Enhancement: Cross-Modality Attentive Feature Fusion (CMAFF) for Object Detection in Multispectral Remote Sensing Imagery

Aditya Khandelwal
Roll Number: 210059

1 Introduction

This report presents implementation and evaluation of the enhancements over my original code of the paper based on multispectral object detection. My task was to use both RGB and thermal images to detect objects more accurately. Github link : IPR_Project_Part1

2 Proposed Enhancemnts & Implementation Details

The following enhancements were implemented in the code (as proposed in the previous document) to improve the performance of multispectral object detection by integrating RGB and thermal images more effectively.

2.1 Sequential Arrangement of CMAFF Modules

The Common Selective Module (CSM) and Differential Enhancive Module (DEM) were restructured sequentially as opposed to parallely in the original code.

2.2 Dimensionality Reduction with 1×1 Convolution

To reduce dimensionality and enhance computational efficiency, a 1×1 convolution is applied after the sequential fusion of CSM and DEM outputs. It reduced the number of channels in the combined feature map without affecting spatial resolution, making feature extraction more efficient for the downstream layers.

2.3 Adaptive Histogram Equalization on Thermal Images

Adaptive histogram equalization is applied as a preprocessing step to enhance the contrast of thermal images. It improved visibility in low-contrast areas, helping the Differential Enhancive Module (DEM) in extracting thermal-specific features more effectively.

3 Results Comparison

3.1 Performance Matrix

Precision, recall, and mAP (mean Average Precision) were used as the evaluation metrics. The model originally achieved a precision of 0.62, recall of 0.31, and mAP@0.5 of 0.20. With the enhancement the model achieved a high precision of 74.87, recall of 0.2849 and mAP@0.5 of 0.2164.

Table 1: Models Comparison on VEDAI Dataset

Model	Precision	Recall	mAP@0.5
YOLOv5 + CMAFF (Before)	**0.6255**	**0.3098**	**0.2082**
YOLOv5 + CMAFF (After)	**0.7487**	**0.2849**	**0.2164**

3.2 Computational efficiency

The training and evaluation was performed on a machine with NVIDIA GeForce RTX 4050. The training time of initial model was 162m 9.0s. With the changes the training time is significantly reduced to 91m 27.1s.

4 Conclusion

The YOLOv5 model's performance for multispectral object detection was greatly enhanced by the changes made. The precision rose up from 0.62 to 0.74 and the computational time reduced from 162 minutes to 91 minutes.