Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery

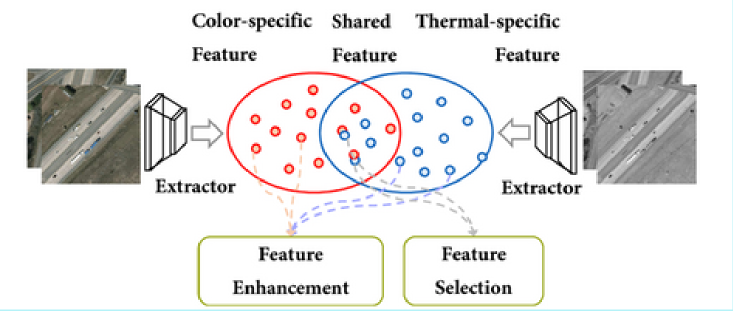
Authors: Fang Qingyun, Wang Zhaokui
Aditya Khandelwal (210059)

Introduction

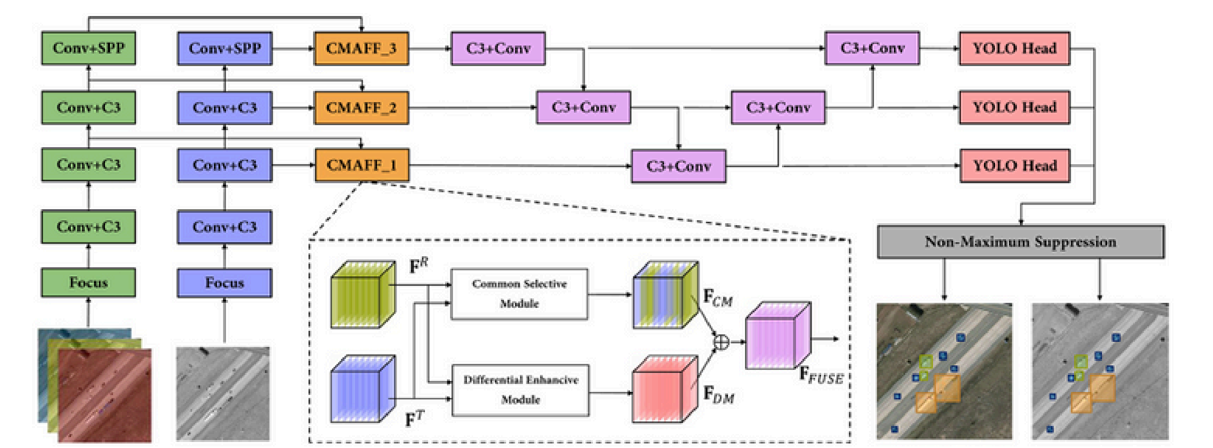
In remote sensing, accurately detecting objects is essential for applications across both civilian and industrial fields. However, conventional methods relying on RGB imagery are often limited by environmental conditions that affect visibility. Recent advancements show that integrating additional data sources, such as thermal imagery, provides complementary information that enhances detection capabilities.

Idea and Motivation

This paper introduces Cross-Modality Attentive Feature Fusion (CMAFF), a method that combines RGB and thermal imagery to enhance object detection. CMAFF selectively boosts unique features from each modality while preserving shared information, addressing the limitations of traditional methods under low visibility, like nighttime conditions. This approach improves detection reliability, essential for applications in security, traffic monitoring, and urban planning that demand consistent performance in diverse environments.

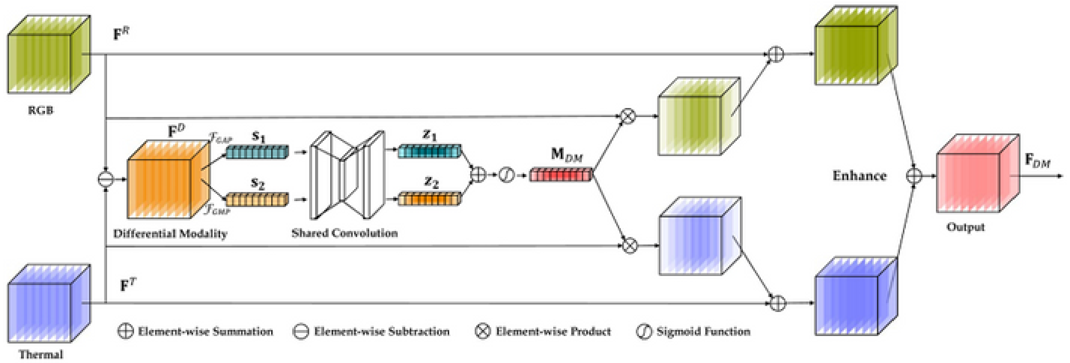


Architecture

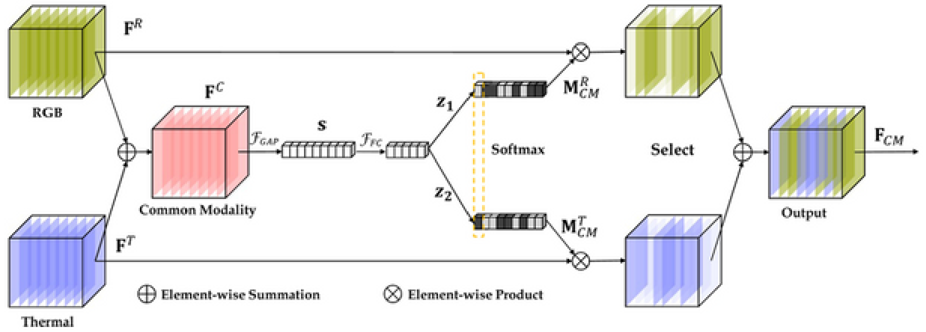


The CMAFF architecture enhances object detection by combining features from RGB and thermal images, each processed through distinct CNN streams. Key components:

- Differential Enhancive Module (DEM): Enhances unique features in each modality—color and texture in RGB, heat signatures in thermal—by focusing on differences between RGB and thermal data. These modality-specific features are amplified to improve detection accuracy.



- Common Selective Module (CSM): Selects shared features, such as shapes, common to both modalities. Using a SoftMax attention mechanism, CSM emphasizes the most reliable features, reducing redundancy and focusing on relevant information.



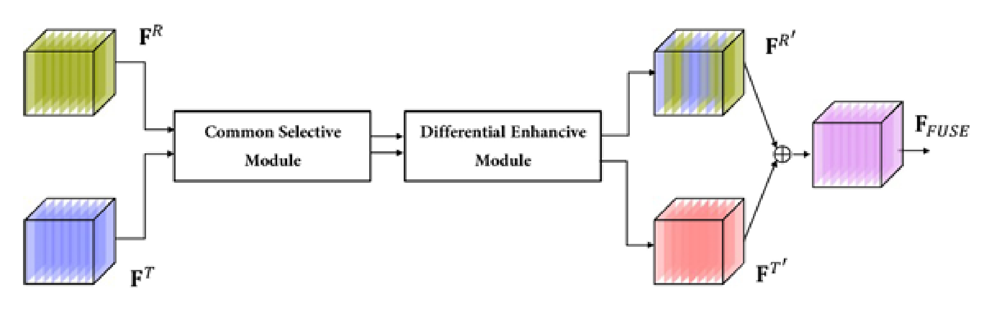
- Fusion in YOLOFusion: Outputs from DEM and CSM are combined and fed into the YOLOv5-based YOLOFusion detection framework. The fused features undergo multi-scale processing for accurate object detection across diverse conditions.

Results (Paper vs mine)

	Precision	Recall	MAP0.5
Paper:	0.81	0.697	0.786
My(Part 1):	0.63	0.309	0.208

Enhancements to the approach

- Sequential Arrangement of Modules: Unlike the original parallel setup, the Common Selective Module (CSM) and Differential Enhancive Module (DEM) were arranged sequentially. First, CSM filters irrelevant details from shared features, refining the common information from RGB and thermal images. DEM then enhances modality-specific features, resulting in sharper and more accurate detection by reducing redundancy.



(b) Common-first Sequential Manner

- Dimensionality Reduction with 1x1 Convolution: A 1x1 convolution is applied after sequential processing to reduce the dimensionality of the combined feature map. This step focuses on informative channels, making the model more computationally efficient without compromising spatial detail.
- Thermal Image Contrast Enhancement: Adaptive histogram equalization is applied to the thermal images, enhancing contrast and making thermal-specific features more prominent. This preprocessing step improves DEM's ability to detect unique thermal features, especially useful under low-contrast conditions.

Results (Before vs After)

	Precision	Recall	MAP0.5	Training Time
Part 1:	0.63	0.309	0.208	162 minutes
Part 2:	0.75	0.285	0.216	91 minutes

Limitations

The high precision with low recall and mAP suggests that the model is overly selective, detecting only the most obvious cases, which is likely due to insufficient training data, limiting recall and overall performance. With adequate data, the model should achieve results comparable to those reported in the paper.