies in the development of capsule networks. Traditional convolutional neural networks (CNNs) often rely on pooling layers to reduce the size of feature maps, but this process can discard important spatial information, which is particularly useful for detecting small objects. Capsule networks, in contrast, are designed to capture spatial hierarchies and entity-specific features, such as pose, orientation, and scale. By preserving these relationships, capsule networks maintain more detailed connections between image features,

DOI: 10.54254/2753-8818/79/2025.20099

allowing for better generalization across different viewing angles and object deformations. Importantly, capsule networks make these relationships explicit, thereby enhancing the interpretability of model predictions. This allows for clearer insights into how the model identifies and distinguishes objects, pote.

5. Conclusion

In this paper, we have provided a comprehensive evaluation of state-of-the-art object detection models, particularly focusing on their performance in detecting small objects across challenging datasets like TinyPerson and DOTA. Through our analysis of models such as YOLOv2, Faster R-CNN, SSD, FPN, and NAS-FPN, we highlighted the trade-offs between speed and accuracy, revealing how these models fare when faced with complex tasks like small object detection. YOLOv2, for instance, performs well in real-time applications but struggles with small objects, while models like Faster R-CNN and NAS-FPN offer higher precision but are limited by slower speeds. Balancing these factors remains a significant challenge in advancing object detection technologies.

A core issue identified is the underrepresentation of small objects in large datasets such as MS COCO, Pascal VOC, and ImageNet, which biases models toward detecting medium and large objects. Data augmentation techniques, including oversampling and copy-paste strategies, have shown promise in increasing small object representation, but further improvements are needed. Methods like dynamic scale training and enhanced semantic segmentation, as explored by Chen et al., have begun to address these gaps, offering more robust solutions for improving small object detection accuracy.

We also discussed the importance of maximizing semantic information extraction. Techniques such as high-resolution feature representation and pixel-level semantic segmentation have proven effective in retaining spatial details and improving small object detection. However, model interpretability remains a significant challenge, particularly in critical applications like surveillance and aerial analysis. Capsule networks offer a potential solution by capturing spatial hierarchies and enhancing transparency in model decision-making, which is crucial for building more trustworthy systems.

Looking ahead, future research should focus on creating more diverse and representative datasets that better capture the spectrum of object sizes and contexts. Continued improvements in data augmentation and semantic feature extraction are essential for building models that generalize across varied detection tasks. Moreover, enhancing model interpretability, particularly through capsule networks and explainable AI techniques, will be crucial for developing transparent and reliable object detection systems for real-world applications.

References

[1] *Ghiasi, G., Lin, T. Y., & Le, Q. V. (2019). Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7036-7045).*

[2] *Zoph, B. (2016). Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578.*

[3] *Ross, T. Y., & Dollár, G. K. H. P. (2017, July). Focal loss for dense object detection. In proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2980-2988).*

[4] *Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., & Sun, J. (2018). Detnet: A backbone network for object detection. arXiv preprint arXiv:1804.06215.*

[5] *Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).*

[6] *Liu, S., & Huang, D. (2018). Receptive field block net for accurate and fast object detection. In Proceedings of the European conference on computer vision (ECCV) (pp. 385-400).*

[7] *Li, Y., Chen, Y., Wang, N., & Zhang, Z. (2019). Scale-aware trident networks for object detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6054-6063).*

DOI: 10.54254/2753-8818/79/2025.20099

[8] Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., & Yan, S. (2017). Perceptual generative adversarial networks for small object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1222-1230).

[9] Kisantal, M. (2019). Augmentation for Small Object Detection. arXiv preprint arXiv:1902.07296.

[10] Chen, Y., Zhang, P., Li, Z., Li, Y., Zhang, X., Qi, L., ... & Jia, J. (2020). Dynamic scale training for object detection. arXiv preprint arXiv:2004.12432.