Multispectral Object Detection

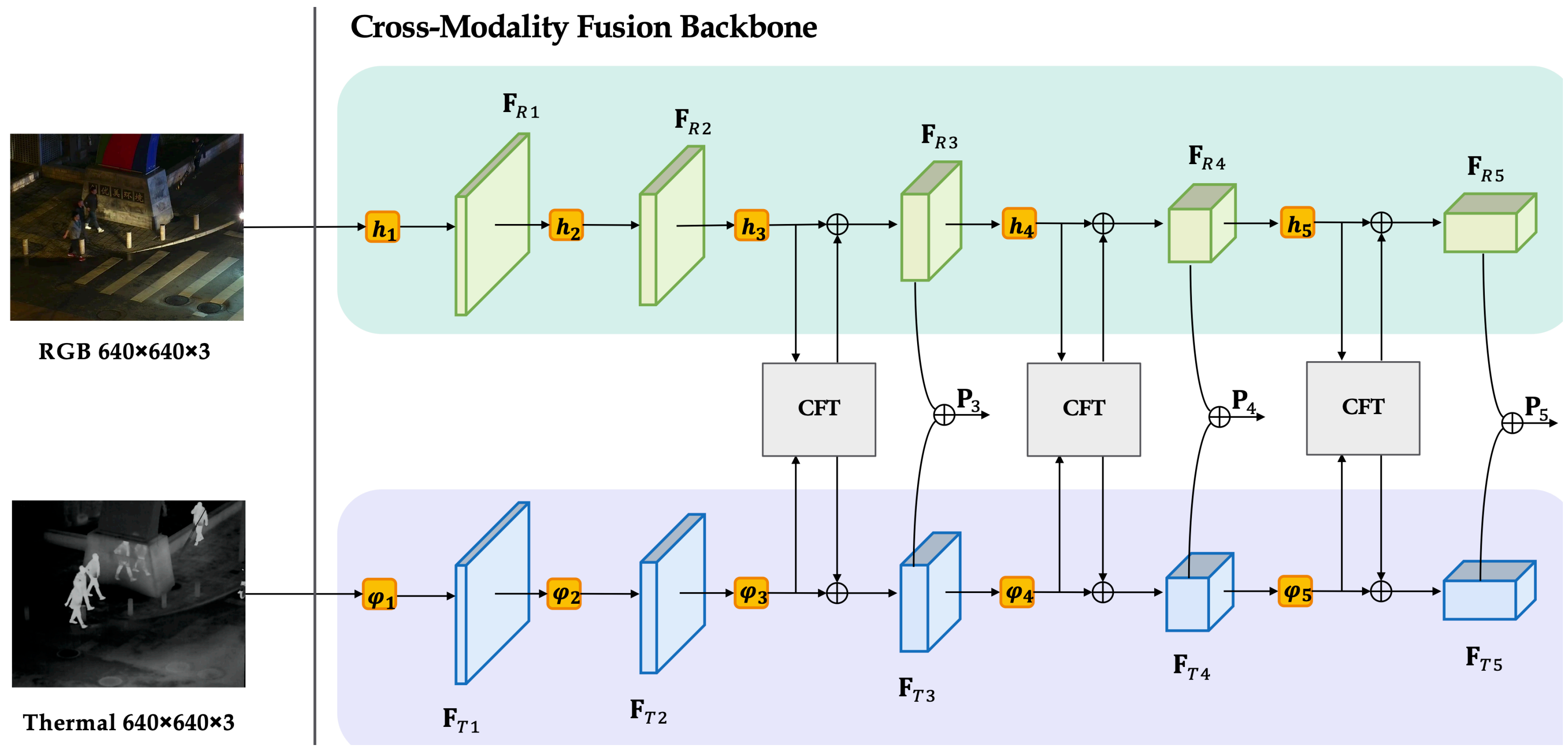
Bachhav Aryan Kishor

Indian Institute of Tecnhology Kanpur

Abstract

- Multispectral image pairs enhance 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.
- Extensive 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.

Methodology



To demonstrate the effectiveness of proposed **CFT** fusion module, we extend the framework of **YOLOv5**, to enable multispectral object detection. To be precise, we redesign the YOLOv5 feature extraction network as a twostream backbone, which is similar to GFD-SSD and embedded the CFT modules to facilitate modal fusion and modal interaction, named as Cross-Modality Fusion Backbone (**CFB**). An illustration of our Cross-Modality Fusion Backbone and CrossModality Fusion Transformer

Use of **SPPF** and **CrossConvolution** enhanced the accuracy and speed of Model compare to simple convolution.

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.

Result

Ablation Studies

On LLVIP, CFT shows gains of 1.7% in mAP50, 1.5% in mAP75, and 1.3% in mAP.

One-Stage and Two-Stage Detector Comparison: When integrated with YOLOv5, YOLOv3, and Faster R-CNN, CFT enhances detection performance: YOLOv5: CFT raises mAP50 by 5.7%, mAP75 by 3.5%, and mAP by 2.8%.

YOLOv3: CFT adds 4.0% in mAP50, 1.4% in mAP75, and 2.2% in mAP.Faster R-CNN: CFT improves mAP50 by 4.3%, mAP75 by 2.6%, and mAP by 2.1%.

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
			74.3	60.7	56

Conclusion

- **Proposed Approach**: Introduced Cross-Modality Fusion Transformer (CFT) to enhance multispectral object detection by learning long-range dependencies and integrating 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 modalities
- **Detector Integration**: Successfully applied CFT to popular detectors like YOLOv5, YOLOv3, and Faster R-CNN, enhancing both one-stage and two-stage detectors in multispectral object detection.
- **General Applicability**: CFT's simplicity and effectiveness suggest it could be adapted for other multispectral and multimodal tasks, including RGB-LiDAR, RGB-D, and stereo image applications..