Multispectral Object Detection

Bachhav Aryan Kishor

Indian Institute of Tecnhology Kanpur

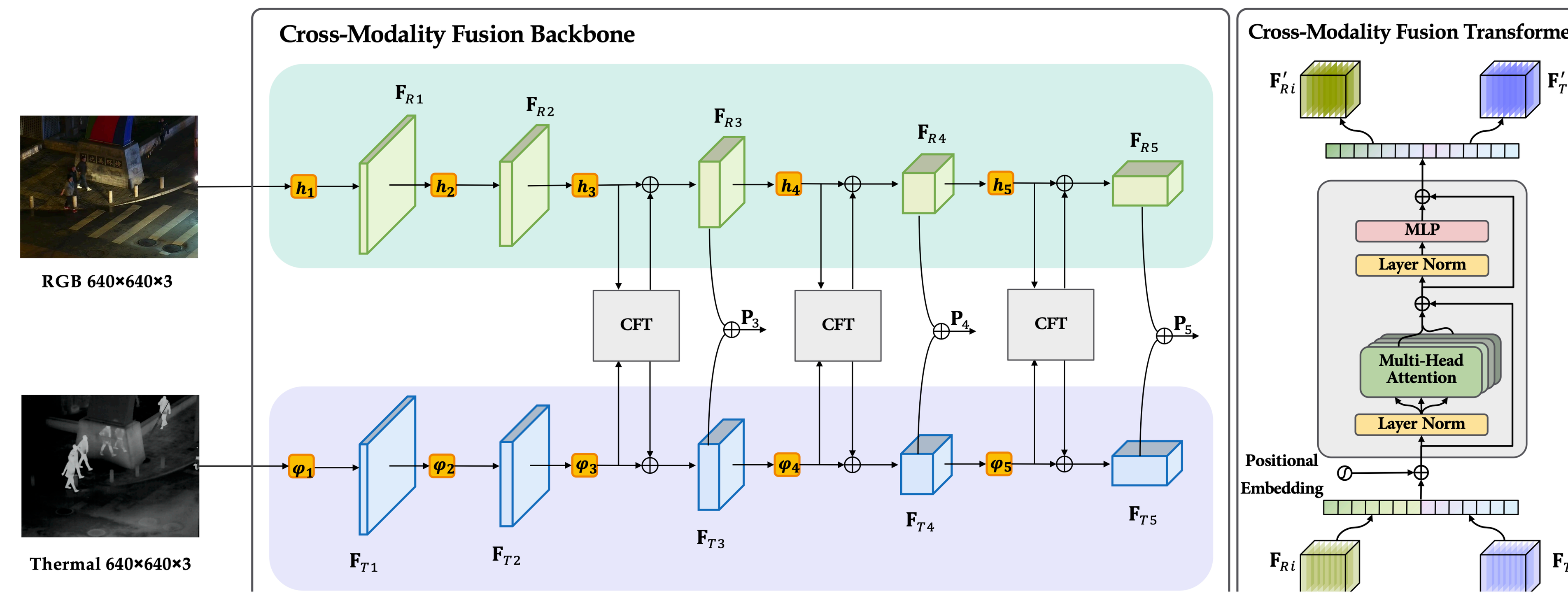
Abstract

- Multispectral image pairs increase object detection by combining RGB and Thermal information for reliability.
- Proposed Cross-Modality Fusion Transformer (CFT) utilizes the Transformer framework, unlike CNN-based approaches.
- CFT leverages self-attention to enable simultaneous intra- and inter-modality fusion.
- Captures interactions between RGB and Thermal domains, improving multispectral detection performance.
- Experiments show CFT achieves state-of-the-art results in multispectral object detection.
- CFT's design allows for effective integration of long-range dependencies, providing enhanced contextual awareness across modalities.

Goal

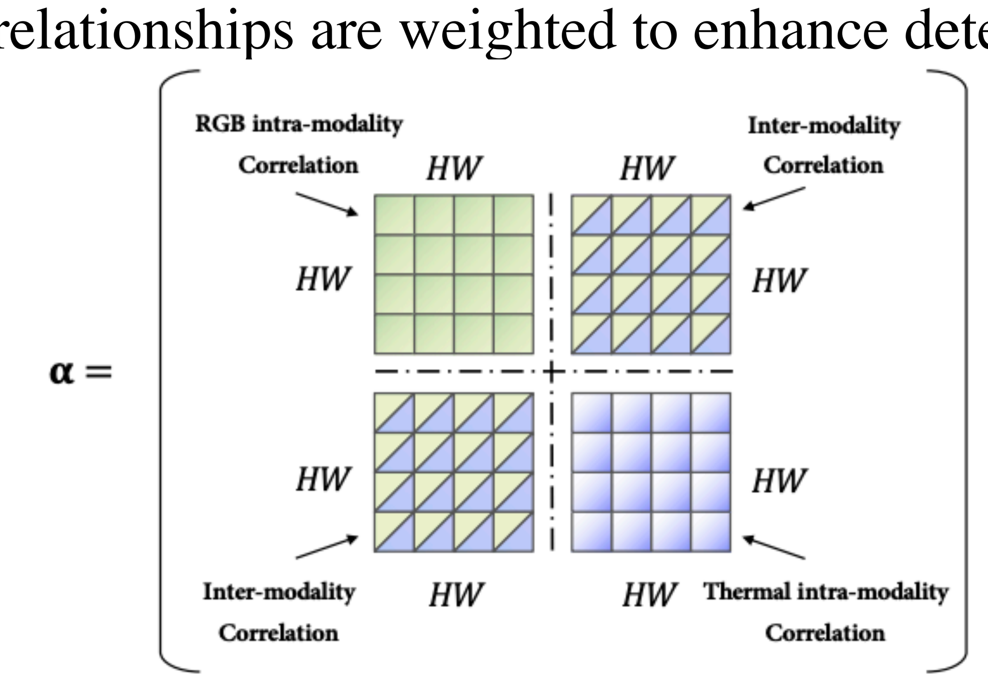
- Enhance detection accuracy by combining complementary information from multiple spectra
- Improve robustness in challenging environments, such as low-light or adverse weather
- Capture unique features across different modalities to detect a broader range of objects
- Enable more effective and adaptable object detection systems for diverse applications.
- Energy efficiency by selectively using specific spectra based on environmental needs, preserving computational resources.
- Facilitate long-range detection by employing spectra that penetrate environmental obscurants, improving accuracy for distant objects

Methodology



Proposed Approach: The YOLOv5 feature extraction network is redesigned as a two-stream backbone, to process both RGB and thermal images.
Cross-Modality Fusion Backbone (CFB): The two-stream backbone integrates the proposed CFT modules for efficient fusion and interaction between modalities
Input Transformation: RGB and thermal feature maps are flattened and reordered for Transformer input.

Transformer for Multispectral Fusion: The self-attention mechanism helps the network learn the relationships between RGB and thermal modalities. The correlation matrix (α) shows how intra- and inter-modality relationships are weighted to enhance detection performance.



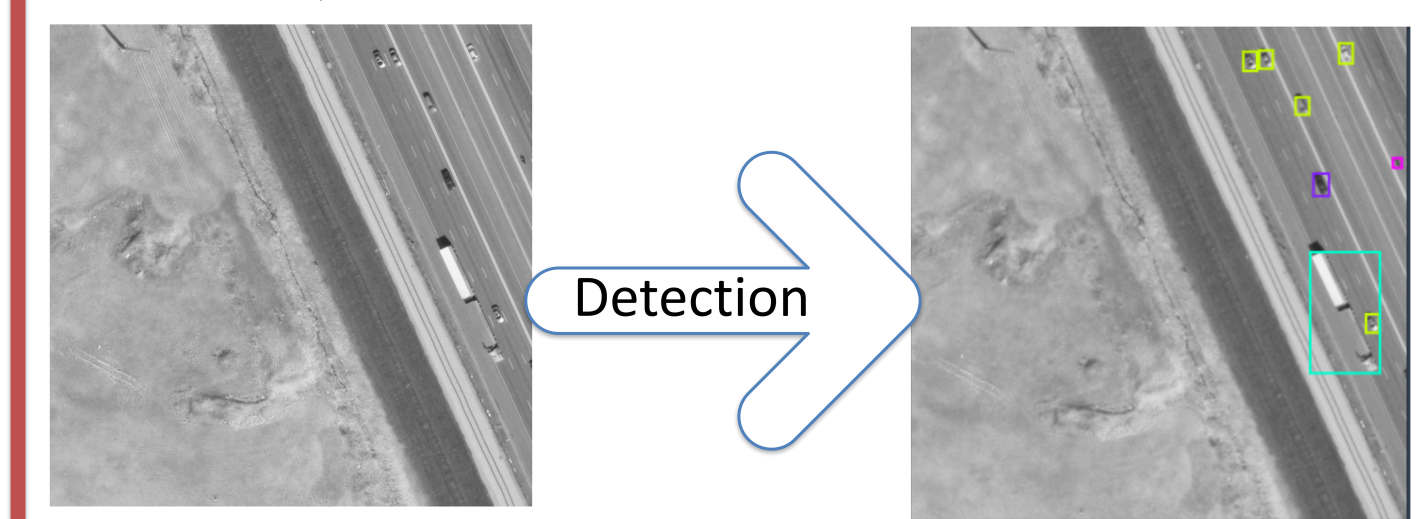
Efficiency and Speed improvement:

To reduce training time, we replaced SPP with SPPF (Simple Path Pooling Fast) for faster feature extraction .
To reduce computation further I used CrossConv instead of standard Conv layers in some parts of the architecture.which requitre less feature parameter compare to stadard convolution .

Result

Ablation Studies

On LLVIP, CFT shows gains of 0.7% in mAP50, 0.9% in mAP75, and 0.1% in mAP.
On Vedai, CFT shows gains of 0.7% in mAP50, 0.9% in mAP75, and 0.1% in mAP.



Dataset	Modality	Method	mAP50	mAP75	mAP
LLVIP	RGB+T	YOLOV5	95.8	68.4	60
		CFT	96.5	69.3	60.1
VEDAI	RGB+T	YOLOV5	70.4	47.7	46.8
		CFT	73.4	52.7	51.6

Conclusion

- **Proposed Approach**: Introduced Cross-Modality Fusion Transformer (CFT) to increase multispectral object detection by learning long-range dependencies and global contextual information.
- **Enhanced Backbone**: CFT modules are densely integrated within the backbone to maximize feature fusion and leverage complementary information between RGB and Thermal features.
- **Detector Combination**: Successfully applied CFT to popular detectors like YOLOv5, YOLOv3, and Faster R-CNN, increasing in both one-stage and two-stage detectors in multispectral object detection.