Multimodal Object Detection by Channel Switching and Spatial Attention

Alec Trela, Sridevi Kaza, Sayan Mondal

01

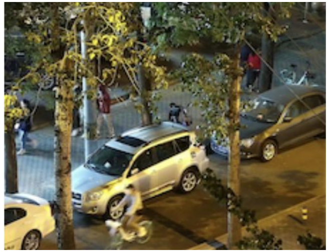
Introduction

Multimodal Data

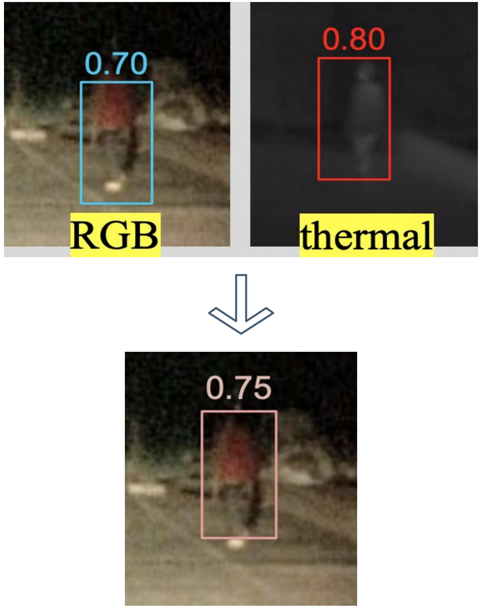
RGB data can capture details of an object with sufficient light but are insufficient in low-light conditions. IR data can ensure that the contour of the object can be provided in poor lighting conditions or obscured scenarios.

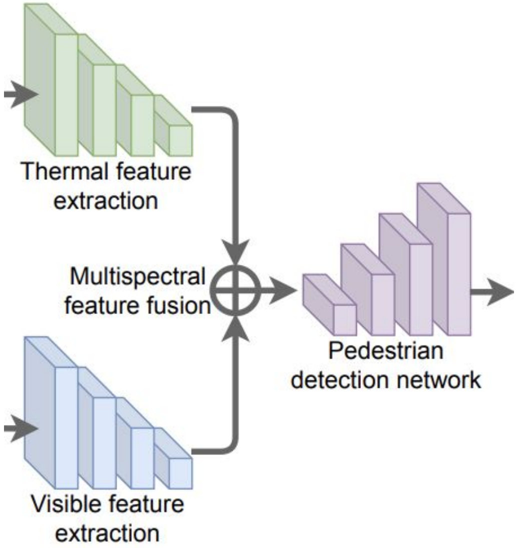
Multimodal Fusion Methods

RGB



IR





Early Fusion

Late Fusion

Mid Fusion

02

Task

Goals

Reaffirm Detection Metrics

Reimplement the model proposed in *Multimodal Object Detection by Channel Switching and Spatial Attention* and verify their results on the LLVIP dataset.

Explore Peripheral Areas Previously Unexplored

New data augmentations, loss propagation methods, and parameter sharing

Dataset: LLVIP

Aligned Pairs of RGB/IR Images

15488

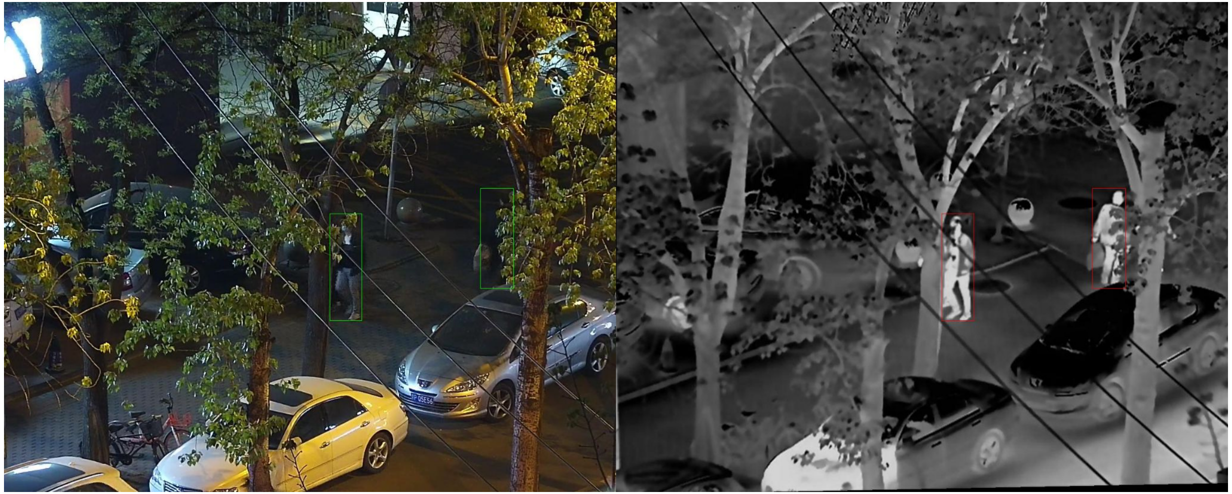
1920x1080 RGB Images & 1280 × 720 IR Images

High Quality

Images taken between 6pm-10pm in 26 Locations

Dimly Lit

Dataset: LLVIP



Aligned Pairs of RGB/IR Images

15488

1920x1080 RGB Images & 1280 × 720 IR Images

High Quality

Images taken between 6pm-10pm in 26 Locations

Dimly Lit

Dataset: LLVIP

Aligned Pairs of RGB/IR Images

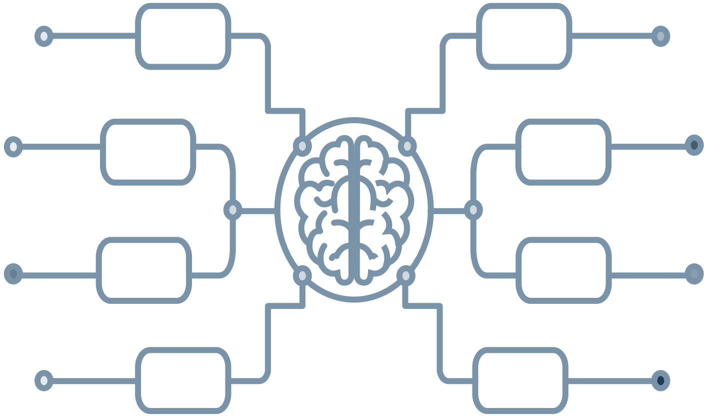
15488

1920x1080 RGB Images & 1280 × 720 IR Images

High Quality

Images taken between 6pm-10pm in 26 Locations

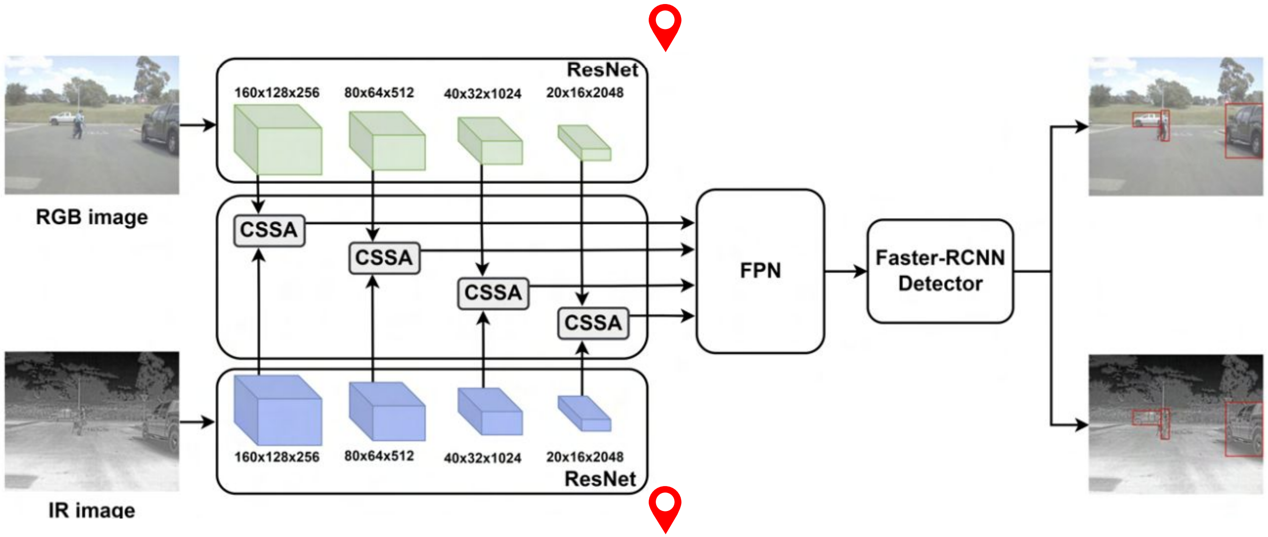
Dimly Lit



03

Approach & Methods

Model Overview



ResNet-50 Backbones

Feature Extraction

CSSA Block

Selecting Important Features

Feature Pyramid Network

Generate Scale Invariant-Feature Map

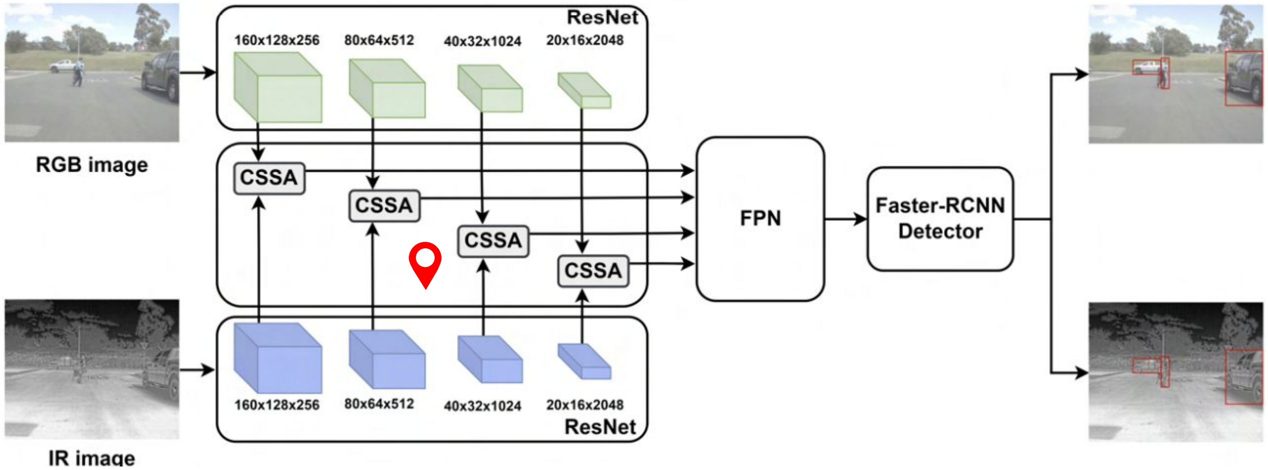
Faster-RCNN

Final Detection

Seen In:

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Model Overview



ResNet-50 Backbones

Feature Extraction

CSSA Block

Selecting Important Features

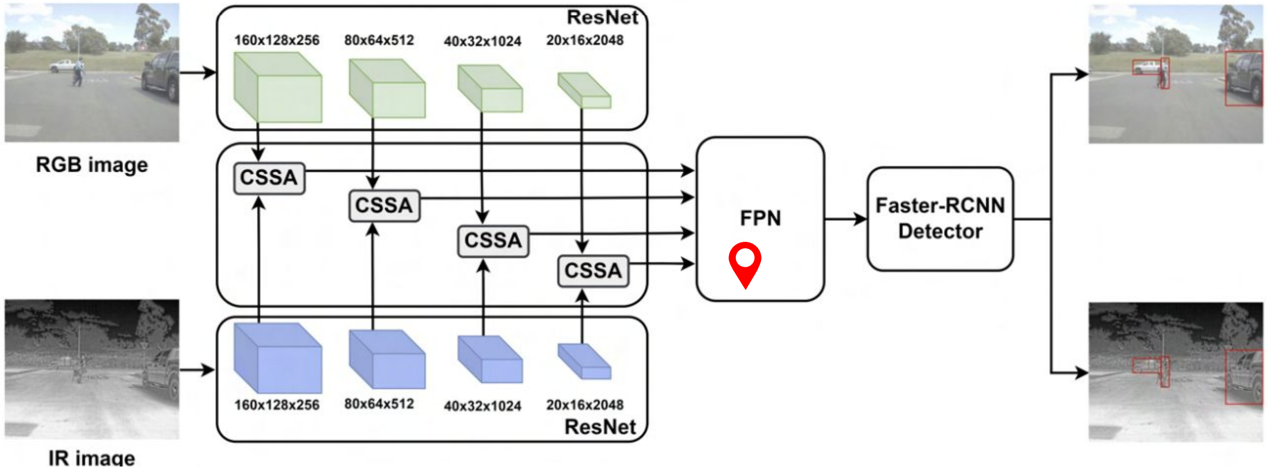
Feature Pyramid Network

Generate Scale Invariant-Feature Map

Faster-RCNN

Final Detection

Model Overview



ResNet-50 Backbones

Feature Extraction

CSSA Block

Selecting Important Features

Feature Pyramid Network

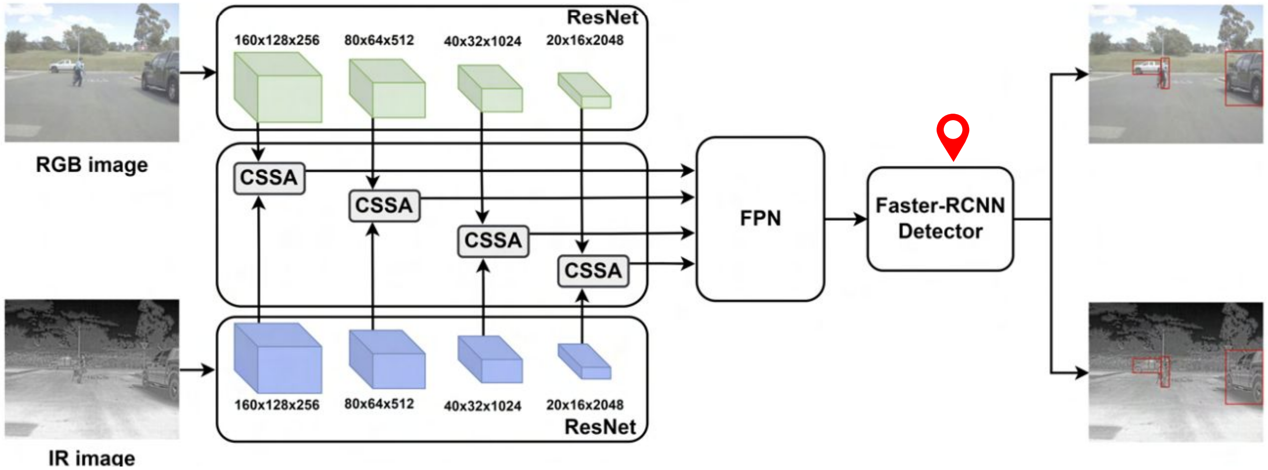
Generate Scale Invariant-Feature Map

Faster-RCNN

Final Detection

Seen In:
Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. CoRR abs/1612.03144, 2016.

Model Overview



ResNet-50 Backbones

Feature Extraction

CSSA Block

Selecting Important Features

Feature Pyramid Network

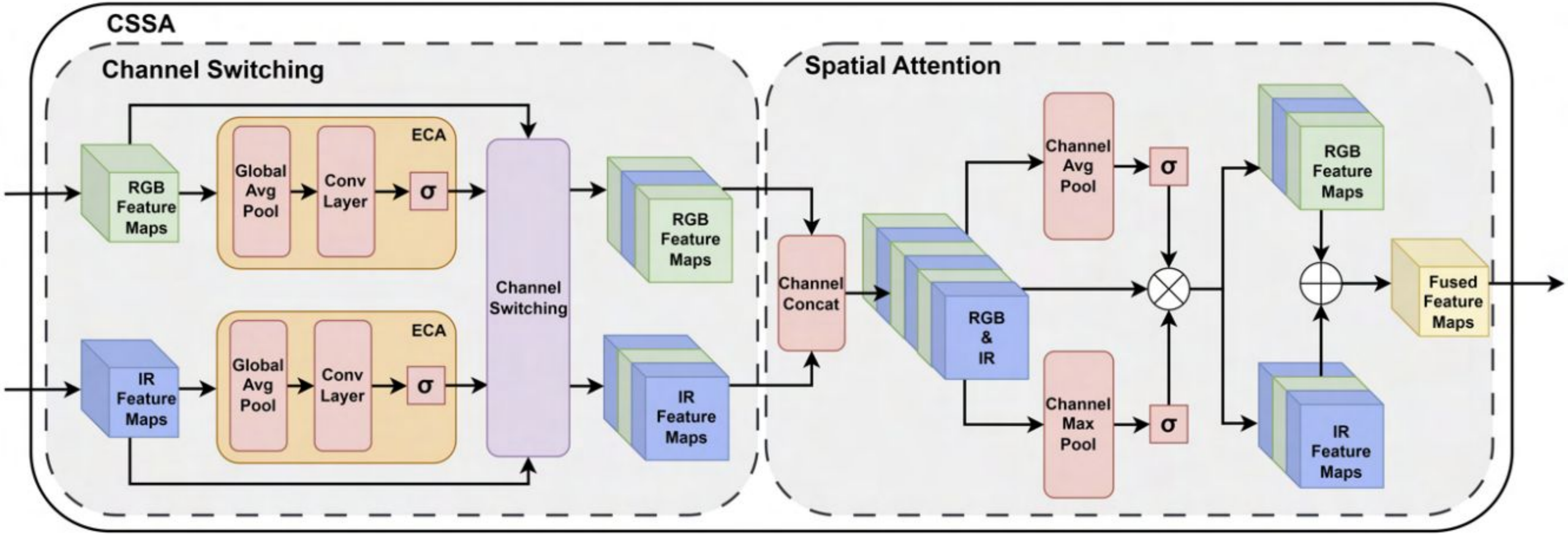
Generate Scale Invariant-Feature Map

Faster-RCNN

Final Detection

Seen In:
Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016

UNDERSTANDING CSSA



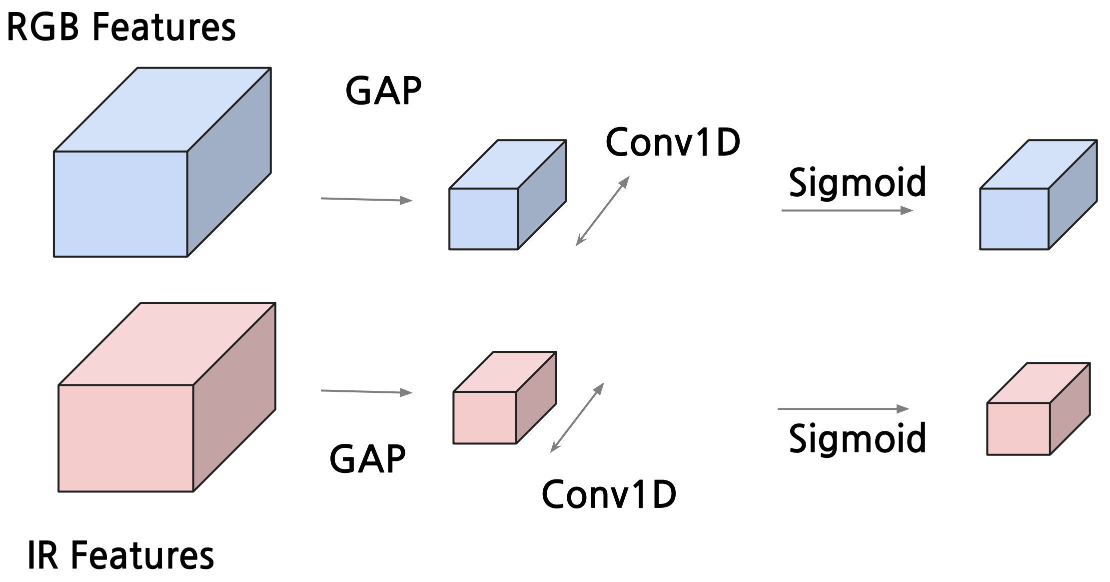
Channel Switching

Which of these channels is important?

Spatial Attention

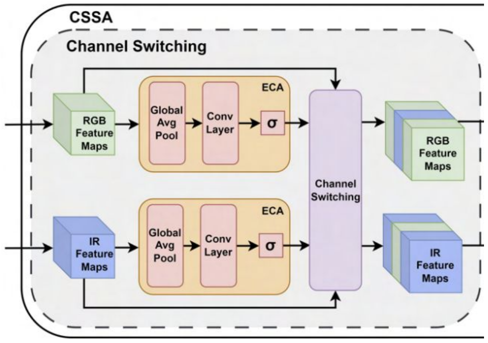
Which of these locations is important?

Channel Switching Block: ECA



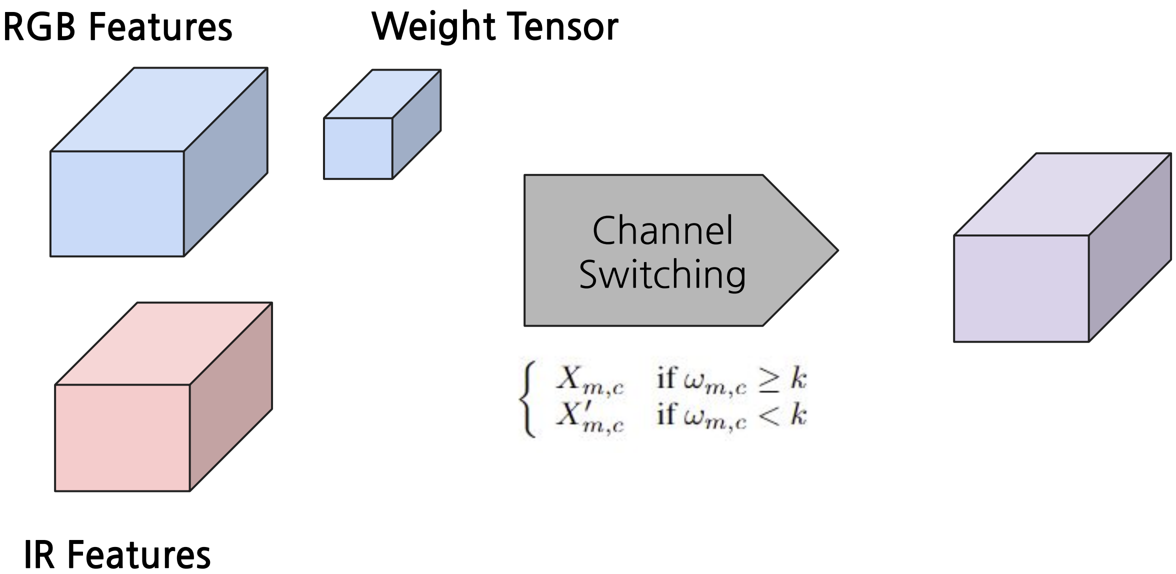
Channel Switching

Which of these channels is important?



Seen In:
Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks, 2020.

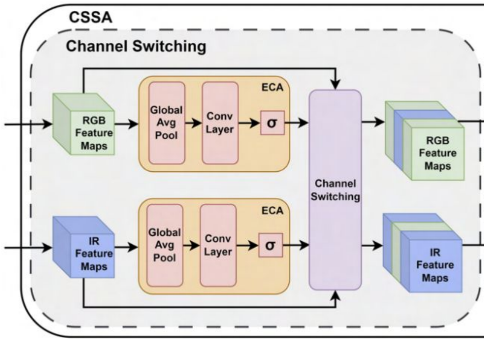
Channel Switching Block: Channel Switching



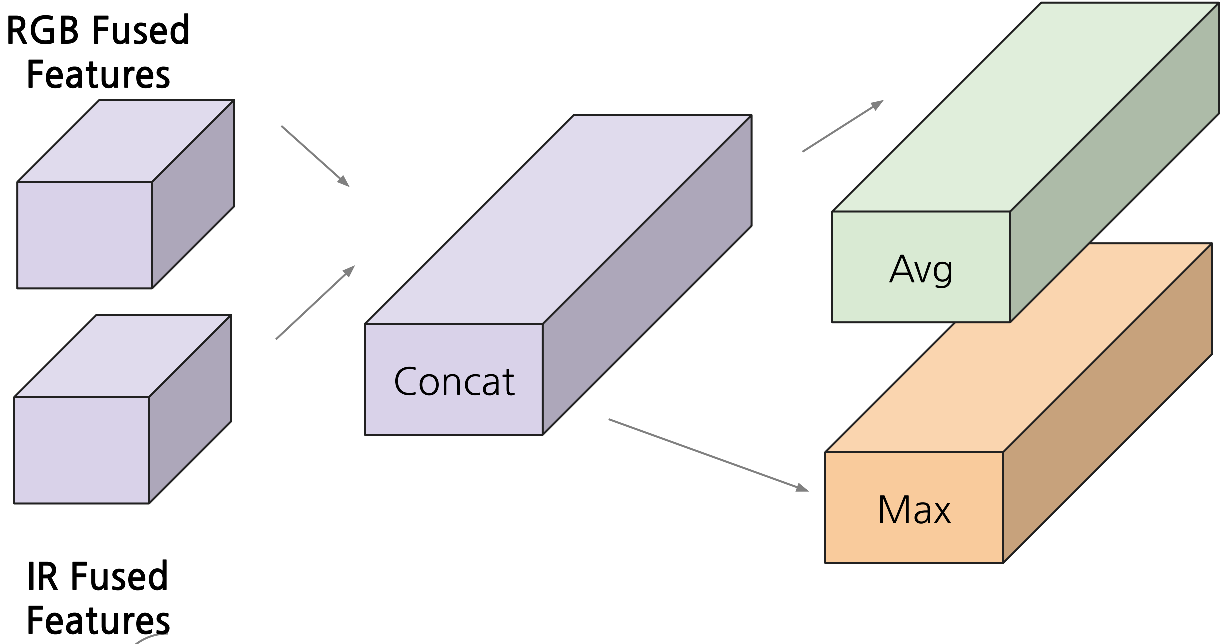
$$\begin{cases} X_{m,c} & \text{if } \omega_{m,c} \geq k \\ X'_{m,c} & \text{if } \omega_{m,c} < k \end{cases}$$

Channel Switching

Which of these channels is important?

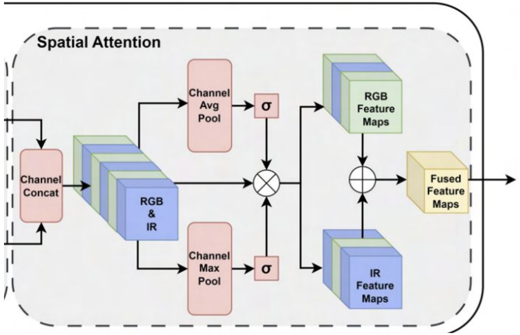


Spatial Attention

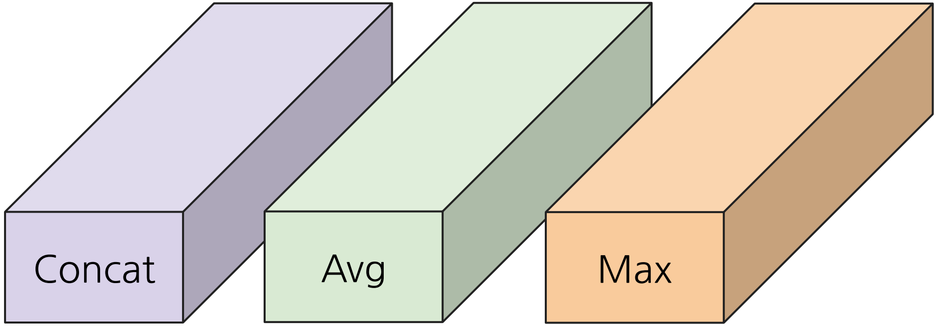


Spatial Attention

Which of these locations is important?

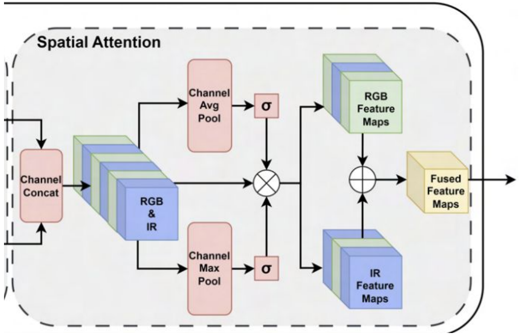


Spatial Attention

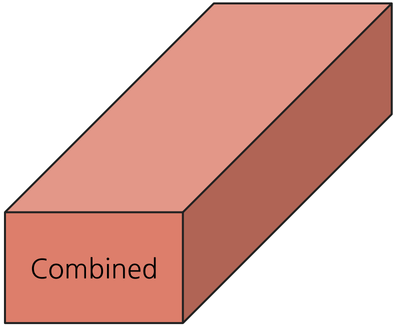


Spatial Attention

Which of these locations is important?

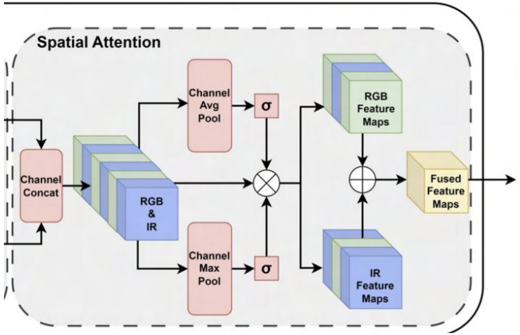


Spatial Attention

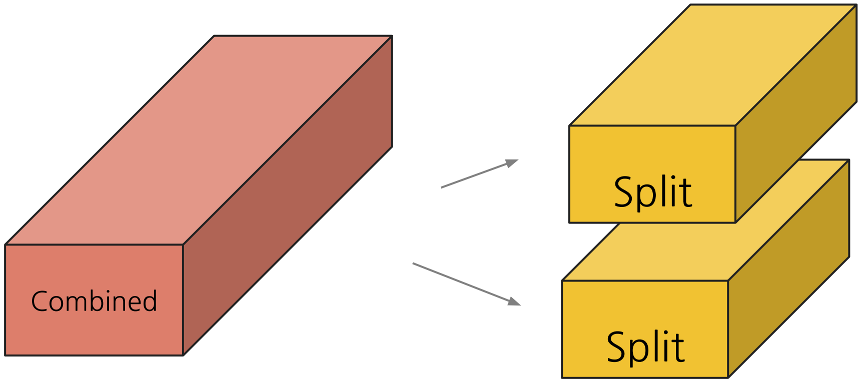


Spatial Attention

Which of these locations is important?

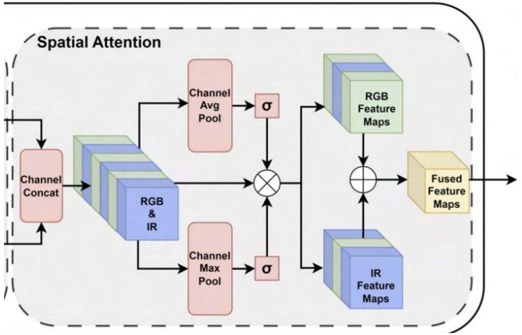


Spatial Attention

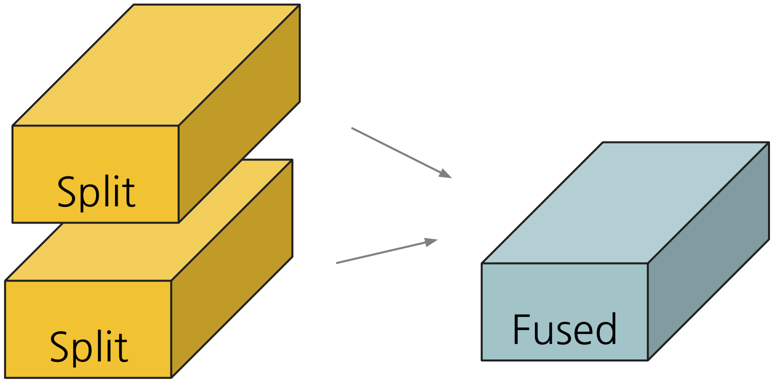


Spatial Attention

Which of these locations is important?

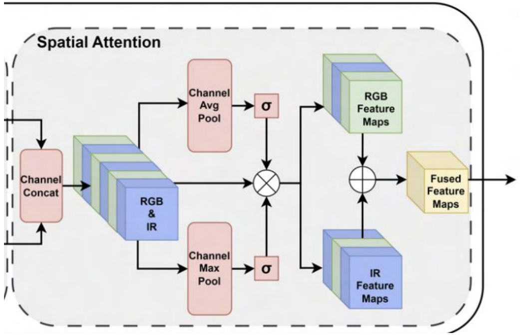


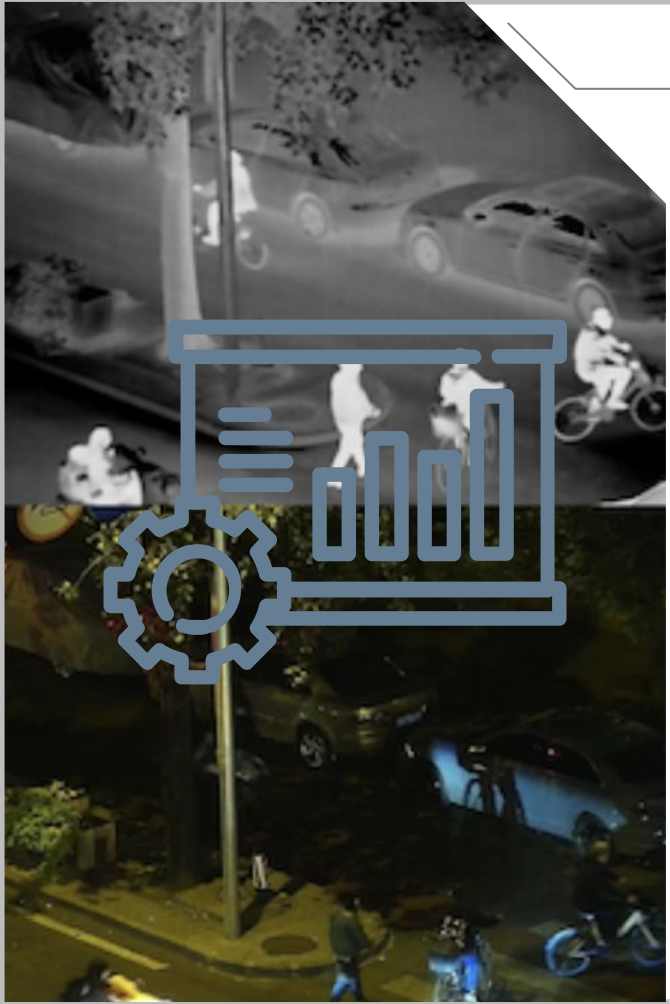
Spatial Attention



Spatial Attention

Which of these locations is important?





04

Results & Discussion

Overview of Results

	Model	Results From	Modality	AP50	AP75	mAP	Avg Inference Time (ms)
1	Faster-RCNN Baseline	Our Implementation	IR	92.1	53.4	51.9	55
2	Faster-RCNN Baseline	Our Implementation	RGB	76.7	28.1	35.4	56
3	CSSA Pipeline: Benchmark	Our Implementation	RGB + IR	89.34	48.2	48.5	—
4	CSSA Pipeline: Best	Our Implementation	RGB + IR	93.8	56.7	53.6	59
5	Faster-RCNN Baseline	Official Paper	IR	92.6	48.8	50.7	23
6	Faster-RCNN Baseline	Official Paper	RGB	88.8	45.7	47.5	23
7	CSSA Pipeline	Official Paper	RGB + IR	94.3	66.6	59.2	31

Hyperparameter Tuning

Channel Switching Threshold

Reducing the threshold for channel switching showed improved performance

Proposal Aspect Ratios

Modifying the number of proposal box aspect ratios had negligible effects

Detections per Image

Modifying the number of proposal box detections per image had negligible effects

RPN Loss Lambda

Modifying the regression loss scale factor had minimal effects at reasonable scale but negative effects for larger scales

Data Augmentations

Normalization

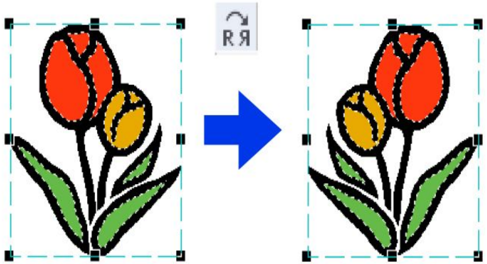
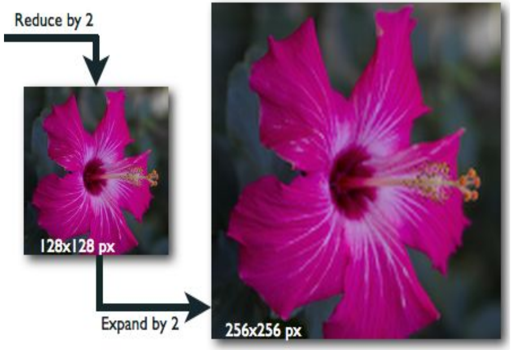
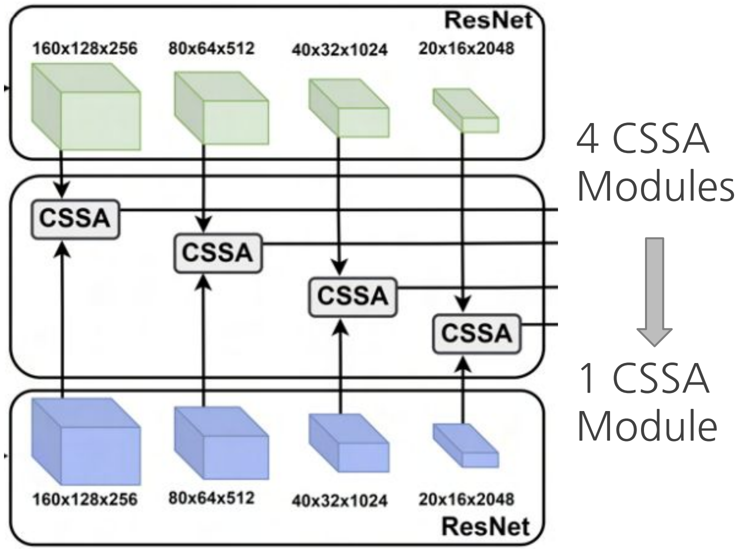


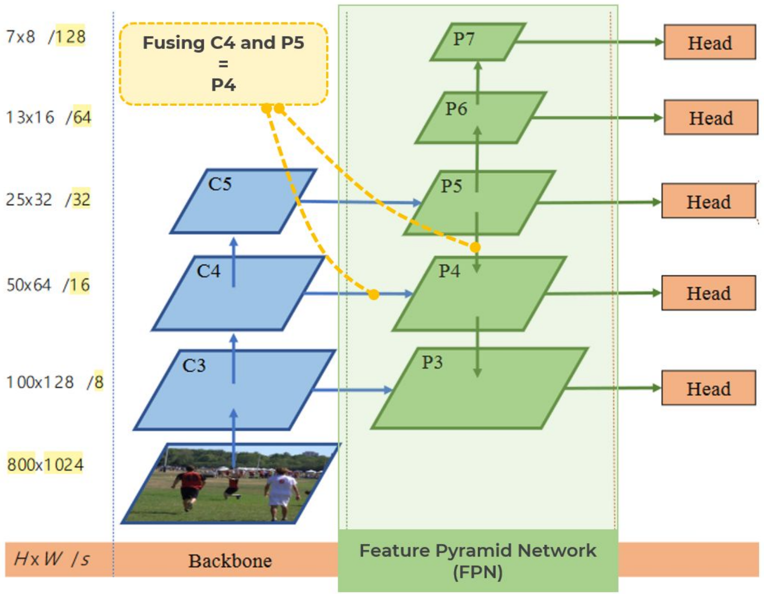
Image Flipping



Resizing

Implementation Modifications





Parameter Sharing

Extra FPN Block

04

Conclusion

Detection Results

Detection Results

Detection Results

Detection Results

Key Takeaways

- Our results showed an improvement in performance when using the proposed CSSA method to fuse RGB and IR data as opposed to unimodal detection methods
- More investigation needs to be done to verify the true efficacy of the CSSA block
- Advantages of parameter sharing
- Future work
    - Channel switching threshold as a learnable parameter
    - Testing on other datasets
    - Testing with different backbones and detector heads