DIP Final Project: Object Detection

Group members and division of works:
R10922084 Yi-Cheng Chen: Method improvement
R10944051 Chia-Sheng Hung: Method comparison
R11944031 Pin-Hsuan Peng: Paper survey, Report

1 Method

In this work, we compare four methods including YOLOv7, KFIoU, DETR, and DINO and try to do some improvement based on YOLOv7. YOLOv7 is the representative method in one-stage method; KFIoU with R3Det is the two-stage method; the last two method use transformer techniques which is very pioneering in object detection field. As for the dataset, KFIoU uses DOTA dataset which provides aerial image, while YOLOv7, DETR, and DINO uses COCO dataset.

1.1 YOLOv7

YOLOv7[1] is the latest method of YOLO series. As previous YOLO algorithms, YOLOv7 features one-stage and real-time detection. The authors proposes a new framework, E-ELAN, which adds expand, shuffle, and merge cardinality (see Figure 1). They don't add more transition layers but only change the structures of computational blocks. What's more, they proposes compound model scaling for their model scaling techniques. They scale the network depth and width when concatenating layers together to keep the model architecture optimal.

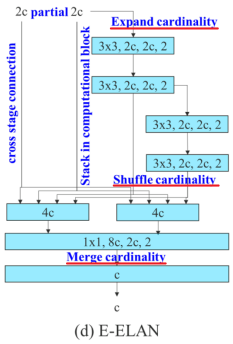

Figure 1: Framework of E-ELAN

1.2 KFIoU

KFIoU(Keypoint-based Free-Form Intersection over Union)[2] is based on the concept of IoU. Unlike IoU, which only considers the bounding boxes of the objects, KFIoU takes into account the keypoints of objects as well. Keypoints are specific points that are important for identifying or distinguishing the object, such as corners or points along the edges. With keypoints, KFIoU can be more accurate and more robust for rotated object and objects with different shape. In our work, we choose R3Det[3] as our backbone.

1.3 DETR

DETR(End-to-End Object Detection with Transformers)[4] is the first method in object detection that utilizes transformer-based network, which derives from Natural Language Processing(NLP). Figure 2 shows the pipeline of DETR. Different from traditional object detection method, DETR has no need of proposals or anchors for possible object position. Instead, it regards object detection as set detection,

using attention in both encoder and decoder to identify the relation in each part and thus focus on the important part automatically.

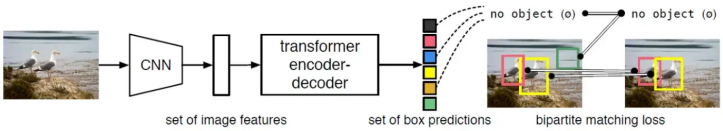

Figure 2: Pipeline of DETR

1.4 DINO

DINO(DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection)[5] further improves the performance of DETR by two main approaches: (1) contrastive way for denoising training and (2) mix query selection for anchor initialization. As Figure 3 , denoising training is that they put noise as negative sample in the input data to let the model identify positive and negative samples. Mix query selection is that they select a subset of position queries in encoder feature to let the query related to the input data.

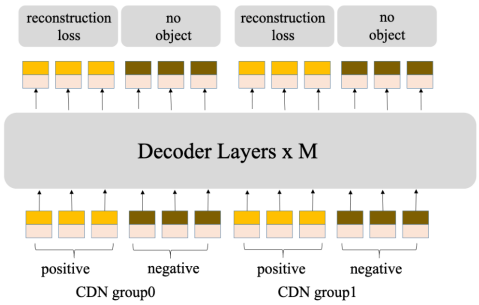

Figure 3: Denoising Training of DINO

2 Result

2.1 YOLOv7

From Figure 4, Figure 5 and Figure 6, We can find that with E-ELAN framwork, the performance of YOLOv7 in both daytime and night is significantly enhanced. The far objects are more possible to be detected, and the overlapping bounding boxes are reduced.

Figure 4: YOLOv7 Detection Result in Daytime Image - 1 The image on the left side is the result of YOLOv7 with traditional framework, and the other on the right side is the result of YOLOv7 with E-ELAN

Figure 5: YOLOv7 Detection Result in Daytime Image - 2 The image on the left side is the result of YOLOv7 with traditional framework, and the other on the right side is the result of YOLOv7 with E-ELAN

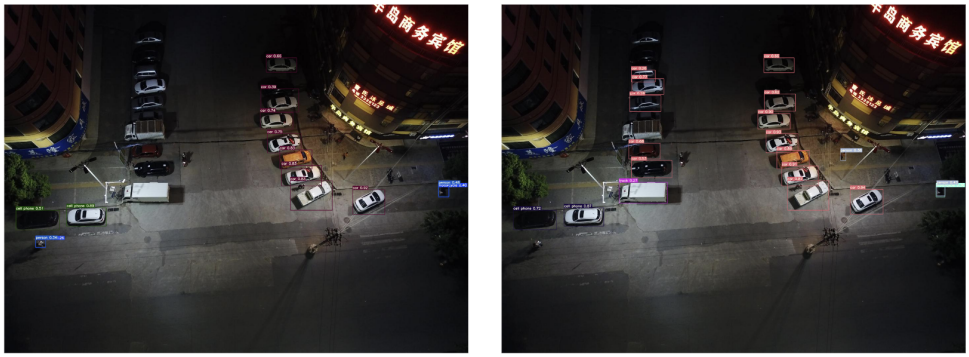

Figure 6: YOLOv7 Detection Result in Night Image The image on the left side is the result of YOLOv7 with traditional framework, and the other on the right side is the result of YOLOv7 with E-ELAN

2.2 KFIoU+R3Det

Largely affected by the training dataset, the accuracy of KFIoU performs better in specific objects. As Figure 7 shows, KFIoU focuses more on the objects with large-scale which are common in aerial images. While some details like people, KFIoU have bad performance. Same reason as previous result, when the shooting angle are smaller, the accuracy is extremely decreased. For example, KFIoU cannot recognize the side of the car but the top of the car (see Figure 8).

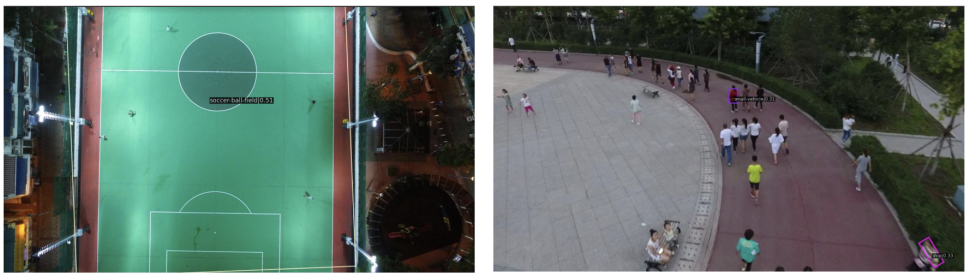

Figure 7: KFIoU Detection Results with different object scale The image on the left side is the result of KFIoU with large-scale object, and the other on the right side is the result with small-scale object

2.3 DETR

Comparing Figure 9 and Figure 4, we can find that DETR can detect more details as the result of YOLOv7 with E-ELAN. DETR is not real-time method, so this result meets out expectation.

Figure 8: KFIoU Detection Results with different shooting angle The image on the left side is the result of KFIoU with shooting angle near 90 degree, and the other on the right side is the result with smaller shooting angle

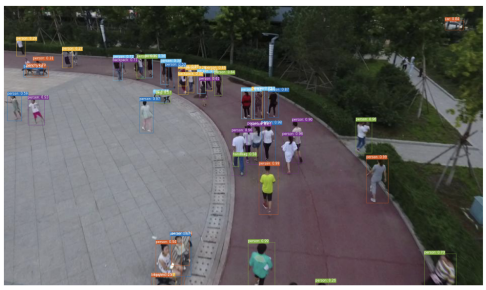

Figure 9: DETR Detection Results

2.4 DINO

Compare with YOLOv7 and DETR which are also trained with COCO dataset, DINO is more sensitive in aerial images (Figure 10) and in night images (Figure 11). In Figure 12, we can also find that the accuracy of DINO is highly improved (YOLOv7 detects all cars as "cellphone").

Figure 10: DINO Detection Results compare to other methods From left to right, the images shows the results of YOLOv7 with E-ELAN, DETR and DINO

3 Discussion

3.1 Comparison of each methods

DINO presents the best performance among these methods with wide range detection in different types of images. Nevertheless, it cannot support real-time detection like YOLOv7. With mix angles testing images, other methods still have some way to go.

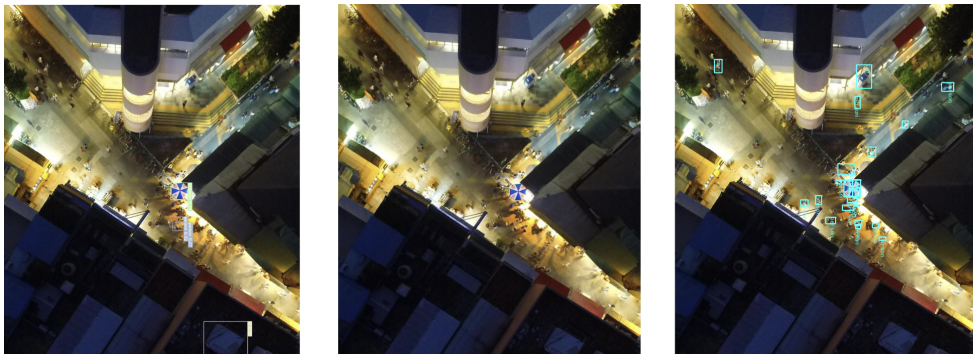

Figure 11: DINO Detection Results compare to other methods at night From left to right, the images shows the results of YOLOv7 with E-ELAN, DETR and DINO

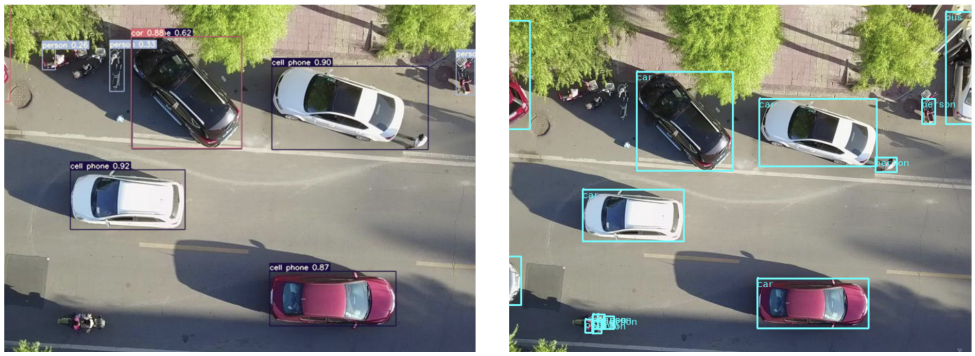

Figure 12: DINO Detection Results compare to YOLOv7 From left to right, the images shows the results of YOLOv7 with E-ELAN and DINO

3.2 Improvement

To reduce the effect of the training data and train a model which meets out testing data, we take parts of 3 datasets: VisDrone, Stanford Drone Dataset and MS COCO dataset and combine to our new training dataset. Based on YOLOv7, we do some finetune on the last hidden layer.[6] We finally train our own weight, combine VisDrone and COCO dataset, and thus significantly improving the performance. Our customized weight and some representative results are here: Finetune based on YOLOv7.

3.3 Future Work

Huge training data makes good performance, but it consumes too much resource. In the case of these testing dataset, we have to handle images from large-scale to small-scale, from 0 shooting degree to 90 shooting degree. If we can classify the images by shooting angle or their scale and put it to the corresponding pre-train model, we will improve accuracy and robustness. Maybe this is a direction that can be researched in the future.

References

[1] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022.

[2] Xue Yang, Yue Zhou, Gefan Zhang, Jitui Yang, Wentao Wang, Junchi Yan, Xiaopeng Zhang, and Qi Tian. The kfiou loss for rotated object detection. *arXiv preprint arXiv:2201.12558*, 2022.

[3] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3163–3171, 2021.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[5] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[6] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020.

Repositories: YOLOv7, KFIoU, DETR, DINO