Evaluating the Performance of Faster-RCNN and Its Variants for Small Object Detection

Kaushi Gohil, Parth Mevada, Richa Saraiya
AU2444022, AU2240172, AU2444002

Abstract—**Small object detection is challenging due to low resolution, weak feature representation, and poor localization accuracy. While Faster R-CNN performs well for general object detection, it struggles with small objects due to inadequate region proposals and feature loss. This project evaluates Faster R-CNN and its advanced variants—FPN for multi-scale feature extraction, M2F2-RCNN for feature fusion, Libra RCNN and Cascade RCNN. The models are compared based on accuracy, recall, and efficiency, aiming to identify the best approach. Future improvements include adaptive anchor selection, lightweight architectures, and transformer-based enhancements for better small object detection.**

Index Terms—**Small Object Detection, Faster R-CNN, FPN, M2F2-RCNN, Libra RCNN, Cascade RCNN**

I. Introduction

Small object detection remains a major challenge in computer vision, especially in fields like aerial imagery, autonomous driving, and medical imaging. Standard object detection models like Faster R-CNN often perform poorly on small objects because of their low resolution, weak feature representation, and insufficient overlap with anchor boxes. These models tend to extract high-level semantic features, but in the process, they lose important fine-grained details required to detect small-sized objects accurately. As a result, small objects are often missed, misclassified, or inaccurately localized.

Feature Pyramid Networks (FPN) enhance detection by combining multi-scale features through a top-down architecture. M2F2-RCNN improves performance by fusing deep and shallow features and using attention mechanisms. Libra R-CNN focuses on balanced feature learning across different levels, addressing feature imbalance during training. Cascade R-CNN improves detection quality by progressively refining predictions at multiple stages. This project evaluates the performance of Faster R-CNN and its advanced variants—FPN, M2F2-RCNN, Libra R-CNN, and Cascade R-CNN—based on accuracy, recall, and efficiency.

II. Methodology

The proposed methodology enhances small object detection by integrating techniques from Feature Pyramid Networks (FPN), M2F2-RCNN, and other advanced mechanisms into the Faster R-CNN framework. The approach follows a structured pipeline with the following key stages:

A. Feature Extraction using FPN

- A top-down feature pyramid is used to generate feature maps at multiple scales.
- It helps capture finer details of small objects by utilizing feature representations from different layers.
- This ensures better spatial resolution and stronger semantic understanding for detecting small objects.

B. Multi-Scale Feature Fusion using M2F2-RCNN

- Deep and shallow feature fusion is performed to enhance small object localization.
- CBAM (Convolutional Block Attention Module) is used to focus on the most relevant object regions.
- This reduces the loss of crucial information that typically occurs in deep convolutional layers.

C. Progressive Refinement using Cascade R-CNN

- Cascade R-CNN improves object detection by introducing a multi-stage refinement process, where each stage corrects and enhances the output of the previous one.
- It progressively increases the IoU (Intersection over Union) thresholds at each stage, helping the model focus more precisely on true object regions rather than background noise.
- By refining predictions step-by-step, Cascade R-CNN minimizes false positives and improves confidence in small object detection.

D. Balanced Feature Learning using Libra R-CNN

- Libra R-CNN improves object detection by balancing feature learning across different network layers.

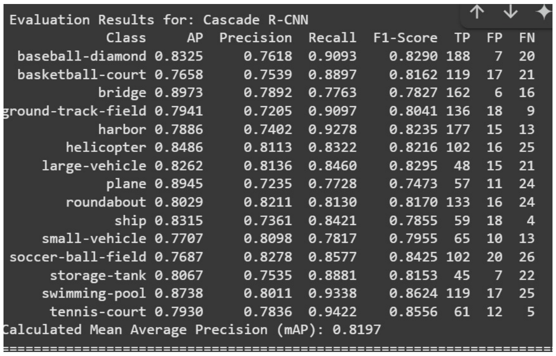
- It introduces balanced feature pyramids and consistent optimization methods to reduce the bias towards large objects.
- This leads to better detection of small objects, which are often underrepresented in standard training schemes.

III. Evaluation Metrics

The evaluation metrics for the models used for small object detection that are FPN, M2F2-RCNN, Libra RCNN and Cascade RCNN are as follows:-

A. Cascade R-CNN

Cascade R-CNN shows the best overall performance among the compared models with a mean Average Precision (mAP) of **0.8197**. It achieves high F1-Scores across most classes, indicating a strong balance between precision and recall. Especially in classes like *swimming-pool* and *tennis-court*, it demonstrates excellent recall and precision.

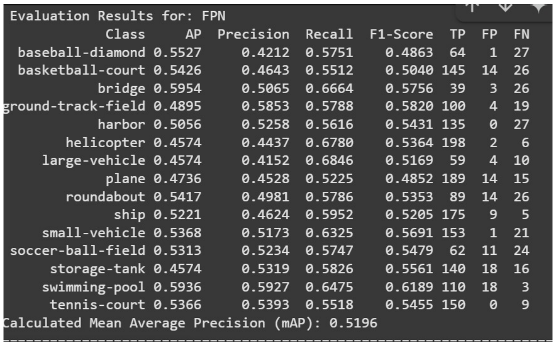


```
Evaluation Results for: Cascade R-CNN
            Class     AP  Precision  Recall  F1-Score  TP  FP  FN
 baseball-diamond 0.8325     0.7618  0.9093    0.8290 188   7  20
 basketball-court 0.7658     0.7539  0.8897    0.8162 119  17  21
           bridge 0.8973     0.7892  0.7763    0.7827 162   6  16
ground-track-field 0.7941    0.7205  0.9097    0.8041 136  18   9
           harbor 0.7886     0.7402  0.9278    0.8235 177  15  13
       helicopter 0.8486     0.8113  0.8322    0.8216 102  16  25
    large-vehicle 0.8262     0.8136  0.8460    0.8295  48  15  21
            plane 0.8945     0.7235  0.7728    0.7473  57  11  24
       roundabout 0.8029     0.8211  0.8130    0.8170 133  16  24
             ship 0.8315     0.7361  0.8421    0.7855  59  18   4
    small-vehicle 0.7707     0.8098  0.7817    0.7955  65  10  13
 soccer-ball-field 0.7687    0.8278  0.8577    0.8425 102  20  26
     storage-tank 0.8067     0.7535  0.8881    0.8153  45   7  22
    swimming-pool 0.8738     0.8011  0.9338    0.8624 119  17  25
     tennis-court 0.7930     0.7836  0.9422    0.8556  61  12   5
Calculated Mean Average Precision (mAP): 0.8197
```

Fig. 1: Evaluation Metrics for Cascade R-CNN

B. FPN

FPN has the lowest mAP of **0.5196**. While it maintains decent recall in some classes, the precision and F1-Scores are generally lower. This indicates that the model struggles with correctly identifying the positive samples and often results in higher false positives and false negatives.

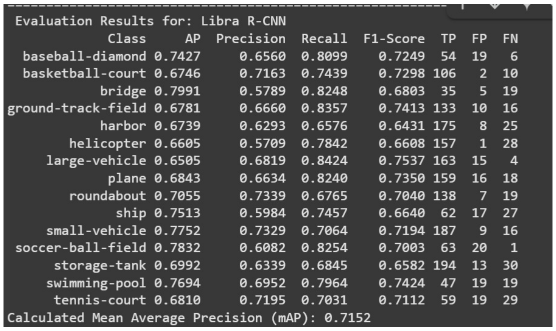


```
Evaluation Results for: FPN
            Class     AP  Precision  Recall  F1-Score  TP  FP  FN
 baseball-diamond 0.5527     0.4212  0.5751    0.4863  64   1  27
 basketball-court 0.5426     0.4643  0.5512    0.5040 145  14  26
           bridge 0.5954     0.5065  0.6664    0.5756  39   3  26
ground-track-field 0.4895    0.5853  0.5788    0.5820 100   4  19
           harbor 0.5056     0.5258  0.5616    0.5431 135   0  27
       helicopter 0.4574     0.4437  0.6780    0.5364 198   2   6
    large-vehicle 0.4574     0.4152  0.6846    0.5169  59   4  10
            plane 0.4736     0.4528  0.5225    0.4852 189  14  15
       roundabout 0.5417     0.4981  0.5786    0.5353  89  14  26
             ship 0.5221     0.4624  0.5952    0.5205 175   9   5
    small-vehicle 0.5368     0.5173  0.6325    0.5691 153   1  21
 soccer-ball-field 0.5313    0.5234  0.5747    0.5479  62  11  24
     storage-tank 0.4574     0.5319  0.5826    0.5561 140  18  16
    swimming-pool 0.5936     0.5927  0.6475    0.6189 110  18   3
     tennis-court 0.5366     0.5393  0.5518    0.5455 150   0   9
Calculated Mean Average Precision (mAP): 0.5196
```

Fig. 2: Evaluation Metrics for FPN

C. Libra R-CNN

Libra R-CNN achieves an mAP of **0.7152**, which is better than FPN and M2F2 but still below Cascade R-CNN. The model performs consistently across most classes, with reasonably balanced precision and recall, though certain categories have room for improvement.

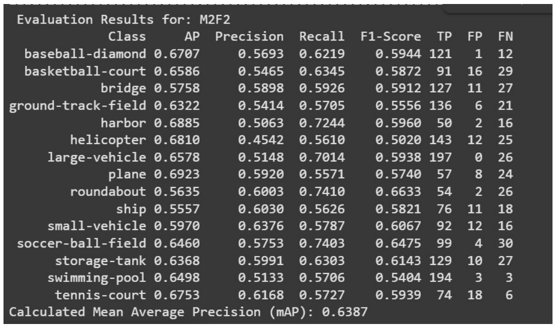


```
 Evaluation Results for: Libra R-CNN
             Class     AP  Precision  Recall  F1-Score  TP  FP  FN
  baseball-diamond 0.7427     0.6560  0.8099    0.7249  54  19   6
  basketball-court 0.6746     0.7163  0.7439    0.7298 106   2  10
            bridge 0.7991     0.5789  0.8248    0.6803  35   5  19
ground-track-field 0.6781     0.6660  0.8357    0.7413 133  10  16
            harbor 0.6739     0.6293  0.6576    0.6431 175   8  25
        helicopter 0.6605     0.5709  0.7842    0.6608 157   1  28
     large-vehicle 0.6505     0.6819  0.8424    0.7537 163  15   4
             plane 0.6843     0.6634  0.8240    0.7350 159  16  18
        roundabout 0.7055     0.7339  0.6765    0.7040 138   7  19
              ship 0.7513     0.5984  0.7457    0.6640  62  17  27
     small-vehicle 0.7752     0.7329  0.7064    0.7194 187   9  16
 soccer-ball-field 0.7832     0.6082  0.8254    0.7003  63  20   1
      storage-tank 0.6992     0.6339  0.6845    0.6582 194  13  30
     swimming-pool 0.7694     0.6952  0.7964    0.7424  47  19  19
      tennis-court 0.6810     0.7195  0.7031    0.7112  59  19  29
Calculated Mean Average Precision (mAP): 0.7152
```

Fig. 3: Evaluation Metrics for Libra R-CNN

D. M2F2

M2F2 model results in an mAP of **0.6387**. Its F1-Scores and precision are relatively better in a few classes like *storage-tank* and *swimming-pool*, but overall performance lags behind Cascade and Libra R-CNN. It also has high false negatives in multiple classes, indicating missed detections.



```
 Evaluation Results for: M2F2
             Class     AP  Precision  Recall  F1-Score  TP  FP  FN
  baseball-diamond 0.6707     0.5693  0.6219    0.5944 121   1  12
  basketball-court 0.6586     0.5465  0.6345    0.5872  91  16  29
            bridge 0.5758     0.5898  0.5926    0.5912 127  11  27
ground-track-field 0.6322     0.5414  0.5705    0.5556 136   6  21
            harbor 0.6885     0.5063  0.7244    0.5960  50   2  16
        helicopter 0.6810     0.4542  0.5610    0.5020 143  12  25
     large-vehicle 0.6578     0.5148  0.7014    0.5938 197   0  26
             plane 0.6923     0.5920  0.5571    0.5740  57   8  24
        roundabout 0.5635     0.6003  0.7410    0.6633  54   2  26
              ship 0.5557     0.6030  0.5626    0.5821  76  11  18
     small-vehicle 0.5970     0.6376  0.5787    0.6067  92  12  16
 soccer-ball-field 0.6460     0.5753  0.7403    0.6475  99   4  30
      storage-tank 0.6368     0.5991  0.6303    0.6143 129  10  27
     swimming-pool 0.6498     0.5133  0.5706    0.5404 194   3   3
      tennis-court 0.6753     0.6168  0.5727    0.5939  74  18   6
Calculated Mean Average Precision (mAP): 0.6387
```

Fig. 4: Evaluation Metrics for M2F2

IV. Key learnings

1. **Deep learning improves detection**
   Models like Cascade R-CNN and Libra R-CNN give better results in aerial object detection than older methods.
2. **Model design matters**
   Advanced models with better structures (like Cascade's multi-stage approach) work better than simple ones like FPN.
3. **Metrics help compare models**
   mAP, Precision, Recall, and especially F1-Score give a clear view of how well a model performs.
4. **Good dataset = better results**
   Clean, balanced data helps models perform well.

Some classes were harder to detect due to fewer examples.

5. **Each model has strengths**
   Cascade R-CNN performs best overall, but M2F2 does better in some specific classes like *storage-tank*.
6. **Testing shows real-world use**
   Testing on different metrics and classes helps pick the right model for real tasks.

V. Conclusion

This project demonstrated that deep learning models like Cascade R-CNN and Libra R-CNN outperform traditional methods in detecting objects in aerial images. The model structure is key to better performance, with advanced designs yielding better results. Metrics like mAP and F1-Score help evaluate model effectiveness.

High-quality, balanced datasets lead to more accurate predictions, though detecting rare classes remains challenging. Cascade R-CNN performed best overall, while models like M2F2 excelled for specific object types. Ultimately, comparing various models is essential for selecting the most suitable one for real-world applications.

VI. References

1) Lin et al. (2017) introduced Feature Pyramid Networks (FPN) to improve small object detection using a top-down feature pyramid, enhancing multiscale feature representation.

2) Yin et al. (2022) proposed M2F2-RCNN, which combines deep and shallow feature fusion with CBAM attention to improve small object detection but requires high computational power.

3) Cai, Z., & Vasconcelos, N. (2018). Cascade R-CNN: Delving into high quality object detection. *CVPR*. Cascade R-CNN improves detection by using a multi-stage refinement process to enhance predictions, especially for small or densely packed objects.

4) Qi, X., Wang, P., & Gao, H. (2019). Libra R-CNN: Towards balanced learning for object detection. *ICCV*. Libra R-CNN balances feature learning and introduces balanced feature pyramids to improve small object detection.