DIP Final Project - Group 5
Object Detection on UAV images

R11922029 吳泓毅, R10922026 吳勝濬, R10922102 林正偉

Dataset

We collected our data from two datasets: VisDrone2019, DIOR. Our dataset include 10 classes: People, Bicycle, Car, Van, Truck, Tricycle, Awing-tricycle, Bus, Motor, Ship. We take ship data from DIOR, and take other classes from DIOR. We merge people and pedestrians in VisDrone2019 into one class to form 10 classes since these two classes are indifferent to our task. The following images show the samples of these two dataset. VisDrone2019 is collected by drone, and DIOR is collected from Google Earth. Therefore, our dataset contains every class needed for this task, and its context is similar to testing data.

VisDrone2019

DIOR

Original Class	In our situation
pedestrian	people
people	people
bicycle	bicycle
car	car
van	van
truck	truck
tricycle	tricycle
awning-tricycle	awning-tricycle
bus	bus
motor	motor

From VisDrone

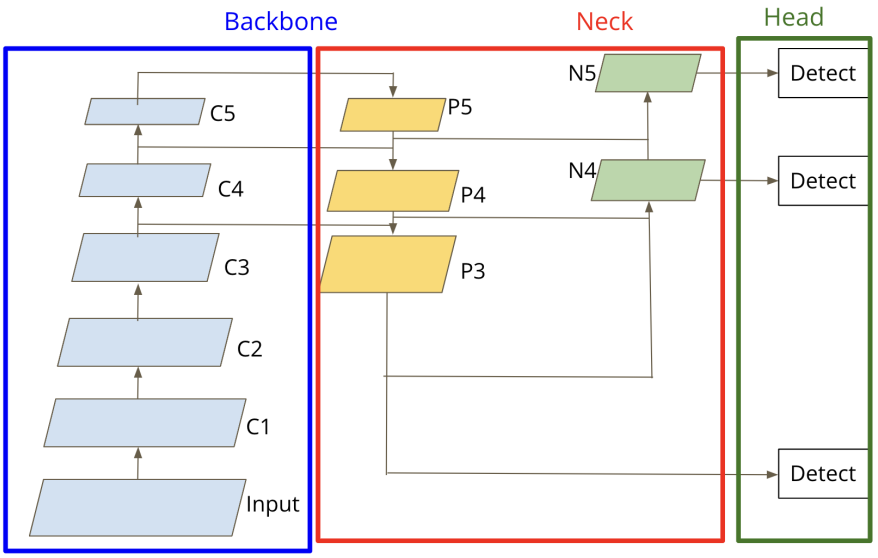
Original Class	In our situation
Airplane	-
Airport	-
Baseball field	-
Basketball court	-
Bridge	-
Chimney	-
Dam	-
Expressway service area	-
Expressway toll station	-
Golf field	-
Ground track field	-
Harbor	-
Overpass	-
Ship	Ship
Stadium	-
Storage tank	-
Tennis court	-
Train station	-
Vehicle	-
Windmill	-

From DIOR

Methodology

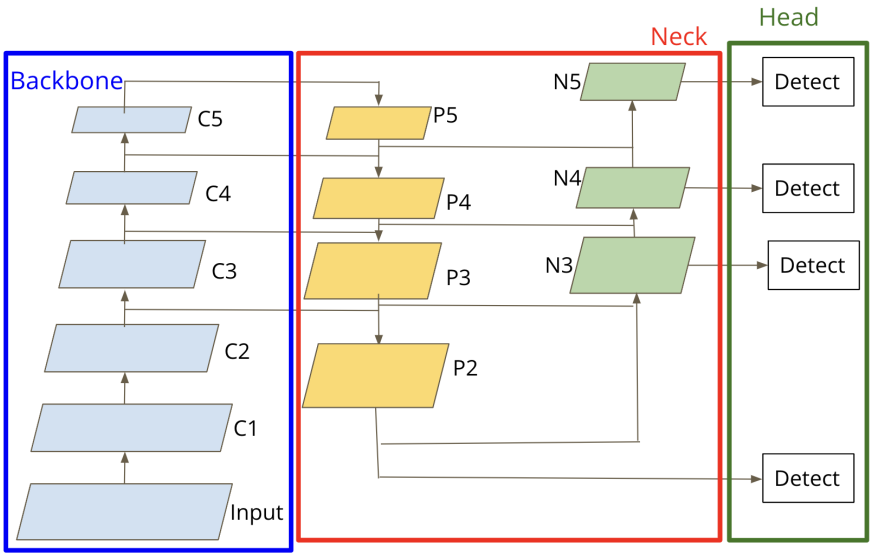
- YOLOv7

  YOLOv7 is a real time object detection model, it can be mainly separated into three parts, backbone, neck and head. Backbone is used to extract features, neck is used to do feature fusion, and head is used to do prediction. In the head part, there are three detectors which are used to detect multi scale objects. We chose YOLOv7 as our baseline.

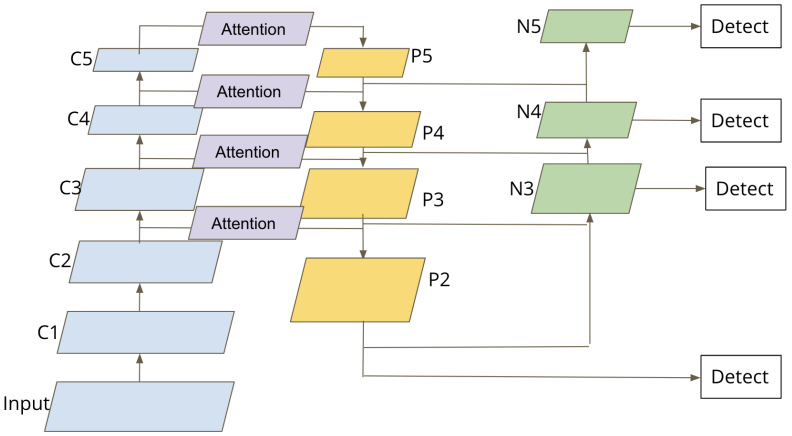


- YOLOv7 with additional detection head

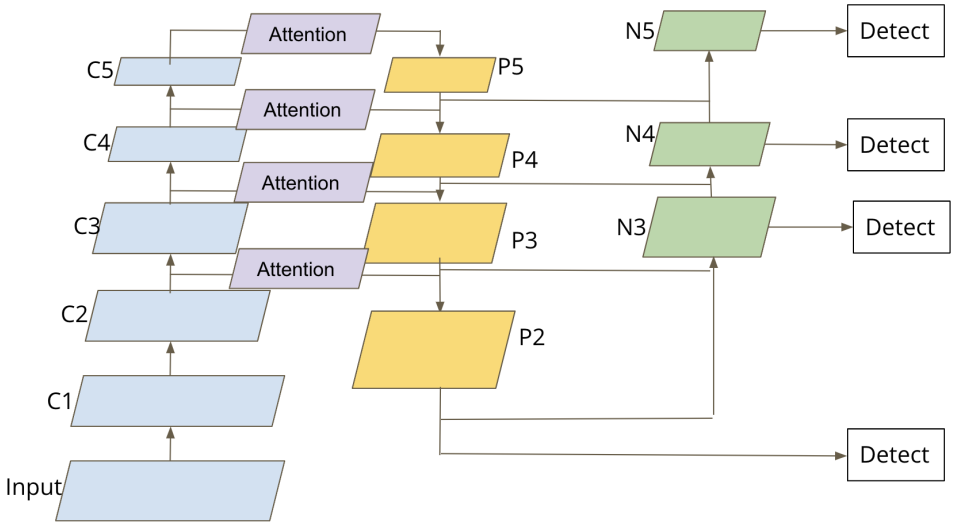
  It has been proved that the lower layers of convolution are used to extract texture information and deeper layers are used to extract semantic information. For small object detection, It is believed that texture information may improve the result of the model, so we extract one more feature map from backbone to do feature fusion and add one more head to do small object detection.



- YOLOv7 with attention module
  Attention module is used to make CNN learn and focus more on important features based on spatial and channel information. We considered that this can help the model to find objects that are occluded or placed in a cluttered background. So we add three attention modules to our model. The position where the module is placed is also a research region. We have tried a lot to find the best position, the following model structure can have the best mAP.



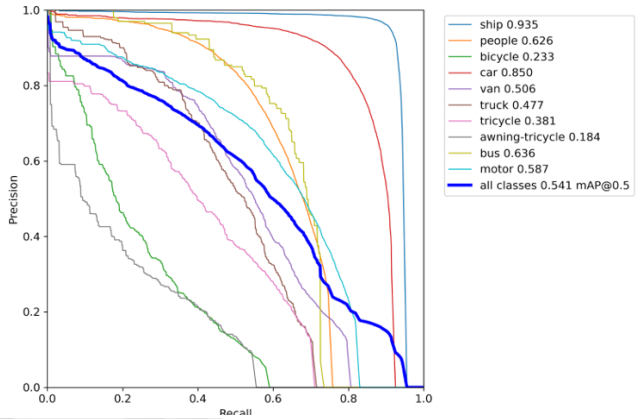
- YOLOv7 with additional detection head and attention module
  We mixed two methods mentioned before. We expect it may have more mAP improvement compared to when only one method is used.

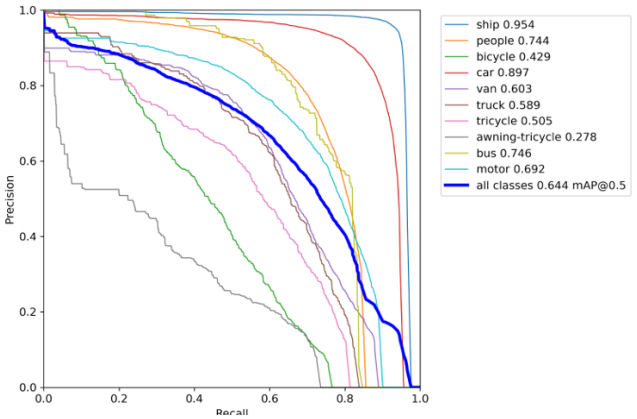


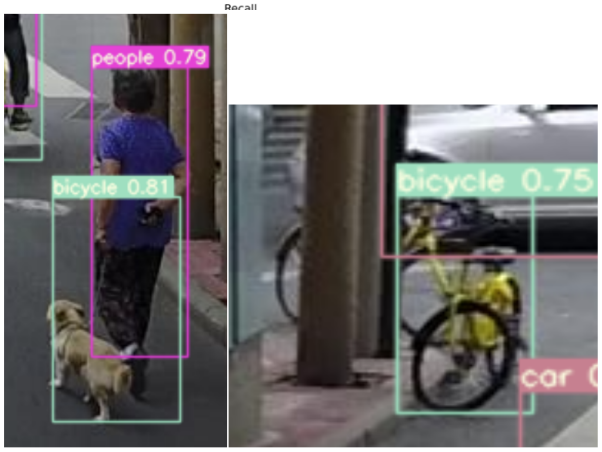
Experiment

- **Different image size: 640, 1280**

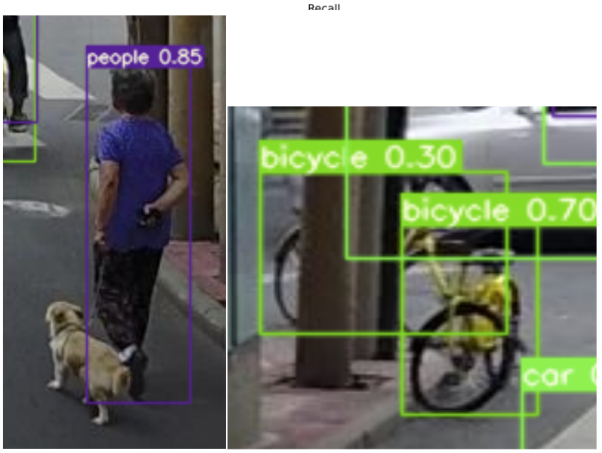
	mAP@.5	mAP@[.5, .95]
YOLOv7 - 1280	**64.4%**	**41.1%**
YOLOv7 - 640	54.1%	32.5%







640



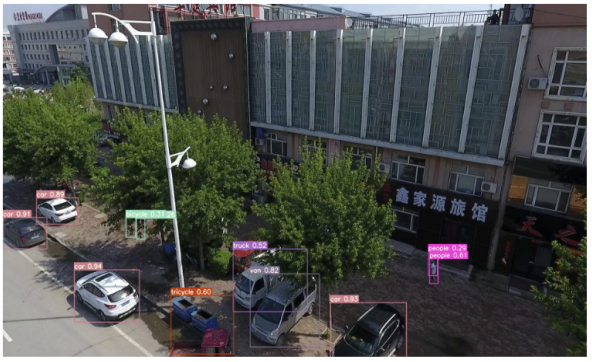
1280

640

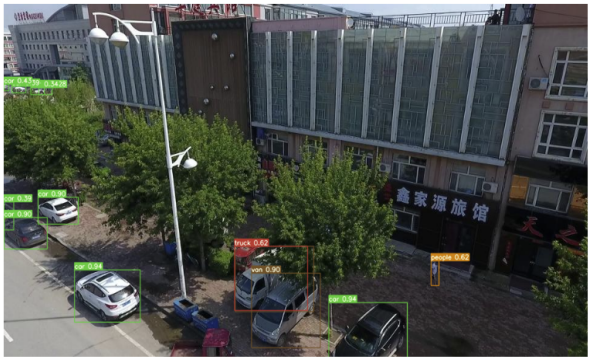
1280

- **Attention Module**

image size - 640	mAP@.5	mAP@[.5, .95]
YOLOv7	54.1%	32.5%
YOLOv7+CBAM	54.2%	32.2%
YOLOv7+RGCAB	54.4%	32.6%
YOLOv7+RGCAB3	**54.5%**	**32.6%**
YOLOv7+SE	**54.5%**	**32.6%**



YOLOv7 (640)



YOLOv7+SE (640)

YOLOv7 (640)

YOLOv7+SE (640)

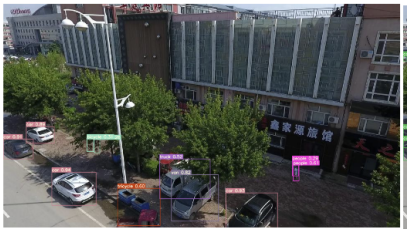
image size - 1280	mAP@.5	mAP@[.5, .95]
YOLOv7	**64.4%**	**41.1%**
YOLOv7+SE	64%	40.7%

YOLOv7 (1280)

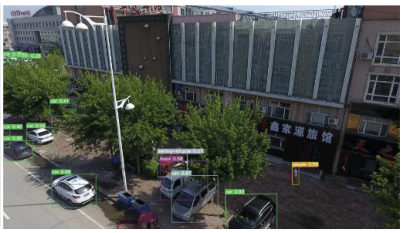
YOLOv7+SE (1280)

- **One more detection head + (Attention Module)**

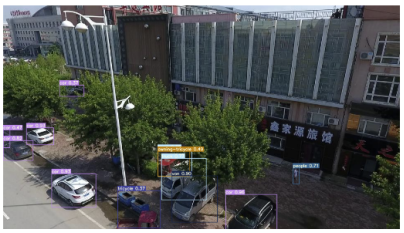
image size - 640	mAP@.5	mAP@[.5, .95]
YOLOv7	54.1%	32.5%
Four heads	55%	33.3%
Four heads + SE	**55.1%**	**33.6%**



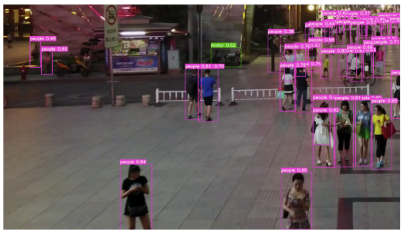
YOLOv7 (640)



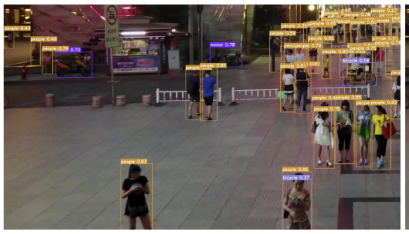
YOLOv7+Four heads (640)



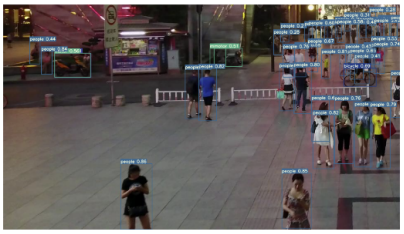
YOLOv7+Four heads+SE (640)



YOLOv7 (640)

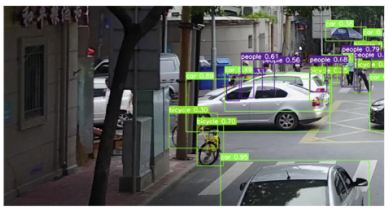


YOLOv7+Four heads (640)

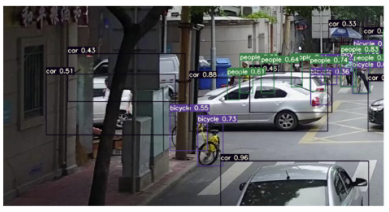


YOLOv7+Four heads+SE (640)

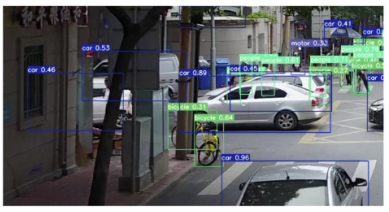
image size - 1280	mAP@.5	mAP@[.5, .95]
YOLOv7	64.4%	41.1%
Four heads	**64.6%**	**41.4%**
Four heads + SE	64.5%	41.1%



YOLOv7 (1280)



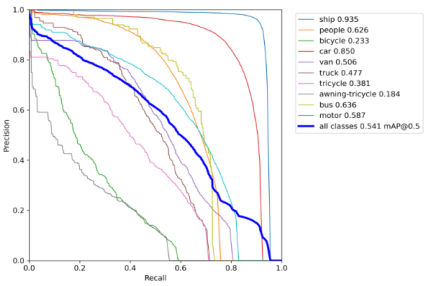
YOLOv7+Four heads (1280)



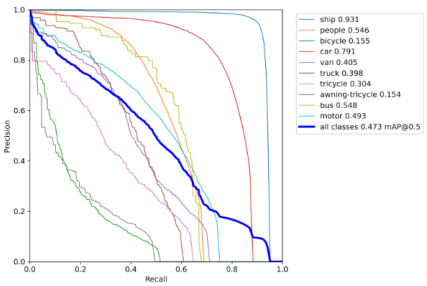
YOLOv7+Four heads+SE (1280)

- **Freeze Layers During Training**
  Since our dataset is a bit noisy, one technique to prevent the model from over-fitting the noisy data is transfer learning. We tried this technique by fixing layers to make our model more robust, but the performance dropped at the end. One possible reason is that our model is pre-trained on the COCO dataset, which is completely different from our dataset. So the model cannot fit our target domain well.

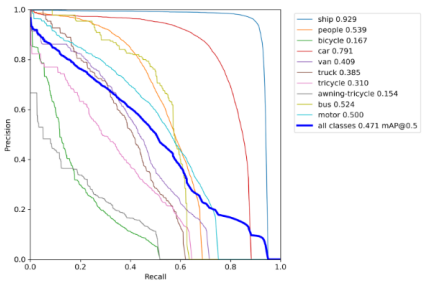
image size - 640	mAP@.5	mAP@[.5, .95]
YOLOv7	**54.1%**	**32.5%**
YOLOv7 Fix Backbone	47.3%	27.3%
YOLOv7 Fix Backbone and Neck	47.1%	27.3%



YOLOv7 - 640



YOLOv7 - Fix Backbone



YOLOv7 - Fix Backbone and Neck

Conclusion

- Shrinking may reduce texture information of the image, which is important for small objects and clustered objects.
- Adding attention modules is helpful for detecting objects that are occluded or placed in cluttered background when input size is 640, but may mislead the detector when input size is 1280.
- Extracting one more layer of feature map from backbone for feature fusion is effective to improve model accuracy.
- Freezing layers during training will reduce the performance.

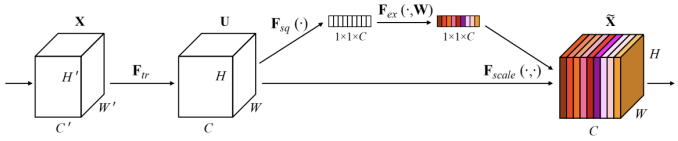
Division of Work

- Model training: 吳泓毅,吳勝濬,林正偉
- Presentation and Report: 吳泓毅,吳勝濬,林正偉
- Dataset: 吳泓毅

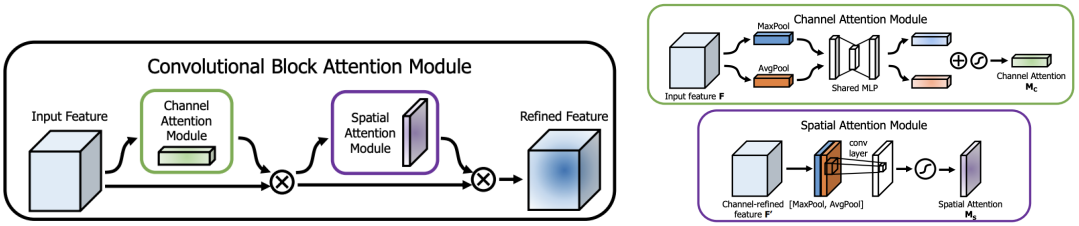
Appendix

- Module of attention module we used:

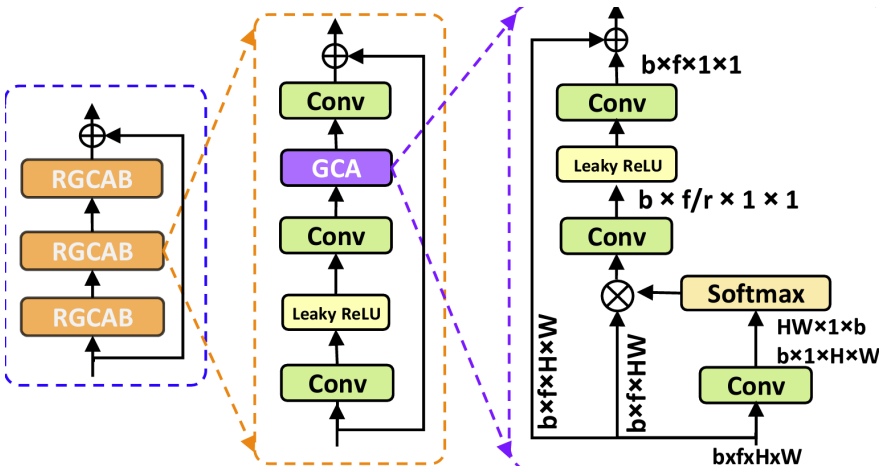
  Squeeze-and-Excitation(SE)



  Convolutional Block Attention Module(CBAM)



  Residual Global Context Attention Block(RGCAB)



References

[VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results]

[Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark]

[YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors]

[SPB-YOLO: An Efficient Real-Time Detector For Unmanned Aerial Vehicle Images]

[Squeeze-and-Excitation Networks]

[CBAM: Convolutional Block Attention Module]

[Burst Image Restoration and Enhancement]