JAN 4-8 WACV 2024
WAIKOLOA HAWAII

RGB-X Object Detection via Scene Specific Fusion Modules

Sri Aditya Deevi[1], Connor Lee[1*], Lu Gan[1*], Sushruth Nagesh[2], Gaurav Pandey[2], and Soon-Jo Chung[1]*

[1]California Institute of Technology

[2]Ford Motor Company

[*]Equal contribution

Introduction

- Object detection for Autonomous Vehicles (AVs) remains challenging in adverse weather conditions

CVPR 2022 BDD100K Challenges [1]

Ottawa Drivers Face Heavy Snow as 'Massive' Winter Storm Hits Canada and US [2]

[1] https://www.bdd100k.com/challenges/cvpr2022/
[2] https://ca.movies.yahoo.com/ottawa-drivers-face-heavy-snow-210100421.html

Ford Caltech

Introduction

- Multimodal sensor fusion provides an effective way to improve the **robustness** of detection models in various sensing conditions

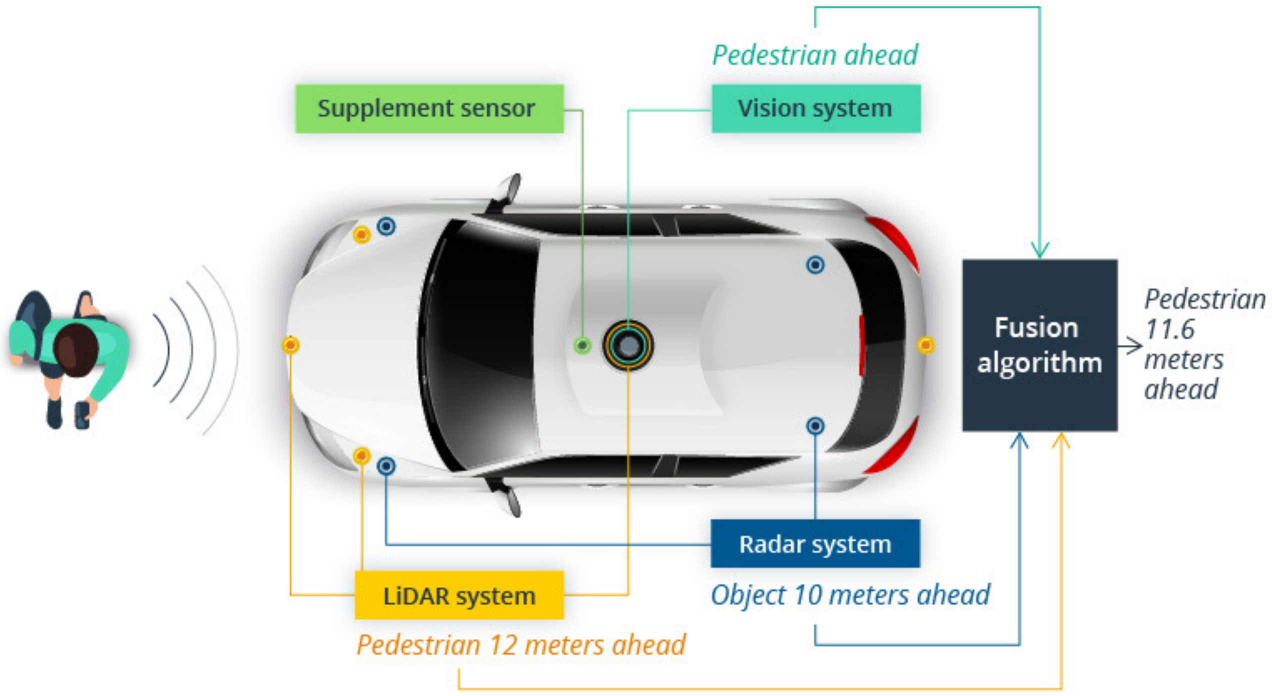
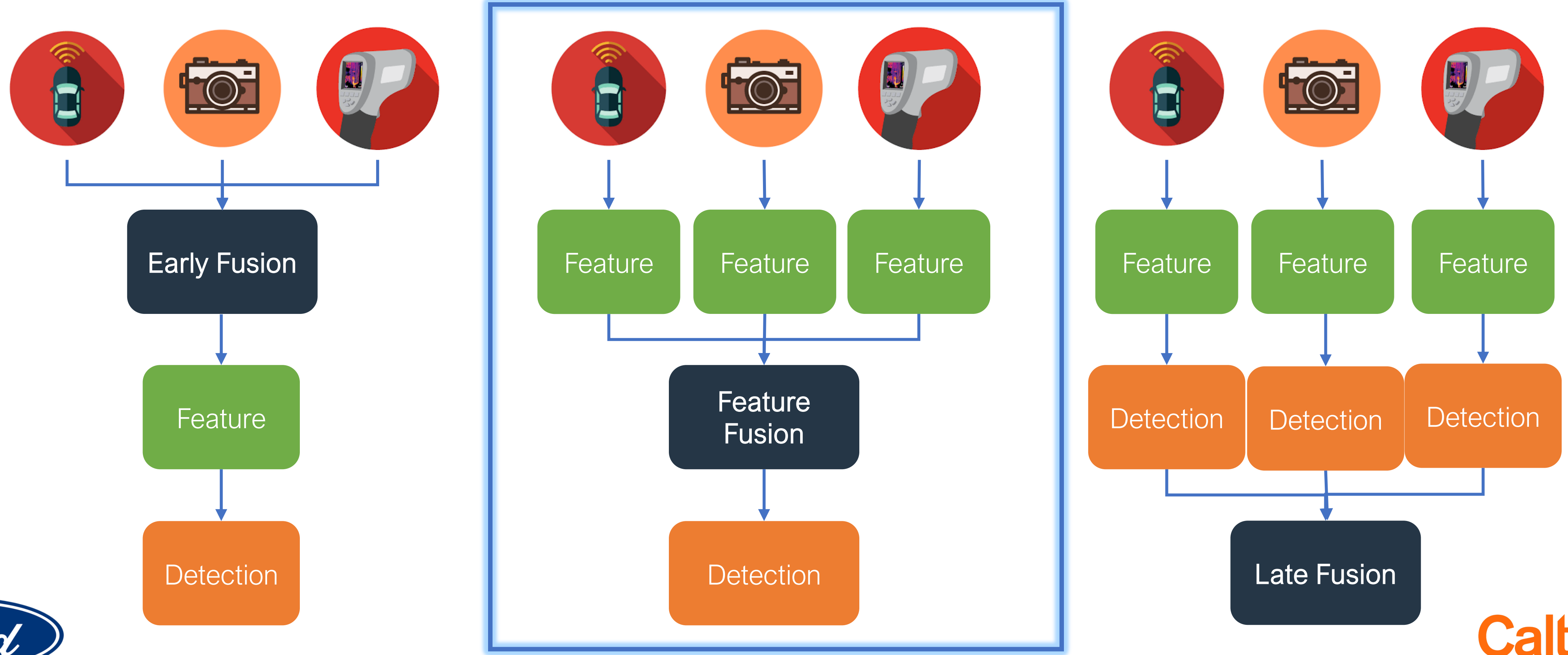


Image source: [Towards Data Science]

Caltech

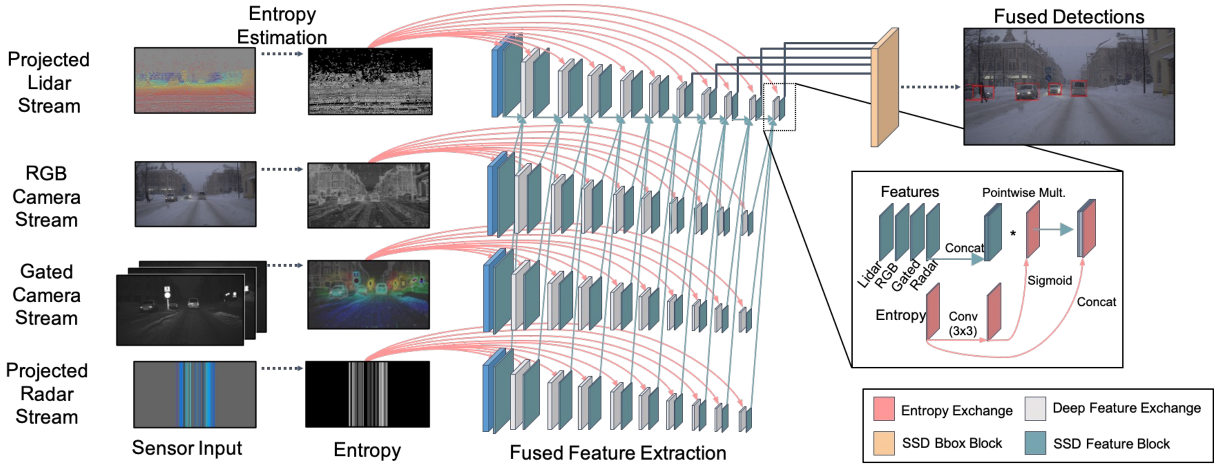
Introduction

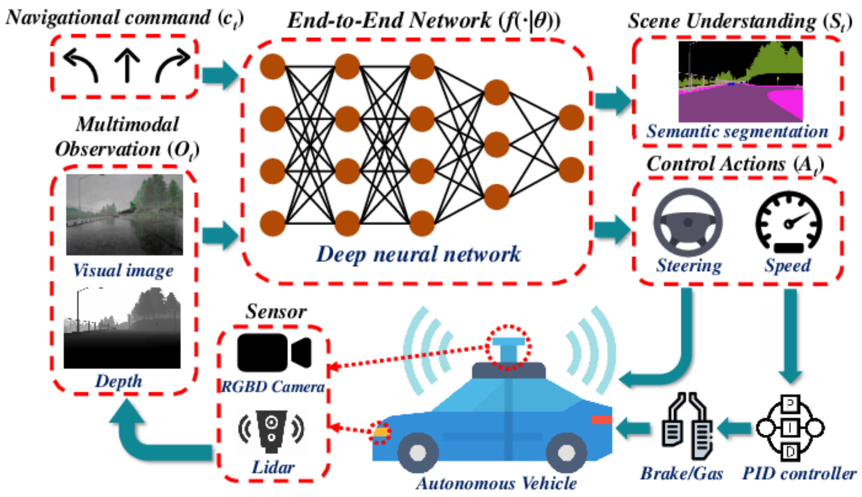
- Deep Sensor Fusion (DSF) on the **feature level** shows better performance



Motivation

- Existing DSF techniques for AVs require:
  - Large **coregistered, multimodal** datasets to train
  - Extensive training (fusion) time anytime a sensor component is **changed**



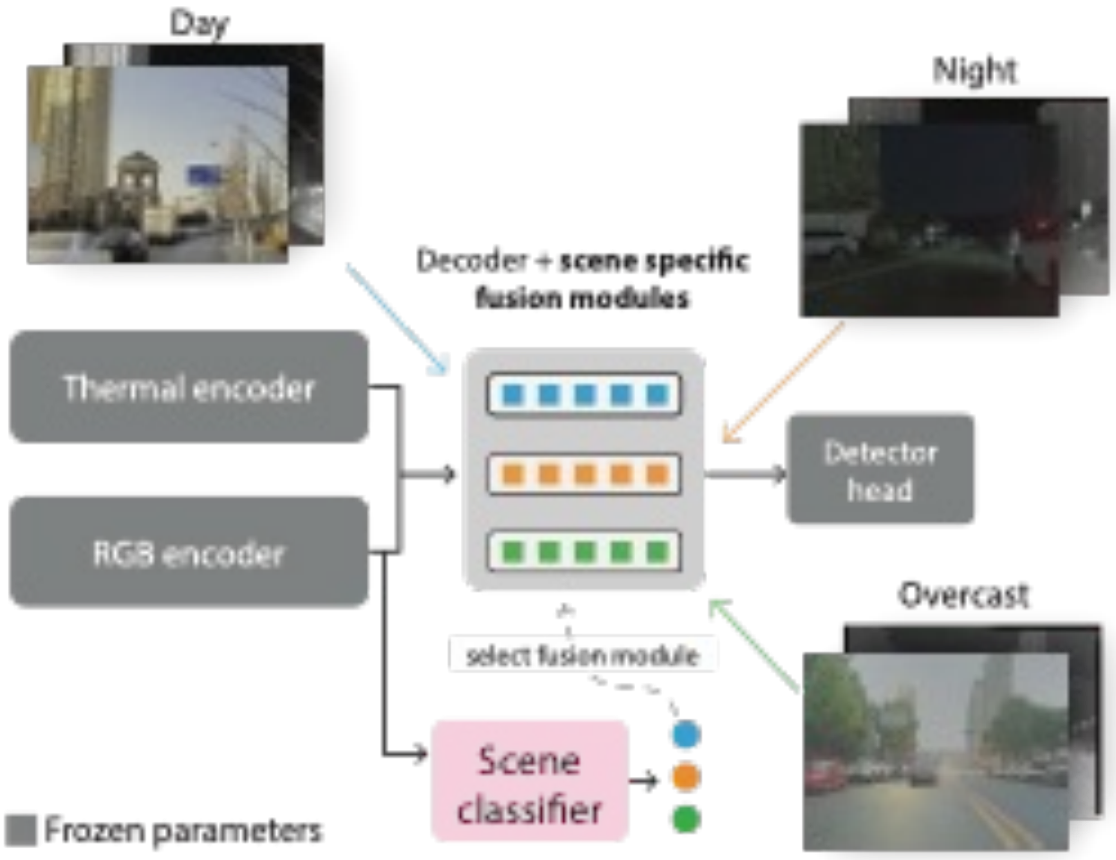


[1] Bijelic, et al. "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather." *CVPR*. 2020.
[2] Huang, et al. "Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding." *IEEE Sensors Journal* 21.10 (2020): 11781-11790.

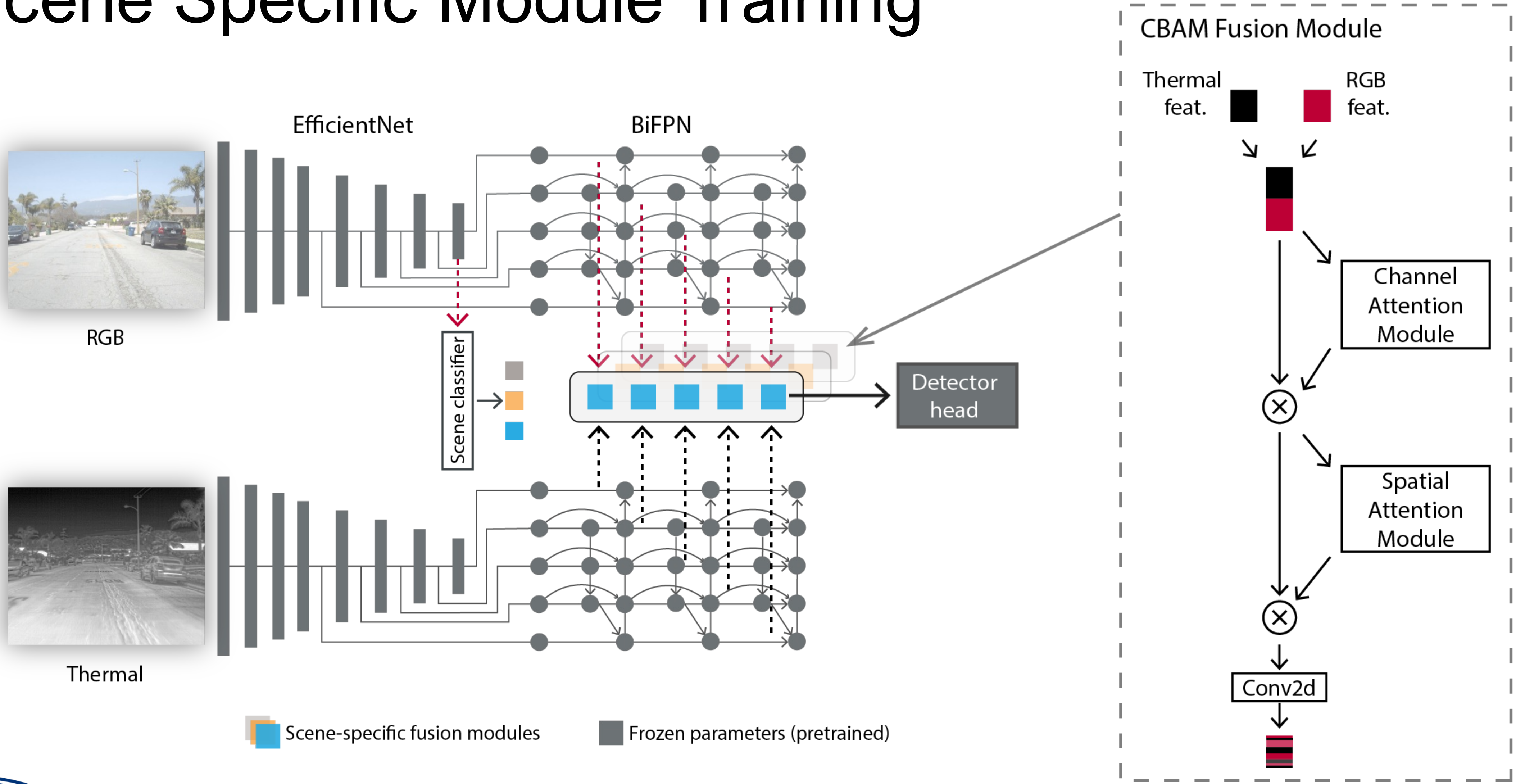
Caltech

Proposed Approach

- An efficient RGB-X fusion network that fuses **pretrained single-modal models** using **scene-specific fusion modules**
- Key advantages:
  - Superior performance over existing object detection methods on RGB-thermal and RGB-gated datasets
  - Comparable results with 75% less coregistered, multimodal training data

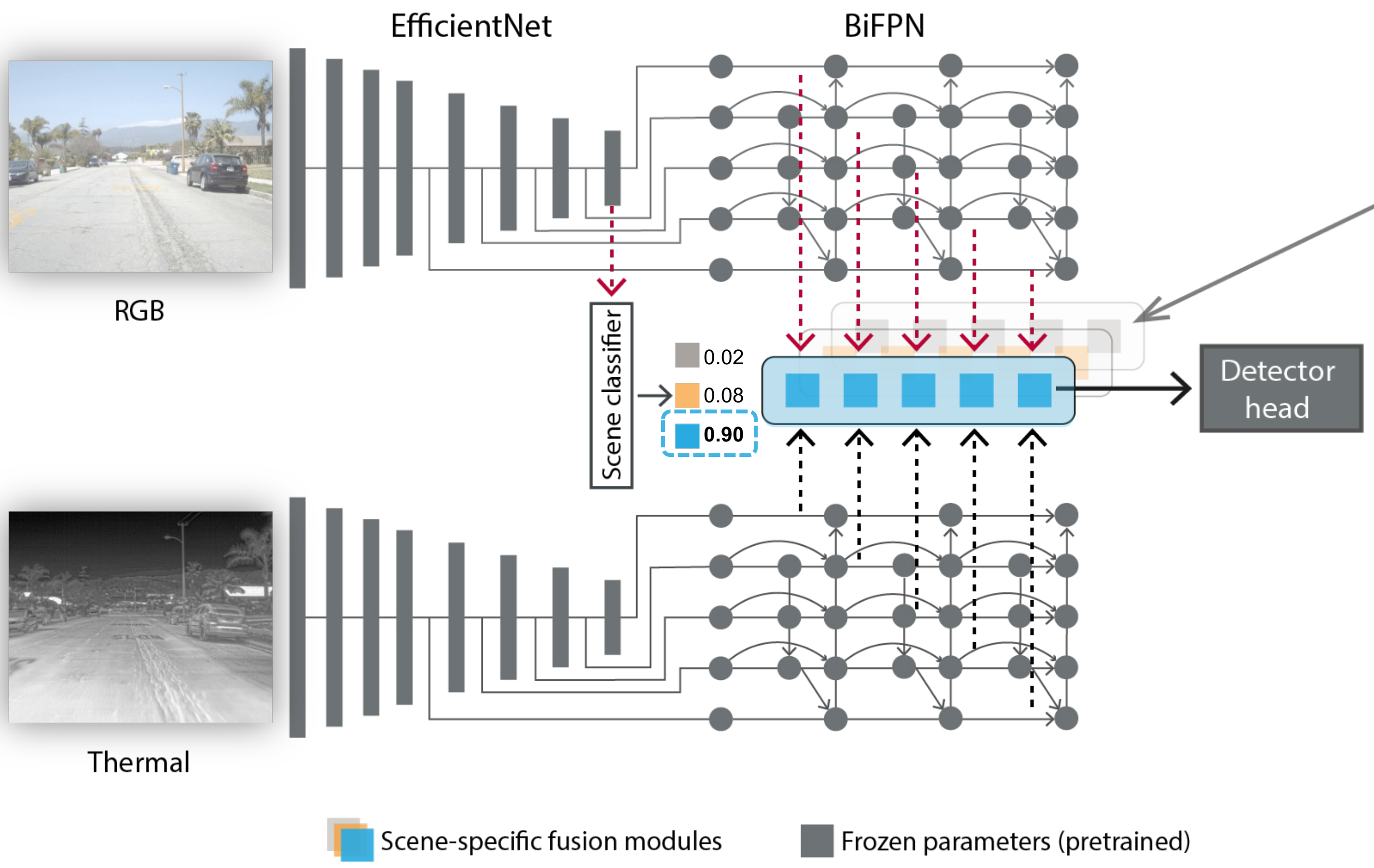


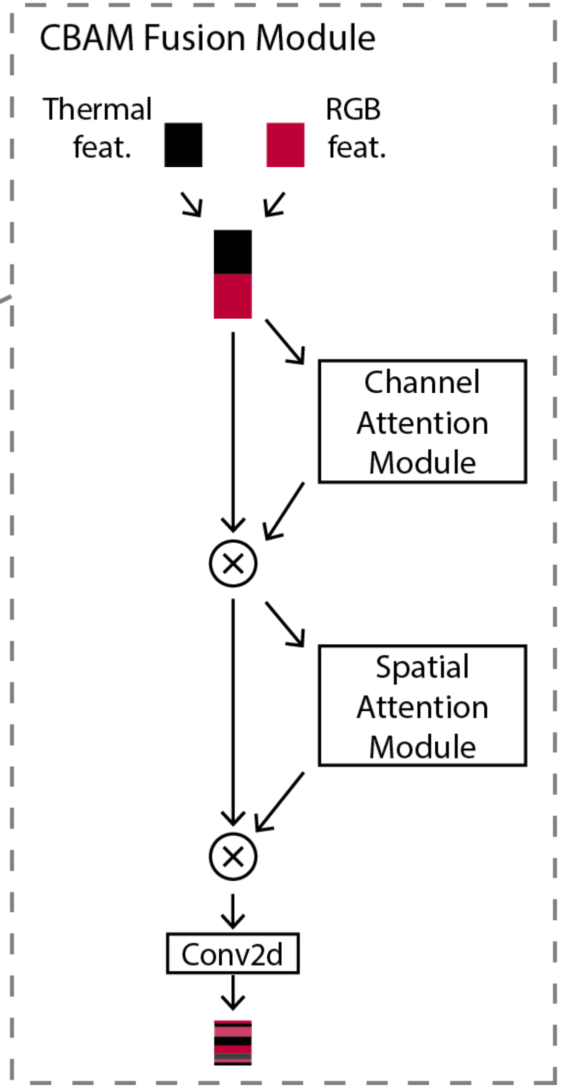
Scene Specific Module Training



[1] https://github.com/rwightman/efficientdet-pytorch
[2] Woo, et al. "CBAM: Convolutional block attention module." *ECCV*. 2018.

Scene Adaptive Model Inference





[1] https://github.com/rwightman/efficientdet-pytorch
[2] Woo, et al. "CBAM: Convolutional block attention module." *ECCV*. 2018.

Caltech

Results

- Improved RGB-T Object Detection Performance on FLIR Aligned Dataset [4]

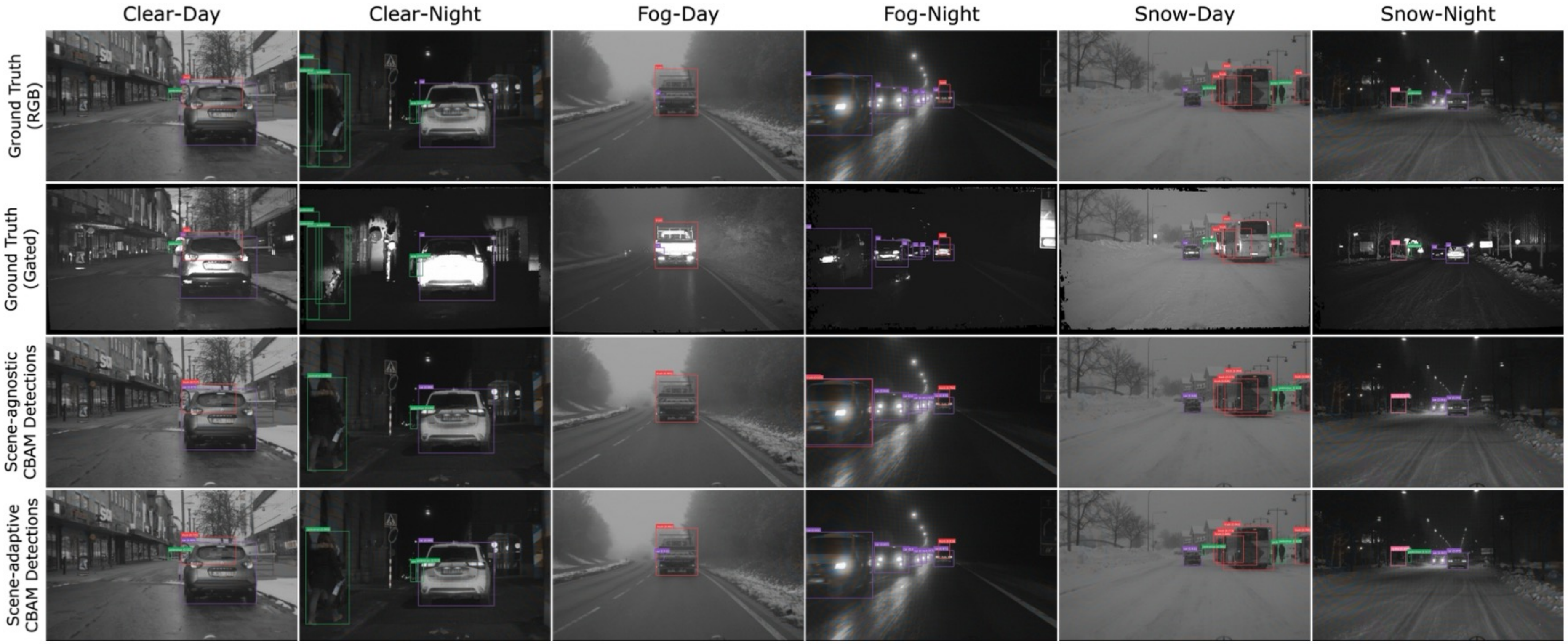
Method	Person	Bicycle	Car	mAP@0.5	mAP	Inference Speed (s)
RGB only	60.79	37.25	73.94	57.32	24.7	0.016
Thermal only	82.86	50.80	82.83	72.16	37.0	0.016
RetinaNet + MFPT [3]	78.1	65.0	87.3	76.80	–	0.050
CFT [2]	–	–	–	78.7	40.2	0.026
FasterRCNN + MFPT [3]	83.2	67.7	89.0	80.00	–	0.080
LRAF-Net [1]	–	–	–	80.50	42.8	–
Scene-agnostic CBAM (ours)	88.26	77.43	90.68	85.45	46.8	0.028
Scene-adaptive CBAM (ours)	88.92	78.61	90.94	86.16	47.1	0.032

[1] Fu, et al. "LRAF-Net: Long-Range Attention Fusion Network for Visible–Infrared Object Detection." *IEEE T-NNLS* (2023).
[2] Fang, et al. "Cross-modality fusion transformer for multispectral object detection." arXiv preprint arXiv:2111.00273 (2021).
[3] Zhu, et al. "Multi-Modal Feature Pyramid Transformer for RGB-Infrared Object Detection." *IEEE T-ITS* (2023).
[4] Zhang, et al. "Multispectral fusion for object detection with cyclic fuse-and-refine blocks." *ICIP*. 2020.

Caltech

Results

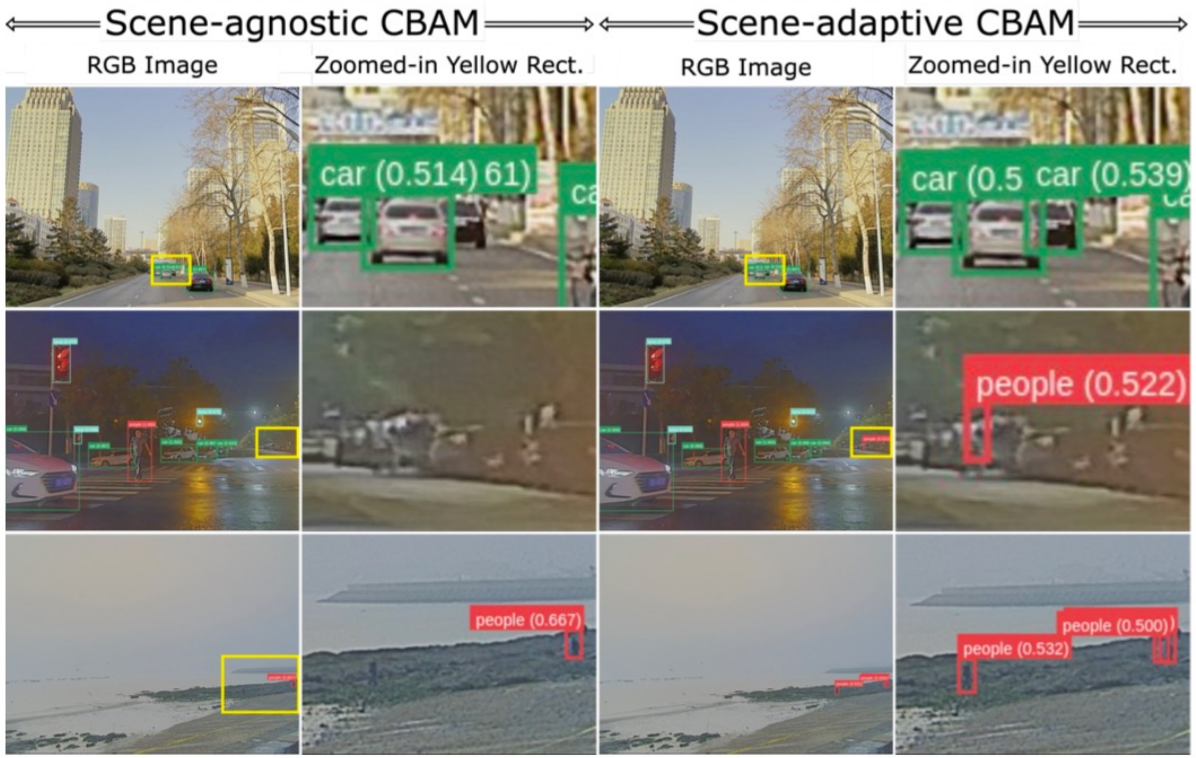
- Visualized RGB-Gated Detections on the STF Dataset [1]

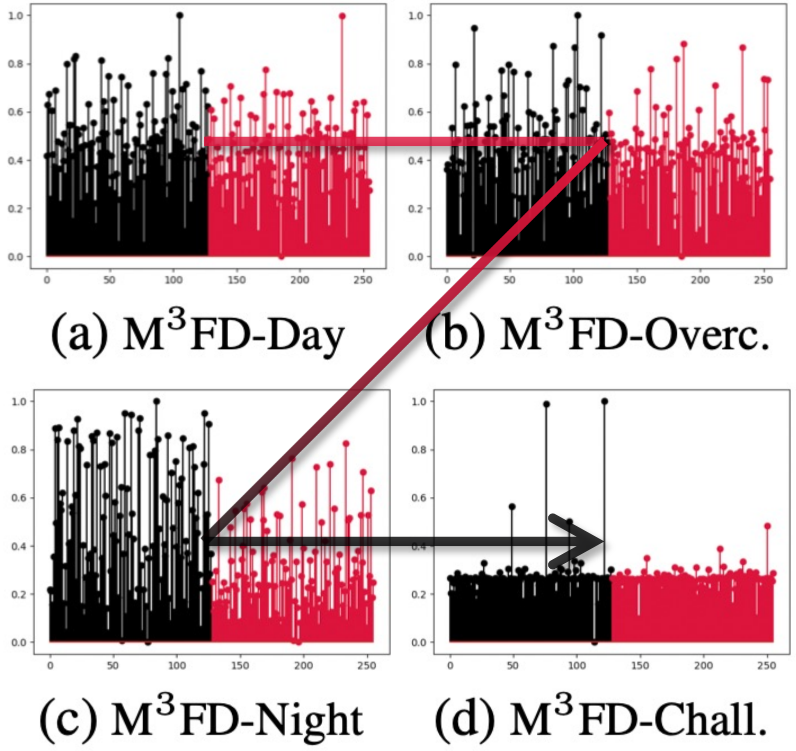


[1] Bijelic, et al. "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather." *CVPR*. 2020.

Results

- Visualized RGB-T Detections on the M^{3}FD Dataset [1]

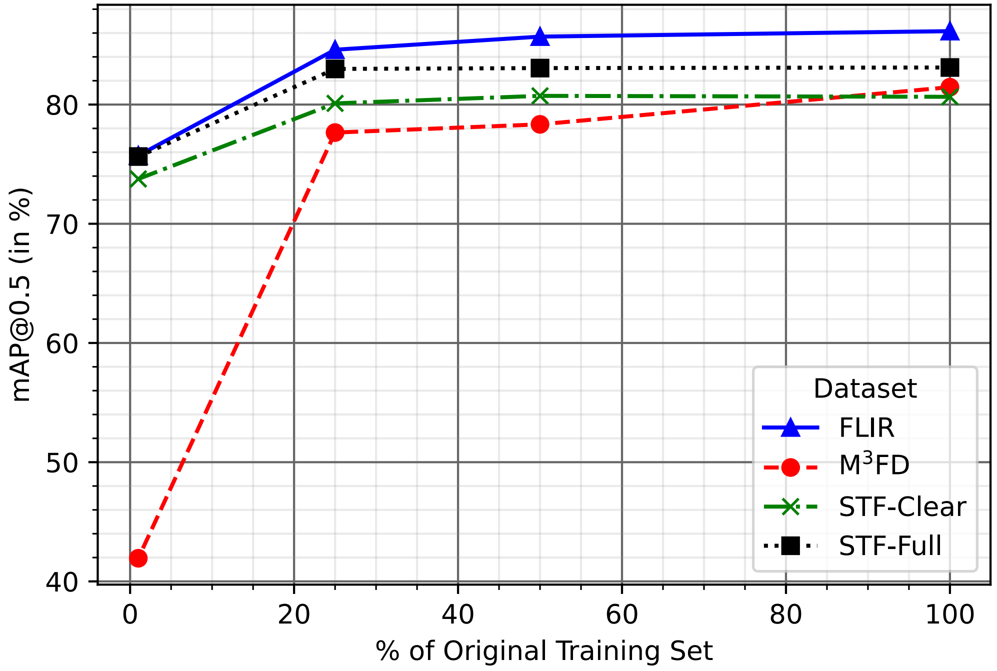




[1] Liu, et al. "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection." *CVPR*. 2022

Results

- Reduced Reliance on Multimodal Training Data for Fusion



Network Part	# Params
Encoders (RGB + X)	24.8 M
Decoders (RGB + X)	0.12 M
Detection Head	1.60 M
Fusion Modules	0.21 M
Total	26.7 M
Total Trainable (per scene)	0.21 M

Conclusions

- We presented an RGB-X object detection framework that fuses off-the-shelf networks using lightweight fusion modules.
- Our approach
  - Reduces the **dependence** on hard-to-obtain coregistered RGB-X datasets
  - Reduces fusion training time when sensors/pretrained networks are swapped out
  - Provides improved **adaptability** via scene-specific fusion modules

ganlu@caltech.edu

https://github.com/dsriaditya999/RGBXFusion

Ford

Caltech