Implementation: 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 my paper on multispectral object detection which uses CMAFF module to enhance detection accuracy. The dataset used for this task is VEDAI. I have based my implementation as given in the paper using differential enhancive module and common selective module then combining them. My task was to use both RGB and thermal images to detect objects more accurately. Github link : IPR_Project_Part1

2 Dataset Description

I have downloaded a part of the VEDAI dataset and used around 400 pairs of images for training and 200 pairs for testing. The dataset consists of paired RGB and IR images, each with their corresponding text file containing annotations. The naming convention for the files is as follows:

- **RGB Image**: Named as `*_co.png`
- **IR Image**: Named as `*_ir.png`
- **Annotation File**: Named as `*.txt`

Each image is resized to a fixed resolution of 640x640 pixels during preprocessing. The object annotations are provided in text files, where each line corresponds to one object in the image. The format of the annotation is as follows:

```
<class_id> <x_center> <y_center> <width> <height>
```

All coordinates and dimensions are normalized between 0 and 1 to ensure consistency across different image resolutions. Data augmentation is applied to the training data using the following techniques:

- **Random Horizontal Flip**: Randomly flips the image horizontally.
- **Random Rotation**: Applies a random rotation (up to 10 degrees).
- **Color Jitter**: Randomly alters the brightness, contrast, saturation, and hue of the image.

These augmentations helped improving the model's generalization ability by introducing variability into training data.
The input to the model consists of concatenated RGB and IR image tensors. The RGB and IR images are combined along the channel dimension, resulting in a 6-channel input tensor:

- **Input Tensor Shape**: [6, H, W], where:
  - The first 3 channels represent the RGB image.
  - The next 3 channels represent the replicated IR image.

3 Implementation Details

I have used YOLOv5s model architecture and combined it with CMAFF module as given in the paper. The CMAFF module consists of two parts: the Differential Enhancive Module (DEM) and the Common Selective Module (CSM). The DEM focuses on enhancing the differential features between RGB and thermal images, while the CSM selects and refines common features.

3.1 Network Architecture

The architecture defines a two-stream backbone network, taking RGB and thermal images, then features of both modalities are fused by the CMAFF module. Finally, bounding boxes along with object classes are predicted from a layer called the detection head, in which the feature maps got fused are passed through.

3.2 Training Parameters

The model was trained using the Adam optimizer with a learning rate of 1e-4, and a batch size of 4. The model was trained only for 20 epochs due to computational limitations only on nearly 400 images. Only the newly added layers (CMAFF and detection head) were fine-tuned, while the backbone network was frozen during training.

Table 1: Performance Metrics

Metric	Value
Precision	0.6255
Recall	0.3098
mAP@0.5	0.2082

4 Results

Precision, recall, and mAP (mean Average Precision) were used as the evaluation metrics. The model achieved a precision of 0.62, recall of 0.31, and mAP@0.5 of 0.20.

5 Comparison with Other Models

I could not directly compare other models with the YOLOv5 + CMAFF module because this would require training those models on the same VEDAI dataset. Given the computational limitations and small dataset size (only 400 images), it was not feasible to train models like SSD, Faster R-CNN, or RetinaNet from scratch. Thus, the performance comparison presented in Table 2 is based on results from similar models applied to other multispectral imagery datasets, as reported in the paper.

Table 2: Model Comparison on VEDAI Dataset

Model	Precision	Recall	mAP@0.5
SSD	0.6090	0.6110	0.6090
Faster R-CNN	0.6460	0.6490	0.6460
YOLOv3	0.7300	0.7340	0.7300
RetinaNet	0.5990	0.6010	0.5990
YOLOv5 + CMAFF (Actually tested)	**0.6255**	**0.3098**	**0.2082**

6 Challenges in Achieving Accurate Results

While the proposed model architecture incorporates the Cross-Modality Attentive Feature Fusion (CMAFF) module to use both RGB and thermal images, several factors contribute to its bad performance:

6.1 Pretrained Weights on YOLOv5s

The YOLOv5s model is pretrained, but it was trained on datasets like MS COCO and not on the aerial imagery datasets like VEDAI. Natural image distributions and contents are very different from those in aerial images whereby the objects appear relatively small or even upside down. This causes the pretrained weights to not fit the task of small object detection well in the multispectral aerial imagery, resulting in poor performance.

6.2 Small Training Dataset

The model was trained on only 400 images from the VEDAI dataset due to computational limitations. With a small dataset, the model is likely to overfit, especially when dealing with the varied and challenging backgrounds present in aerial imagery.

6.3 Limited Computational Resources

Training a large and complex model such as YOLOv5 with CMAFF on the multispectral images requires huge computing, especially memory, and high speeds on GPUs. Complete training of the model on the entire VEDAI dataset, elaborate hyperparameter tuning, or even longer training sessions were impossible due to limited computing capabilities and slow internet speed.

7 Conclusion

The YOLOv5 model's performance for multispectral object detection would have been greatly enhanced by the cross-modality attentive feature fusion technique. However, the model's performance falls short of its potential because of constraints like the pretrained model, a small dataset, and limited computing power.