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

Background

Problem: Existing deep sensor fusion techniques for AV require:

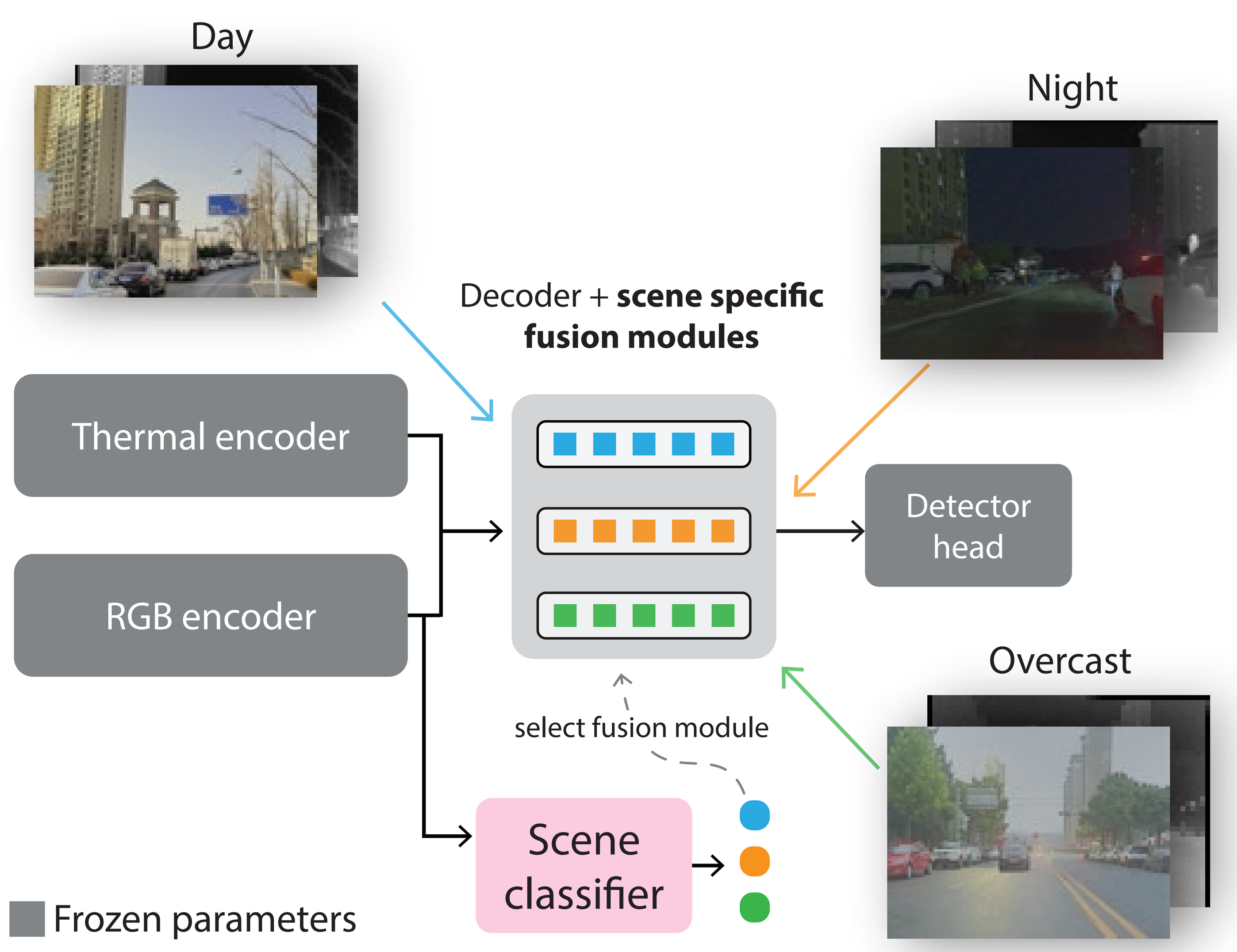
1. Large coregistered, multimodal datasets to train.
2. Extensive training (fusion) time anytime a sensor component is changed.

Proposed Solution: We introduce an efficient RGB-X fusion network that fuses pretrained single-modal models using scene-specific fusion modules. Key advantages include:

- Superior performance over existing object detection methods on RGB-thermal and RGB-gated datasets.
- Overall framework achieves comparable results with 75% less coregistered, multimodal training data.
- Enables creation of DSF models using small multimodal datasets.

Approach

Our modular RGB-X object detection network is built using pretrained single-modal detectors and fused using scene-specific CBAM [5] modules.

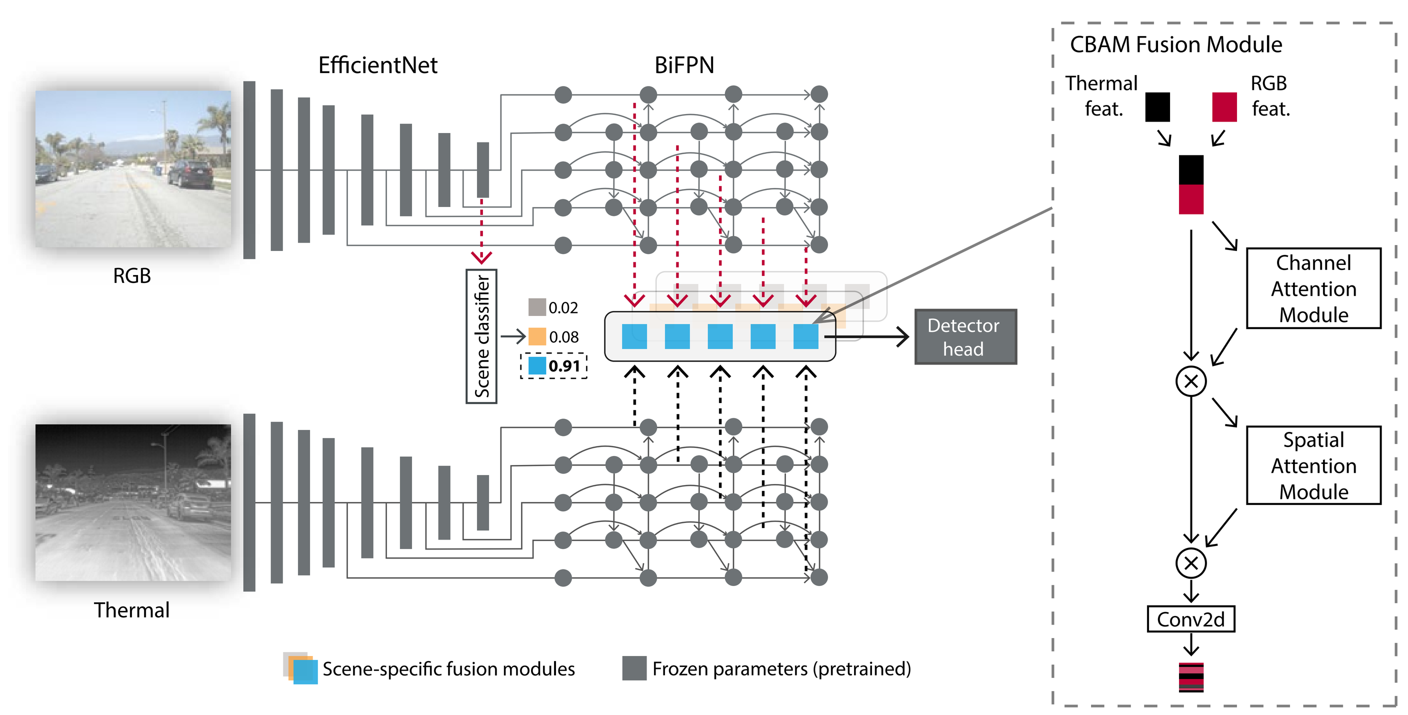


Model Inference:

1. Modality-specific network encoders take in a pair of RGB-X images.
2. A scene classifier operates on RGB encodings to determine the current scene category.
3. Scene-specific fusion modules are retrieved based on the predicted scene and used in decoder of the network.
4. Detector head performs inference on the fused features to generate bounding box outputs.

Model Training:

1. Pretrained object detectors of each modality are obtained or trained on single-modality datasets.
2. A scene-classifier is added to RGB encoder and trained on RGB data.
3. *(Fusion training)* Scene-specific fusion modules are trained per scene category on multimodal datasets while encoders, scene classifier, and object detection head (set to pretrained thermal weights) are frozen.



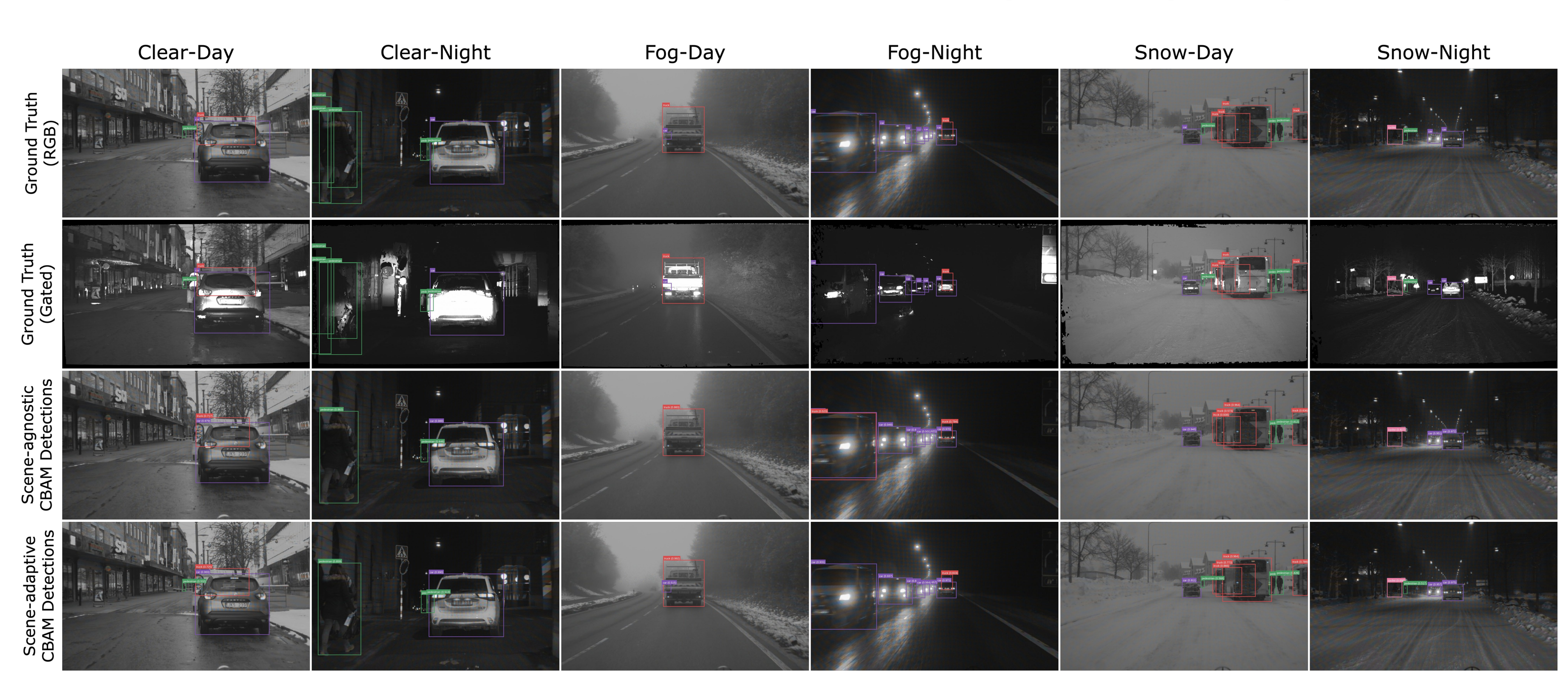
Results

Improved RGB-T Object Detection Performance:

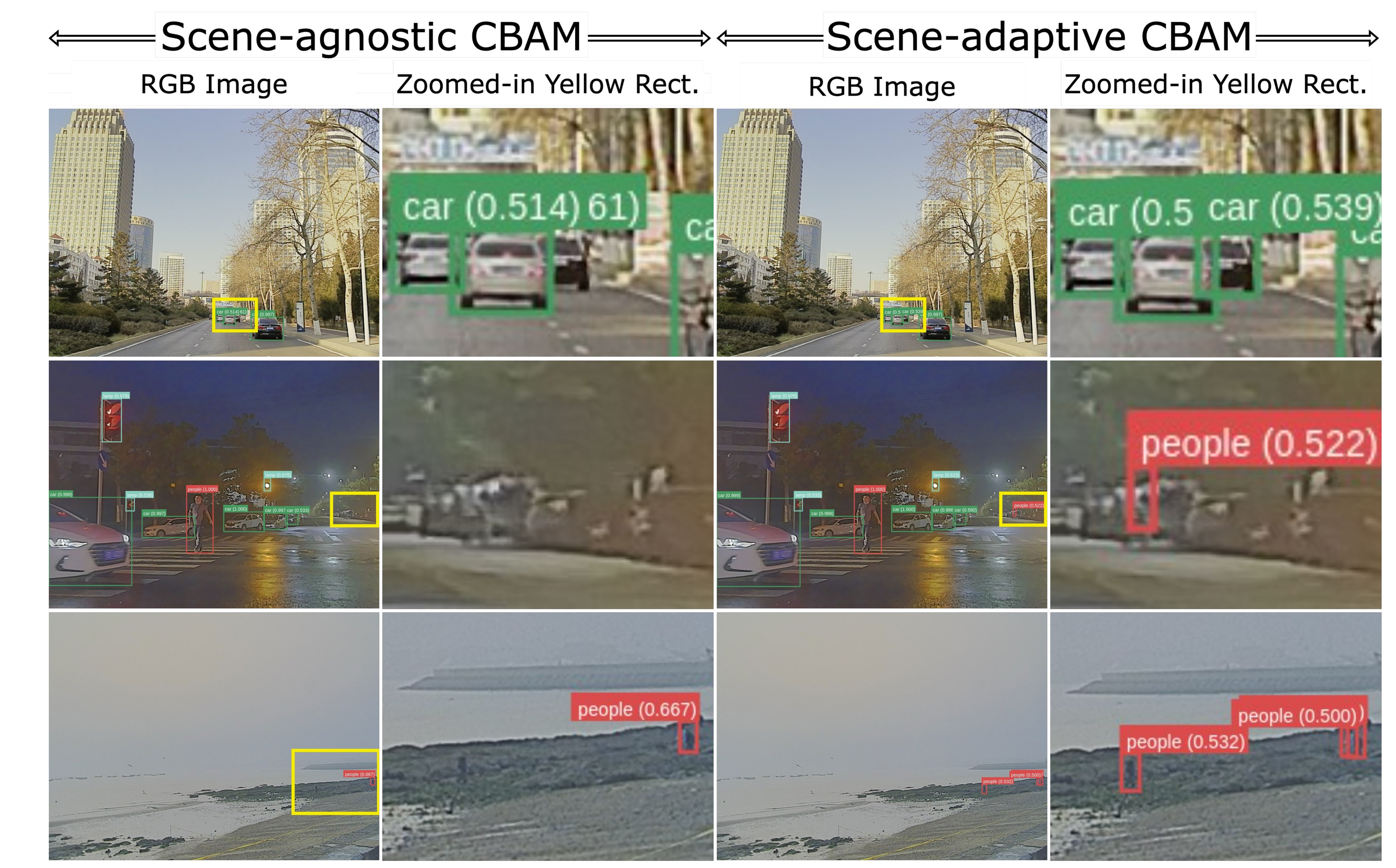
- Our algorithm performs better than existing RGB-T methods on the FLIR aligned object detection dataset.
- Scene-adaptive fusion modules provide a boost over scene-agnostic parameters.

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[6]	78.1	65.0	87.3	76.80	–	0.050
CFT [4]	–	–	–	78.7	40.2	0.026
FasterRCNN + MFPT[6]	83.2	67.7	89.0	80.00	–	0.080
LRAF-Net[2]	–	–	–	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

Visualized RGB-Gated Detections on the Seeing Through Fog Dataset [1]:

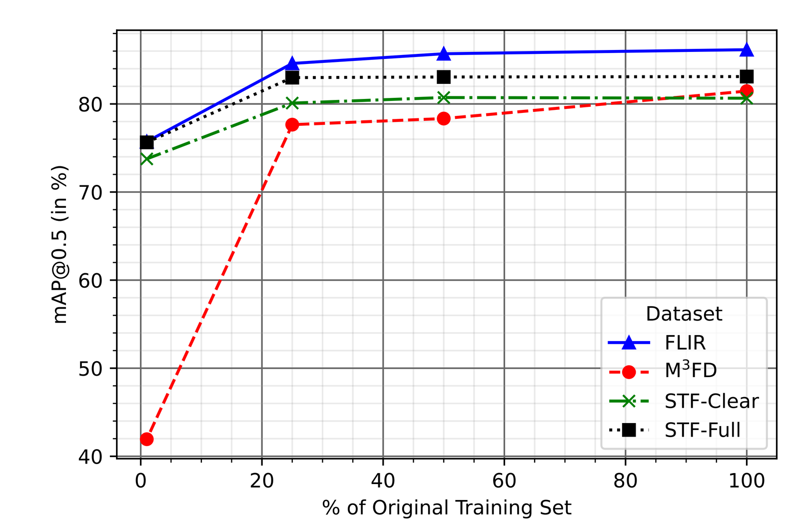


Visualized RGB-T Detections on the M^{3}FD Dataset [3]:



Reduced Reliance on Multimodal Training Data for Fusion:

- Our approach heavily constrains the training process.
- Similar performance can be maintained even with 75% less fusion training data.



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 a 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.

References

[1] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *CVPR*, June 2020. [2] Haolong Fu, Shixun Wang, Puhong Duan, Changyan Xiao, Renwei Dian, Shutao Li, and Zhiyong Li. Lraf-net: Long-range attention fusion network for visible–infrared object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. [3] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, pages 5802–5811, 2022. [4] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021. [5] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. [6] Yaohui Zhu, Xiaoyu Sun, Miao Wang, and Hua Huang. Multi-modal feature pyramid transformer for rgb-infrared object detection. *IEEE Transactions on Intelligent Transportation Systems*, 2023.