Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD29184

A Study of YOLO, Transformer and Diffusion Model for Small Object Detection

Tiancheng Hu

UCL Department of Mathematics, University College London, London, The United Kingdom
zcahth3@ucl.ac.uk

Abstract. In recent years, algorithms in the field of computer vision have been continuously innovated and promoted, and the progress of small object detection has become a key task in the development of this field. However, compared with the detection of medium and large targets, factors such as background interference can easily interfere with the detection of small targets with smaller pixel coverage areas, making progress more difficult. In recent years, researchers have proposed various methods to address these challenges, and the three most representative frameworks are algorithms developed using YOLO, Transformer, and Diffusion models. This article provides a detailed overview and comparison of three models. The YOLO based method is superior in improving real-time detection through multi-scale feature enhancement, structural optimization, and adjusting the loss function. Based on the Transformer, the accuracy and precision of identifying small targets are improved by adjusting the mechanism, using a hybrid structure and multimodal feature fusion. And researchers will adjust the diffusion process, involving the construction of diffusion bounding boxes and diffusion engines, to enable the application of diffusion model algorithms. Finally, this article summarizes the advantages and limitations of these methods and discusses potential future research directions. The significance of this study lies in providing a unified overview of the three main research paradigms, helping researchers understand current progress, identify existing challenges, and explore new possibilities for advancing small object detection.

Keywords: Small Object Detection, YOLO, Transformer, Diffusion Model.

1. Introduction

In recent years, object detection has become a major research and application in the field of computer vision, leading to the development of numerous applications such as autonomous driving, intelligent monitoring, unmanned aerial vehicles (UAVs), and image analysis. Compared to medium to large objects, small objects cover fewer pixels and have weaker features that can be extracted, making algorithm detection more difficult. Therefore, the accuracy of small object detection still lags far behind that of large objects, which poses a research bottleneck for the visual system in the real world.

You Only Look Once (YOLO) is one of the most widely used detection frameworks due to its unique end-to-end system, resulting in extremely high detection performance and efficiency. As

DOI: 10.54254/2755-2721/2025.LD29184

Shao's team pointed out, YOLO is one of the most versatile and practical algorithms in the field of object detection, and has great potential [1]. However, the standard YOLO model has low positioning accuracy in small object detection and weak detection ability in complex backgrounds. However, the researchers have made certain improvements to this and applied it to scenarios such as drone imaging and autonomous driving.

The Transformer based method has a powerful global attention mechanism and is highly popular for its ability to scan up and down information. And the detection performance basically surpasses algorithms based on a convolutional neural network architecture. However, they also face challenges such as high computational overhead, dependence on large-scale datasets, and difficulty in real-time deployment.

The emerging diffusion based methods view object detection as a generative process, providing a new perspective. The diffusion model gradually adds noise to the data and learns the reverse denoising process, which can be applied to bounding box generation or disguised object detection. However, the difficulty of training and the cost and accuracy of algorithms are not balanced,

This article provides a comprehensive overview of small object detection methods based on these three architectures. This paper analyze the algorithm models of each representative, summarizes the current methods for improving applications and their limitations. The significance of this study lies in systematically understanding how three different frameworks adapt to handling small object detection, highlighting the advantages and disadvantages of different strategies, and providing insights for future research to design more accurate, robust, and efficient small object detection algorithm frameworks.

2. Application of YOLO

2.1. Study of YOLO

YOLO is an object detection algorithm proposed in recent years. Researchers have taken a different approach by cropping images and detecting objects separately. Due to its unique end-to-end training and detection method, high accuracy and detection efficiency, and ease of training, YOLO has been widely experimented and developed in computer vision. Scholars have developed small object detection algorithms based on the YOLO framework and conducted experiments and improvements in different environments.

The first method is to design a structure for extracting features from multimodal objects, in other words, to construct a feature fusion module to improve the detection capability of small targets. For example, Liu's team mainly focuses on small target detection for drones and has developed UAV-YOLO [2]. The core of the improvement lies in constructing a pyramid network structure to enhance the ability to perceive pixels and improve the ability to extract small target features, thereby solving the problem of small target size and high density in drone captured images.

The second type is the optimization of network structure. It is to add or adjust a detection module suitable for capturing small target features for the system, which can effectively alleviate the problem of feature loss caused by downsampling while ensuring real-time detection. Benjumea et al. developed YOLO-Z for small object recognition in autonomous driving scenarios [3]. The modified and adjusted structure of each part greatly enhances the ability to recognize small vehicles and pedestrians. The experimental results on autonomous driving related datasets show that the accuracy of YOLO-Z is significantly better than the standard model, while maintaining a high detection speed. This indicates that structural optimization improves and enhances YOLO small object detection in specific scenarios.

DOI: 10.54254/2755-2721/2025.LD29184

The third type is the optimization of the loss function. The loss function of bounding box regression can ensure that the algorithm can handle small targets more accurately and stably, thereby avoiding the problem of fuzzy detection of boundary objects. For example, Hu et al. proposed EL-YOLO, which introduces the α - CIOU loss function to replace the conventional CIOU [4]. This method reduces the weight of low-quality object boundaries by introducing the power parameter α into the IoU loss. The experimental results on the aerial image datasets DIOR and VisDrone show that α - CIOU can greatly improve regression and detection accuracy without increasing the computational complexity and inference time of the model.

2.2. Experimental results and limitations

In terms of performance, UAV-YOLO significantly improves the mAP on drone aerial datasets [2]. Experimental results on drone datasets show that UAV-YOLO performs significantly better on AP than the standard YOLO model, especially in terms of mAP values, achieving the best small target detection accuracy currently available. However, when detecting small objects, the degree of accuracy improvement should be smaller than that of large objects, so it will greatly lack performance in detecting other objects.

The accuracy and recall of small object detection in YOLO-Z in autonomous driving scenarios are significantly better than YOLOv5 [3], but the feature maps in the module may cause noise interference to the model during the experimental process, leading to a decrease in accuracy.

EL-YOLO has achieved extremely high performance results in various aspects of similar models. These results indicate that improvements based on loss functions, continuously increasing the weight and attention to high-quality borders, can indeed improve the performance of small object detection in different scenarios. However, by focusing too much on high LOU samples and reducing attention to low LOU samples, the related recall rate will decrease, and the comprehensiveness of the loss function has certain limitations.

In summary, in recent years, YOLO based small object detection methods have made significant progress in multi-scale feature fusion and network structure optimization. These improvements have shown good performance in application scenarios such as drone aerial photography, autonomous driving, and aviation monitoring. However, these methods still face difficulties in further improving accuracy and expanding the ability to extract object types, especially in improving the accuracy of simultaneously performing large object detection tasks and optimizing the extraction amplitude of small objects. The detection accuracy needs to be further improved.

3. Transformers' application

3.1. Study of transformer

In recent years, Transformers have received widespread attention in the field of computer vision. Through its global attention mechanism, Transformer can better capture contextual information of small targets, thereby improving detection performance and demonstrating stronger potential in small target detection. Therefore, transformer based small object detection has become a research hotspot in both academia and industry.

The first method is to improve small target feature extraction through hierarchical attention and pyramid structure, replacing the backbone network and enhancing multi-scale features [5], so that the network can balance global semantics and local details. Xu's team proposed a DETR small object detection algorithm based on Swin Transformer [5]. This method improves the detection

DOI: 10.54254/2755-2721/2025.LD29184

performance of small targets by designing a novel layer backbone network that combines the advantages of DETR and Swin, and introducing a three-layer feature pyramid to enhance multi-scale information, thereby solving the problem of poor performance of traditional DETR in small target detection.

Next is the design of the hybrid powertrain structure. For example, Chen et al. proposed a hybrid transformer detector HTDet for complex underwater environments [6]. Due to low contrast and uneven lighting in underwater images, the main problem is that the target signal is easily covered and disappears. HTDet adopts a hybrid lightweight network based on transformers, combining the local inductance bias of CNN with the global modeling ability of transformers, and introducing fine-grained FPN to eliminate the problem of losing relevant signals. It demonstrates how to leverage CNN's lightweight and inductive biases, as well as Transformer's contextual modeling capabilities, to maintain high accuracy in complex noisy environments.

The third category is mechanism adjustment and structural optimization. Multiple studies have shown that using standard self attention directly on high-resolution input may result in high computational costs [7]. Therefore, deformable DETR and other methods have proposed sparse attention mechanisms to reduce complexity and improve convergence speed. Rekavandi et al. reviewed over 60 studies on transformers, including most of the testing items [7]. This review points out that the key advantage of Transformer over CNN is that its self attention mechanism can explicitly simulate contextual information, which is particularly important for small object detection.

3.2. Experimental results and limitations

These Transformer based methods perform well in different scenarios. The experimental results show that due to the precise detection advantage of the Swin Transformer structure, its detection accuracy on the VOC and Tiny Person datasets is 11AP higher than normal, and its detection speed is 2.5 fph higher than DETR [5]. This indicates that the DETR algorithm combined with Swin Transformer is more effective and suitable for large-scale complex object detection. However, the Transformer model has a large number of parameters and high computational overhead [5].

HTDet maintains a high real-time detection rate in underwater images, with mAP measured on the underwater dataset increasing by 6.3 percentage points compared to the normal baseline, and a real-time detection speed of 22.6 fps, demonstrating its advantage in high noise scenes. However, it has a strong dependence on data size and is prone to overfitting if there is insufficient training data [6].

This review compares the performance of different methods on datasets such as COCO, remote sensing, underwater, and medical images. The results indicate that this method is generally superior to traditional CNN in detecting small objects in transformer adjustment mechanisms and specific structures. However, its detection accuracy is still limited in extremely small targets or densely occluded scenes [7].

4. Application of diffusion

4.1. Study of diffusion

The core idea of the diffusion model is to gradually add noise to the data and learn the inverse process to restore a clear image from random noise. It is widely used in image recognition tasks for object detection. Since small objects often have weak features and are buried in the background, the diffusion model allows users to not only generate high-quality images but also leverage the inverse

DOI: 10.54254/2755-2721/2025.LD29184

process of the diffusion model for image recognition and feature extraction. This chapter reviews typical small object detection methods based on the diffusion model, analyzes their applications and improvements in different scenarios, and summarizes their strengths and limitations.

The first method is diffusion driven bounding box generation. Chen et al. proposed DiffuseDet, which introduces diffusion models into object detection tasks [8]. This method considers object detection as a "noise to object" generation process: object boxes diffuse from ground truth boxes to random distributions, and the model learns to reverse this noise diffusion process. During the inference process, the randomly generated boxes are iteratively corrected step by step to obtain the final detection result. Unlike traditional methods that rely on anchor boxes or query points, DiffusionDet completely eliminates the need for manually designing candidate boxes and provides flexibility in a dynamic number of boxes and iterative optimization.

The second method is a diffusion process and feature fusion. Chen et al. proposed a diffusion based object segmentation framework diffCOD [9], which applies diffusion models to detect disguised small objects. Due to the high similarity between disguised objects and the background, traditional methods often find it difficult to distinguish them. DiffCOD describes this task as the diffusion process from the noise mask to the object mask. During the training process, noise gradually adds to the real mask, allowing the model to learn denoising and mask restoration. During the inference process, the object region is gradually generated from a random noise mask.

Zhang et al. proposed a diffusion engine to process data generation from the perspective of data generation, and proposed a new method of using diffusion models to extend detection datasets [10]. This framework consists of pre trained diffusion models and detection adapters, which facilitate plug and play generation of scalable, diverse, and universal detection data, thereby achieving better boundary prediction. The extended versions (COCO-DE and VOC-DE) are also built based on the COCO and VOC datasets.

4.2. Experimental results and limitations

The experimental results show that the diffusion model method outperforms traditional CNN and Transformer detectors in various scenarios. DiffusionDet outperforms DETR by several percentage points on COCO and achieves a 5% AP improvement in zero sample transmission on the CrowdHuman dataset. It also outperforms DETR and sparse R-CNN in small object detection and cross domain transmission tasks, further improving accuracy by increasing the number of sampling steps. This demonstrates its unique advantages in small object detection in complex scenes [8]. However, DiffuseDet may be affected by relatively slow sampling speeds, resulting in a decrease in speed performance.

DiffCOD outperforms 11 existing methods comprehensively on four camouflage detection datasets [9]. Especially in terms of texture detail segmentation. This work demonstrates that diffusion models are not only effective in general detection, but can also address the challenge of detecting extremely small objects in complex backgrounds. However, the inference process of diffusion models involves multiple iterations, resulting in significant computational overhead and difficulty in meeting real-time detection requirements. Moreover, the larger the model structure, the more difficult it is to train.

The diffusion engine increased the mAP of COCO-DE and VOC-DE by 3.1% and 7.6%, respectively [10]. These results demonstrate the strong generalization ability and improvement potential of diffusion models in small object detection. Moreover, experiments have shown significant improvements in various scenarios, including those involving various detection algorithms, self supervised pre training, and data scarcity. Its performance in small object detection

DOI: 10.54254/2755-2721/2025.LD29184

is significantly better than several mainstream detectors. This study suggests that diffusion models can not only be used directly as detectors, but also as powerful data engines, providing diverse training samples for small object detection. However, the engine experiment lacks human experience data, and data generation requires text guidance.

5. Future research

Although the above algorithms have achieved improvements in accuracy and speed, they have great advantages and uniqueness compared to other algorithms. But limitations still exist. YOLO's algorithm still has performance bottlenecks in detecting extremely small targets, and its application scope for object detection is not wide. The Transformer model requires a large amount of computation and large-scale datasets, resulting in high costs. The high complexity of training diffusion models is hindered by excessive reliance on simulated data and a lack of human data. Future research should focus on integrating the advantages of these paradigms. Firstly, a lightweight and deployable architecture is crucial for real-time applications, as it can adapt to the changing environment by adjusting the structure. In addition, all three models involve multimodal feature fusion, which is also one of the main paths for future development. People believe that combining efficiency and contextual modeling extraction capabilities can better build a more powerful and efficient small object detection system.

6. Conclusion

This article mainly reviews three frameworks and algorithms developed in small object detection, namely YOLO, the transformer, and the diffusion model. Each framework has unique advantages and solves the challenge of detecting small targets in different fields. The YOLO model mainly enhances feature extraction to improve detection accuracy and ensure real-time detection. This can be achieved by constructing a feature pyramid structure, modifying the network structure, and improving the loss function. However, it still has limitations in terms of accuracy and application scope. Researchers enhance the detection capability of contextual information in Transformer models by extracting multimodal features, using hybrid structures, and adjusting and optimizing mechanisms. However, it has high computational costs and strong data dependencies. The diffusion based model can achieve the most accurate and repetitive detection and improve accuracy by building a diffusion engine, feature fusion, and establishing diffusion bounding boxes. However, the training method is difficult and the cost is high due to the large size of the model. This article provides a detailed overview of three basic algorithm development methods. Future research can continue to improve based on the three algorithms and deploy them in real-time according to their advantages for the application. In addition, combining the standards and shortcomings in the article, continuously improving accuracy, multi-modal and real-time return rates, and reducing training costs and data dependencies will be the future path for small object detection algorithms. This is crucial for the development of this field,

References

[1] Shao, Y., Zhang, D., Chu, H., Zhang, X., & Rao, Y. (2022). A review of YOLO object detection based on deep learning. Journal of Electronics and Information Technology, 44(10), 3697-3708.

[2] Liu, M., Wang, X., Zhou, A., Fu, X., Ma, Y., & Piao, C. (2020). Uav-yolo: Small object detection on unmanned aerial vehicle perspective. Sensors, 20(8), 2238.

DOI: 10.54254/2755-2721/2025.LD29184

[3] Benjumea, A., Teeti, I., Cuzzolin, F., & Bradley, A. (2021). YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles. arXiv preprint arXiv: 2112.11798.
[4] Ji, S. J., Ling, Q. H., & Han, F. (2023). An improved algorithm for small object detection based on YOLO v4 and multi-scale contextual information. Computers and Electrical Engineering, 105, 108490.
[5] Fengchang, X., Alfred, R., Pailus, R. H., Ge, L., Shifeng, D., Chew, J. V. L., ... & Xinliang, W. (2024). DETR novel small target detection algorithm based on Swin transformer. IEEE Access, 12, 115838-115852.
[6] Rekavandi, A. M., Rashidi, S., Boussaid, F., Hoefs, S., & Akbas, E. (2023). Transformers in small object detection: A benchmark and survey of state-of-the-art. arXiv preprint arXiv: 2309.04902.
[7] Chen, G., Mao, Z., Wang, K., & Shen, J. (2023). HTDet: A hybrid transformer-based approach for underwater small object detection. Remote Sensing, 15(4), 1076.
[8] Chen, S., Sun, P., Song, Y., & Luo, P. (2023). Diffusiondet: Diffusion model for object detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 19830-19843).
[9] Chen, Z., Gao, R., Xiang, T. Z., & Lin, F. (2023). Diffusion model for camouflaged object detection. arXiv preprint arXiv: 2308.00303.
[10] Zhang, M., Wu, J., Ren, Y., Yang, J., Li, M., & Ma, A. J. (2025). Diffusionengine: Diffusion model is scalable data engine for object detection. Pattern Recognition, 112141.