Enhancing YOLOv7 with GLF-Trans for Precision in Small Object Detection

Naohito Yoshikawa* and Masaaki Ikehara†
* Electrical and Information Engineering (of Keio University), Kanagawa, Japan
E-mail: yoshikawa@tkhm.elec.keio.ac.jp
† Electrical and Information Engineering (of Keio University), Kanagawa, Japan
E-mail: ikehara@tkhm.elec.keio.ac.jp

Abstract—Object detection techniques are essential for identifying objects and accurately determining their locations in images and videos. This paper proposes an approach to enhance the performance of YOLOv7, the YOLO (You Only Look Once) series, by improving detection accuracy, particularly for small objects. The accurate detection of small objects is crucial in various real-world applications such as autonomous driving, surveillance, and medical imaging, where the identification of small, often critical objects can have significant implications. The proposed method balances the current one-step and two-step algorithms used in object detection, with the goal of improving the accuracy of small object detection. We apply several improvements to the YOLOv7 framework, including the introduction of an attention module, the Global-Local Fusion Transformer (GLF-Trans), and the addition of a tiny-head. These enhancements are designed to further enhance the accuracy of small object detection. The experimental results show that the proposed method achieves higher accuracy than the previous model, with a specific improvement of 3.8% in $mAP_{0.5}$ on the VisDrone2019 dataset.

I. Introduction

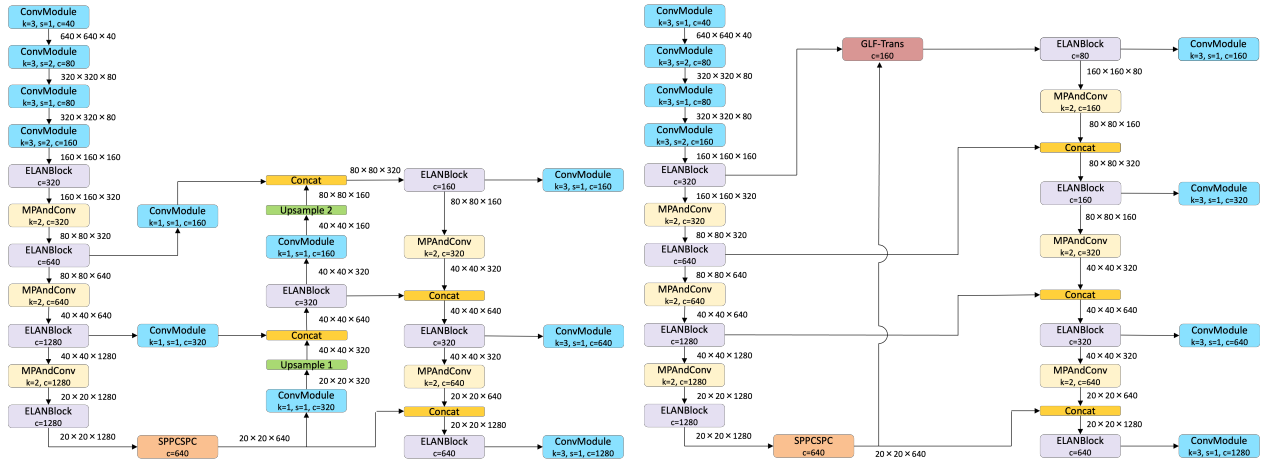
Object detection is a technique used to locate and identify objects in images or videos. This technology is widely used in various applications. In recent years, advances in object detection models based on deep learning have enabled fast and accurate detection of objects. The YOLO (You Only Look Once) series [1][2][3][4][5][6][7] has gained attention due to its balance of speed and accuracy. This study examines YOLOv7 by Wang et al. [7], which has demonstrated superior performance compared to other object detection techniques. However, there is still potential for improvement in detecting small objects.

Object detection algorithms can be broadly divided into two categories: single-stage and two-stage algorithms. The former prioritizes execution speed over accuracy, while the latter achieves higher accuracy at the cost of slower processing times. YOLO is a typical example of a one-stage algorithm, which is suitable for real-time processing due to its high speed. However, it has been noted that its accuracy is inferior to the two-stage approach, particularly in detecting small objects. Small object detection is a challenging task due to the lack of detail and difficulty in distinguishing objects from the background. However, improving the accuracy of small object detection while maintaining real-time performance can have numerous applications, including navigation in automated vehicles, surveillance systems, and robotics. This paper develops new methods that can accurately detect small objects by surpassing the limitations of current one-step algorithms.

In this study, we improved the YOLOv7 [7] object detection framework based on the importance of learning the right balance between detailed information and overall context to enhance the accuracy of small object detection. The existing YOLO model does not learn enough contextual information, lacking the detailed information necessary for small object detection. To address this issue, we introduced two major improvements. First, we added a fourth head, the Tiny Head, to convey sufficient information for detecting small objects. Next, we replaced the neck part of the model with an attention module, the Global-Local Fusion Transformer (GLF-Trans), to reduce the increased parameter count and computational load caused by the Tiny Head while maintaining accuracy and real-time performance. These improvements are expected to enable learning the balance between detailed and contextual information, thereby improving the accuracy of small object detection.

The proposed method reduces the number of parameters from 70.9M to 67.3M without negatively impacting computational cost and efficiency. Experiments on the VisDrone2019 dataset [8] demonstrate a 3.8% accuracy improvement at $mAP_{0.5}$ for small object detection compared to the previous model. The main contributions of this paper are as follows:

- A new Tiny-head specialized for the detection of small objects is added to the existing framework of YOLOv7, enabling more precise identification of small objects. Detection of small objects relies heavily on high-resolution feature maps due to their diminutive size. The implementation of Tiny-head allows for the utilization of finer feature maps to detect small objects, significantly enhancing the accuracy of identification for small objects.
- To maintain accuracy while reducing the number of parameters and computational load, the Global-Local Fusion Transformer (GLF-Trans) is introduced. The detection of small objects necessitates the differentiation of crucial details from background noise



(a) YOLOv7x (b) Proposed model

Fig. 1: Structure of YOLOv7x and the proposed model

and the concurrent understanding of the object and its surrounding context. GLF-Trans enhances this capability by fusing global contextual information with local detail through an attention mechanism, effectively replacing the current neck mechanism.

- By adopting Cluster-NMS in place of the traditional NMS algorithm, we improve the clustering accuracy of detected objects, thereby increasing detection precision, particularly in dense scenes or situations where occlusion occurs.

II. Related Works

A. Advancements in YOLO Architectures for Small Object Detection

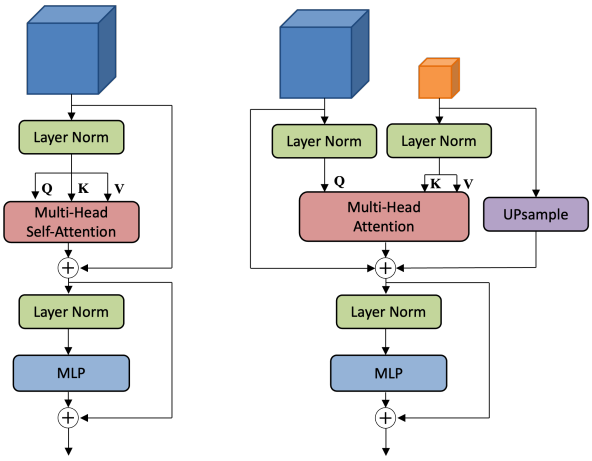
The field of small object detection has significantly evolved alongside the advancements in object detection technologies. Among these developments, the QueryDet approach by Yang et al. [9] represents a specialized methodology for enhancing the detection accuracy of small objects through the utilization of a cascading structure. This approach first identifies the approximate location of objects, then refines the details at a finer scale, thereby improving the detection precision for small objects. The YOLO series is renowned for its simplicity and rapid detection speed, and with each successive version, improvements have been made to enhance small object detection. While early versions struggled with detecting small objects, the introduction of multi-scale detection in YOLOv3 [3], utilizing feature maps of different resolutions, marked a significant improvement in detection accuracy for small objects. YOLOv4 [4] further expanded upon this by adding data augmentation techniques such as CutMix and Mosaic. Additionally, the inclusion of technologies like CSPNet and Spatial Pyramid Pooling further strengthened the detection accuracy and generalization capability of models for small object detection. In YOLOv5 [5], the model architecture was optimized, and the use of anchor boxes was refined, leading to further improvements in detecting performance for small objects. YOLOv7 [7] referenced in this paper, the introduction of the ELAN structure is highlighted, which serves to enhance the network's learning capacity without interrupting the original gradient flow. These improvements have elevated the accuracy of the YOLO series in detecting various objects, including small ones. However, the performance in small object detection has not yet reached a satisfactory level.

B. Attention Modules for Object Detection

The attention mechanism allows models to focus on the important parts of the input, clarifying the distinction between objects and background in object detection. Zhu et al.'s TPH-YOLOv5 [10], developed for drone-captured scenario applications, introduces a Transformer Prediction Head (TPH) leveraging a self-attention mechanism into YOLOv5 to accommodate object scale variations due to different altitudes and motion blur, thereby enhancing the accuracy in dense object detection. This model employs the Convolutional Block Attention Module to concentrate on significant objects while suppressing confusing information in the image. The utilization of attention module to capture detailed information extends beyond object detection. In Han et al.'s BRNet [11], a Global-Detail Fusion Module (GDFM) that combines local and global features was introduced to improve self-supervised monocular depth estimation. This module generates feature representations enriched with more comprehensive information through the fusion of local and global information, enhancing the accuracy of depth estimation.

C. NMS

Non-Maximum Suppression (NMS) is an essential process in object detection, aimed at reducing redundant detection boxes to refine the final detection outcomes. The method proposed by Zheng et al., Cluster-NMS [12],



(a) Transformer Block (b) GLF-Trans Block

Fig. 2: Structure of Transformer Block and GLF-Trans Block

enhances this traditional NMS process by efficiently clustering detection boxes, thereby simultaneously improving the efficiency and accuracy during the inference phase. This technique not only considers the overlapping areas but also incorporates geometric elements such as the distance between the centers of the boxes and their aspect ratios, improving the balance between detection precision (Average Precision, AP) and recall (Average Recall, AR). Notably, it offers performance enhancements in densely populated scenes and situations where occlusion occurs, demonstrating effectiveness not only in detecting large objects but also in small object detection scenarios.

III. Proposed Method

The proposed network is illustrated in Fig. 1(b). While our model builds upon the foundation of YOLOv7x by Wang et al. [7], citing Fig. 1(a) as the basis for improvements, it fundamentally retains the same structure as YOLOv7. YOLOv7 is composed of a Backbone, Neck, and Head, whereas our network consists of a Backbone, GLF-Trans, and Head. The Backbone, similar to traditional YOLOv7, is comprised of ConvModules and ELANBlocks [13], extracting feature maps of various scales from the input image. The GLF-Trans serves as an attention module, replacing the conventional Path Aggregation Network structure in the Neck segment. Feature maps downsampled by factors of 4 and 32 are fed into the attention mechanism. The fusion of feature maps of different scales through attention enables the aggregation of global features required by each pixel and adds these features to the detailed feature map, generating feature representations containing richer information. This allows for the acquisition of detailed information necessary for improving the detection accuracy of small objects, as well as global contextual information, and is also expected to contribute to the suppression of background noise. Traditionally, the Head component utilized three Heads for

TABLE I: Model Performance Comparison

Model	$mAP_{0.5\_val}$	$mAP_{0.5\_test}$	Parameters	GFLOPS
YOLOv5m	36.7	30.9	20.9M	48.3
YOLOv5l	37.2	31.6	46.1M	107.8
YOLOv5x	40.7	32.7	86.2M	203.9
YOLOv7	48.9	41.1	61.9M	103.3
YOLOv7x	50.1	42.3	70.9M	189.1
Ours	53.9	44.5	67.3M	210.3

detecting large, medium, and small objects, respectively; however, an additional Head is essential for detecting even smaller objects. Therefore, a fourth Head dedicated to detecting tiny objects is added. The introduction of the other head could ordinarily lead to a higher computational cost, but the adoption of GLF-Trans as a replacement for the conventional Neck structure offsets this increase in this model. As a result, the addition of the Tiny-head does not significantly escalate the overall computational burden. Furthermore, in the post-detection NMS phase, the adoption of Cluster-NMS [12] is expected to yield higher accuracy in dense scenes and situations with occlusion. In summary, the main improvements from the traditional YOLOv7 can be identified in three areas.

- The ConvModule and ELANBlock, previously utilized in the Neck section of YOLOv7, are removed and replaced with the GLF-Trans.
- Beyond the conventional three Heads, a fourth Tiny-Head is introduced to detect smaller objects.
- The traditional NMS is replaced with Cluster-NMS [12].

A. Global-Local Fusion Transformer(GLF-Trans)

The Global-Local Fusion Transformer (GLF-Trans) is constructed based on the Global-Detail Fusion Module (GDFM), an attention module utilized in Han et al.'s BRNet [11]. Its structure is depicted in Fig. 2. This architecture adapts the multi-head attention of a basic Transformer block to handle two different-sized inputs. The C3 level feature map (downsampled by a factor of 4) serves as the Query, and the C5 level feature map (downsampled by a factor of 32) acts as both Key and Value for input into the multi-head attention mechanism. Subsequently, the output of the multi-head attention is added to the C3 feature map along with the upsampled C5 feature map. Utilizing this module enables the fusion of global and local features and the suppression of background noise. From the perspective of fusing global and local features, this function can replace the PAN structure embedded in the Neck of the traditional YOLOv7. Therefore, by replacing the entire Neck with GLF-Trans, it is possible to maintain the detection accuracy of small objects while reducing parameters and computational load.

B. Tiny-Head

In small object detection, the size of the feature maps processed by the Head constitutes a critical factor. Net-

TABLE II: Ablation Experiment

Model	Tiny-head	GLF-Trans	Neck	Cluster-NMS	$mAP_{0.5}$	mAP_{small}	Parameters	GFLOPS
Base(YOLOv7x)	-	-	✓	-	42.3	11.9	70.9M	189.1
+Cluster-NMS	-	-	✓	✓	43.0	12.2	70.9M	189.1
+Tiny-Head	✓	-	✓	-	44.6	13.4	72.4M	226.8
+Tiny-Head w/o Neck	✓	-	-	-	41.8	12.4	66.9M	198.4
+Tiny-Head+GLF-Trans	✓	✓	-	-	44.4	13.1	67.3M	210.3
Ours	✓	✓	-	✓	44.5	13.3	67.3M	210.3

works in the RetinaNet lineage, such as Yang et al.'s QueryDet [9] and Chen et al.'s YOLOF [14], have identified the increased computational cost associated with a higher number of heads as a significant issue. However, the design of the YOLO series networks allows for improved detection performance while maintaining high computational speed. RetinaNet [15] employs numerous heads for multi-scale detection, leading to increased computational expenses and a decrease in computational speed. In contrast, YOLO processes the entire image in one go, resulting in an increase in computational cost, but it restrains the decrease in computational speed. The newly added Tiny-head, generated from low-level, high-resolution feature maps, exhibits high sensitivity towards small objects. Consequently, it mitigates the adverse effects of object scale variance, particularly enhancing detection accuracy for smaller objects.

IV. Experiment

A. Set Up

Dataset: In this study, we conduct a comprehensive analysis based on the VisDrone-DET2019 [8] dataset. This dataset is comprised of a diverse collection of images captured by drones, focusing on ten categories of objects that are closely related to daily life in urban environments, including pedestrians and various types of vehicles. With a total of 8,599 images, annotated with over 540,000 bounding boxes, the dataset highlights the diversity of objects and their complex overlaps, captured from various altitudes and locations. The dataset is divided into a training set of 6,471 images, a validation set of 548 images, and a test set of 1,580 images, collected from different locations under similar environmental conditions, facilitating a broad evaluation of the challenges in object detection. The utilization of this dataset represents an attempt to address some of the more difficult challenges in object detection, such as variations in object scale and the identification of overlapping objects. The images contain a rich diversity of overlapping objects, as well as a variety of sizes and shapes, making it an ideal dataset for testing the accuracy and robustness of small object detection technologies.

Metrics: In this study, we place particular emphasis on the evaluation metric $mAP_{0.5}$ for assessing performance. $mAP_{0.5}$ represents the mean Average Precision calculated with an Intersection over Union (IoU) threshold set at 0.5, evaluating detections as accurate if the predicted bounding boxes by the algorithm overlap more than 50% with the actual object bounding boxes. Utilizing this metric allows us to directly reflect the accuracy of predictions regarding the location and size of objects, thereby effectively capturing the actual detection performance of the algorithm. Additionally, we assess the detection accuracy for small objects (mAP_{small}) to validate the performance of the algorithm in small object detection. Moreover, as another crucial aspect of performance, we use GFLOPS and the number of parameters as metrics to evaluate processing speed.

Hyperparameters: In this study, we adopted the following hyperparameters for training the model. Stochastic Gradient Descent (SGD) is utilized for optimization, with an initial learning rate set at 0.01. This learning rate decreases gradually through the cosine annealing method to a final value of 0.1, after a warm-up period of 3 epochs. During the warm-up period, the momentum is set at 0.8 and the learning rate for biases at 0.1. Weight decay is established at 5×10^{-4} to mitigate overfitting. Furthermore, the model is trained on images of size 640x640 pixels, with a batch size of 10.

B. Performance Comparison

To evaluate the performance of the proposed model, we conduct a comparative analysis with other state-of-the-art single-stage detectors, including three variants of YOLOv5 [5] and two variants of YOLOv7 [7]. The results are presented in Tab. I. The proposed model has achieved a 3.8% improvement in $mAP_{0.5}$ on the validation set and a 2.2% improvement on the test set compared to the traditional YOLOv7x. The number of parameters has slightly decreased due to the removal of the Neck component. Regarding computational load, there was a slight increase compared to YOLOv7x, but this increase is not as significant as the increase in computational load from YOLOv7 to YOLOv7x. Overall, the proposed algorithm demonstrates its effectiveness in detecting small and densely packed objects contained within aerial photography datasets.

C. Ablation Study

To investigate the effectiveness of the Tiny-head, GLF-Trans, the existing Neck, and Cluster-NMS [12] in small object detection, we conduct an ablation study using YOLOv7x [7]. The results are shown in Tab. II. The

use of the Tiny-head resulted in a 2.3% improvement in mAP0.5, indicating its significant impact on small object detection. However, both the number of parameters and computational load increased. Therefore, replacing the existing Neck with GLF-Trans can reduce the number of parameters and computational load, although it slightly decreases the accuracy. Additionally, the adoption of Cluster-NMS [12] in YOLOv7x [7] improved mAP0.5 by 0.7% and mAPsmall by 0.3%, demonstrating its contribution to detecting small objects. In the final proposed model, mAP0.5 is improved by 2.2%, and $\text{mAP}_{\text{small}}$ is improved by 1.4%.

V. Conclusions

In this paper, we propose a real-time object detector with high accuracy for small object detection, based on YOLOv7x [7]. To optimize the network for small objects, we introduce an attention module, the Global-Local Fusion Transformer (GLF-Trans), and a fourth head, known as the Tiny-head. By adopting GLF-Trans in place of the traditional Neck, we are able to reduce parameters while learning the contextual information necessary for detecting small objects. The addition of the Tiny-head utilizes high-resolution feature maps, thereby enhancing the network's ability to recognize small objects. The proposed model demonstrates its effectiveness for small objects on the VisDrone-2019 [8] dataset, showing a 3.8% improvement in $\text{mAP}_{0.5}$ compared to conventional models.

References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[2] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263–7271.

[3] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[5] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, and Y. K. et al., ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, version v7.0, 2022. doi: 10.5281/zenodo.7347926. [Online]. Available: https://doi.org/10.5281/zenodo.7347926.

[6] C. Li, L. Li, H. Jiang, et al., "Yolov6: A single-stage object detection framework for industrial applications," arXiv preprint arXiv:2209.02976, 2022.

[7] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7464–7475.

[8] D. Du, P. Zhu, L. Wen, et al., "Visdrone-det2019: The vision meets drone object detection in image challenge results," in Proceedings of the IEEE/CVF international conference on computer vision workshops, 2019, pp. 0–0.

[9] C. Yang, Z. Huang, and N. Wang, "Querydet: Cascaded sparse query for accelerating high-resolution small object detection," in Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2022, pp. 13 668–13 677.

[10] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2778–2788.

[11] W. Han, J. Yin, X. Jin, X. Dai, and J. Shen, "Brnet: Exploring comprehensive features for monocular depth estimation," in European Conference on Computer Vision, Springer, 2022, pp. 586–602.

[12] Z. Zheng, P. Wang, D. Ren, et al., "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," IEEE transactions on cybernetics, vol. 52, no. 8, pp. 8574–8586, 2021.

[13] C.-Y. Wang, H.-Y. M. Liao, and I.-H. Yeh, "Designing network design strategies through gradient path analysis," arXiv preprint arXiv:2211.04800, 2022.

[14] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13 039–13 048.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.