Multimodal Object Detection via Probabilistic Ensembling

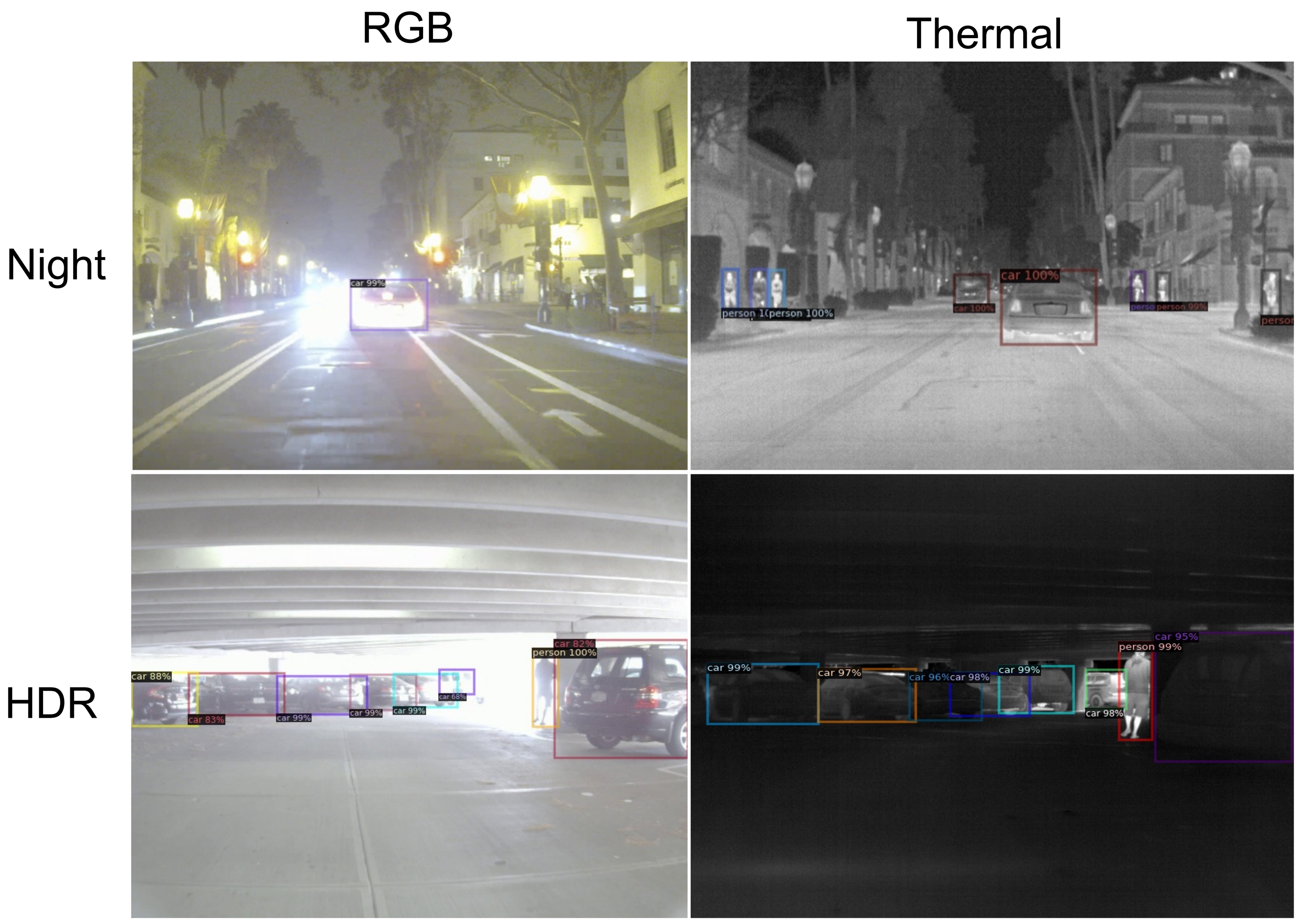
ECCV TEL AVIV 2022

bit.ly/RGBT-detection

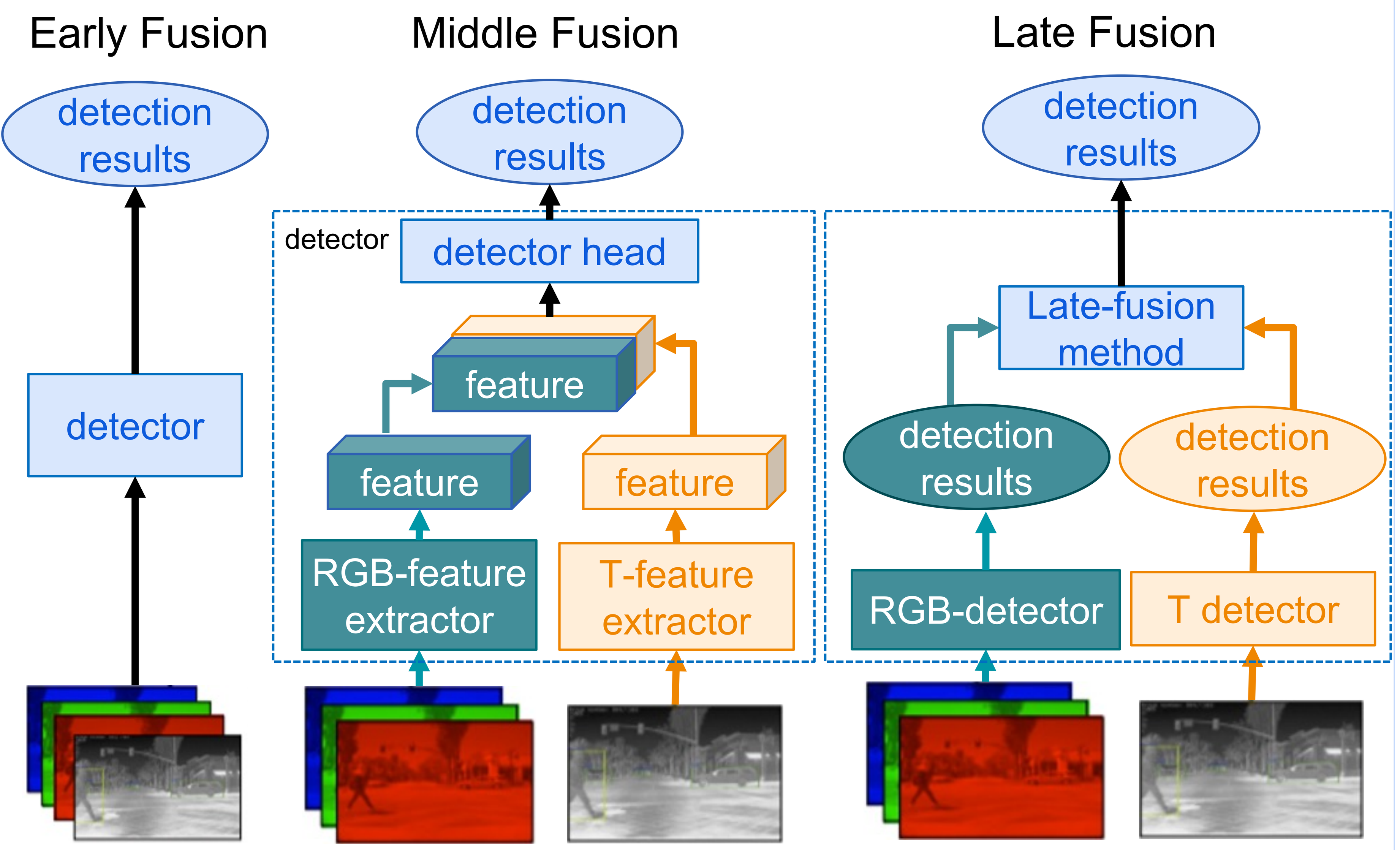
Yi-Ting Chen* Jinghao Shi* Zelin Ye* Christoph Mertz Deva Ramanan# Shu Kong#

Motivation

- RGB and thermal modalities are complementary
- Different scenarios

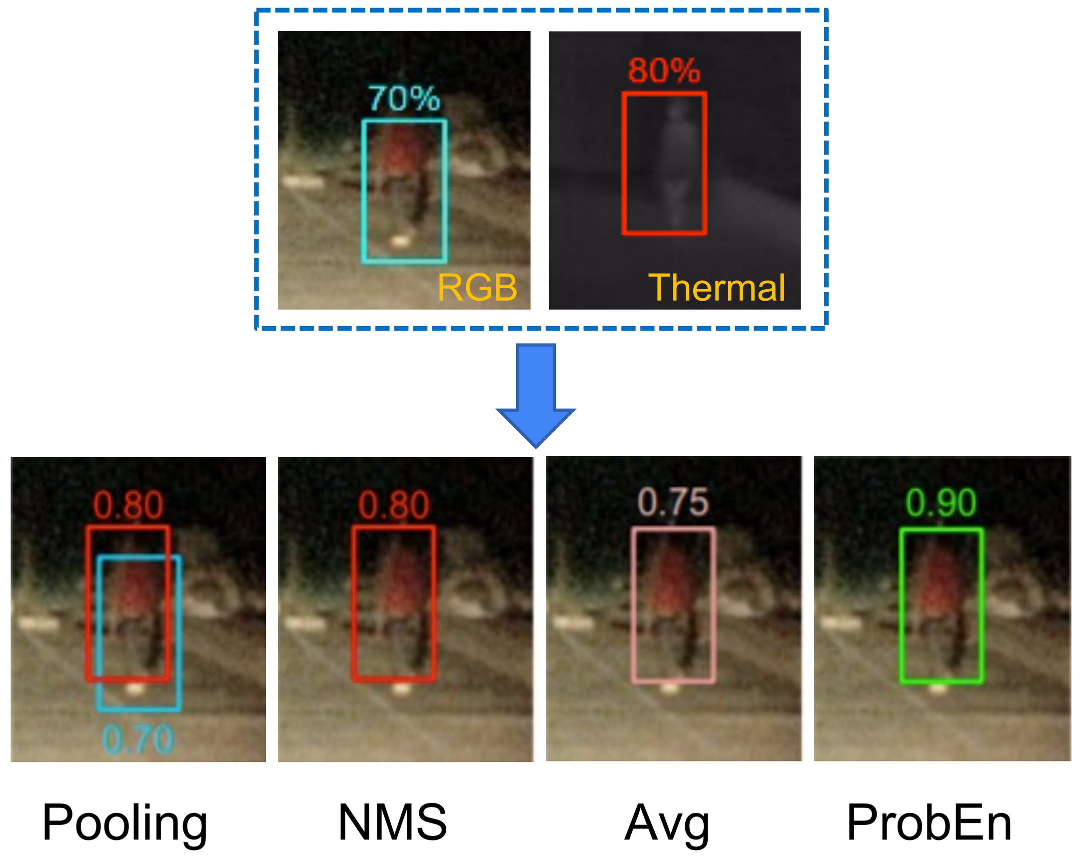


Multimodal Fusion Strategy

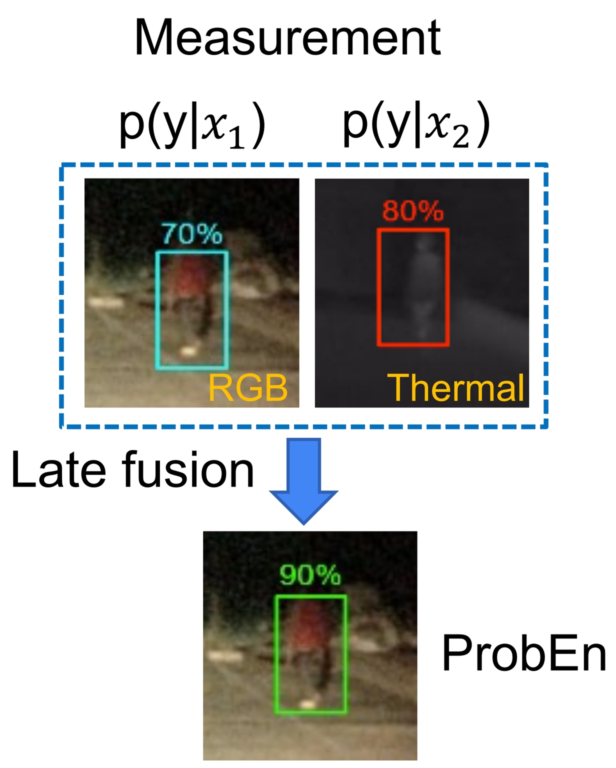


Proposed Method

- Score fusion



- Probabilistic Ensembling



- Assume conditional independence

$$p(x_1, x_2|y) = p(x_1|y)p(x_2|y)$$

- Given multimodal measurement

$$p(y|x_1, x_2) = \frac{p(x_1, x_2|y)p(y)}{p(x_1, x_2)} \propto p(x_1, x_2|y)p(y)$$

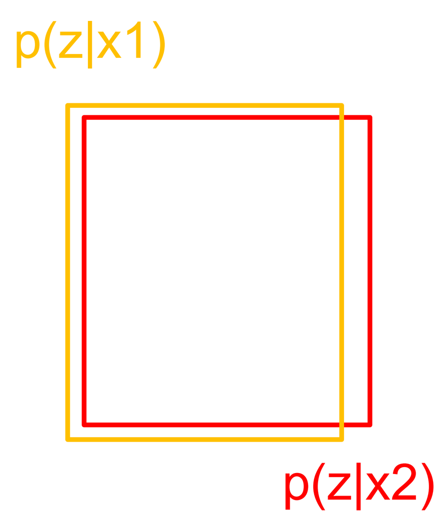
- Bayes rule

$$p(y|x_1, x_2) \propto p(x_1|y)p(x_2|y)p(y)$$

$$\propto \frac{p(x_1|y)p(y)p(x_2|y)p(y)}{p(y)}$$

$$\propto \frac{p(y|x_1)p(y|x_2)}{p(y)}$$

- Box fusion



- Assume Gaussian posterior

$$p(\mathbf{z}|x_i) = \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2\mathbf{I})$$

- Conform with conditional independence

$$p(\mathbf{z}|x_1, x_2) \propto p(\mathbf{z}|x_1)p(\mathbf{z}|x_2)$$

$$\exp\left(\frac{\|\mathbf{z}-\boldsymbol{\mu}_1\|^2}{-2\sigma_1^2}\right) \exp\left(\frac{\|\mathbf{z}-\boldsymbol{\mu}_2\|^2}{-2\sigma_2^2}\right)$$

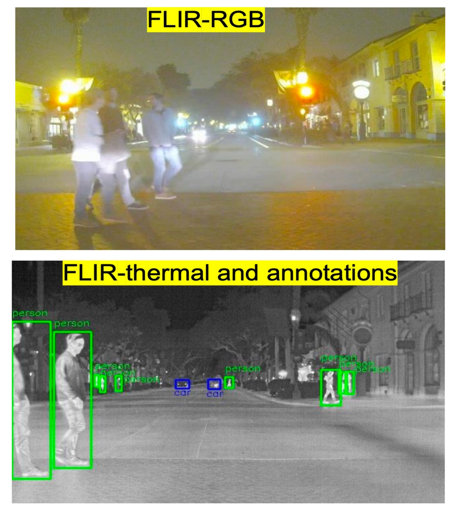
merge

$$\exp\left(\frac{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}}{-2} * \|\mathbf{z}-\boldsymbol{\mu}\|^2\right)$$

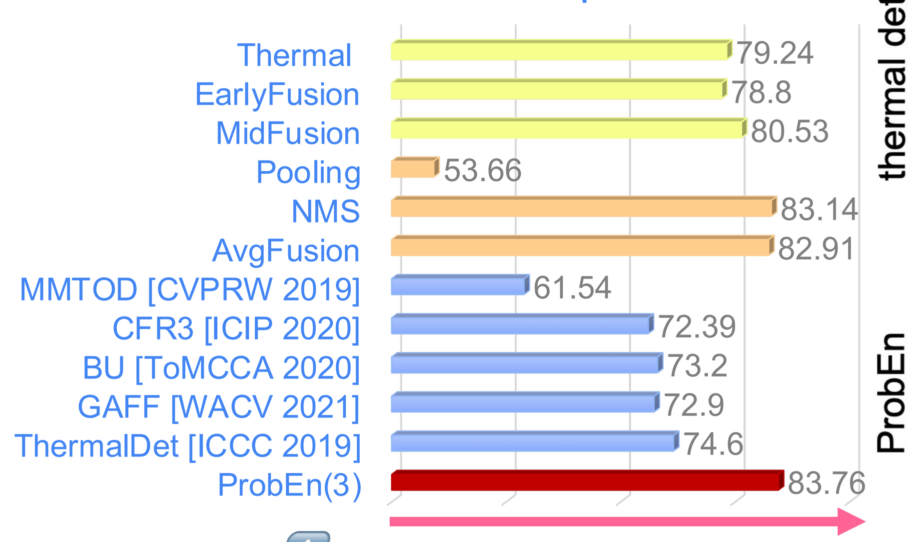
where $\boldsymbol{\mu} = \dfrac{\frac{\boldsymbol{\mu}_1}{\sigma_1^2}+\frac{\boldsymbol{\mu}_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}}$

Experiment

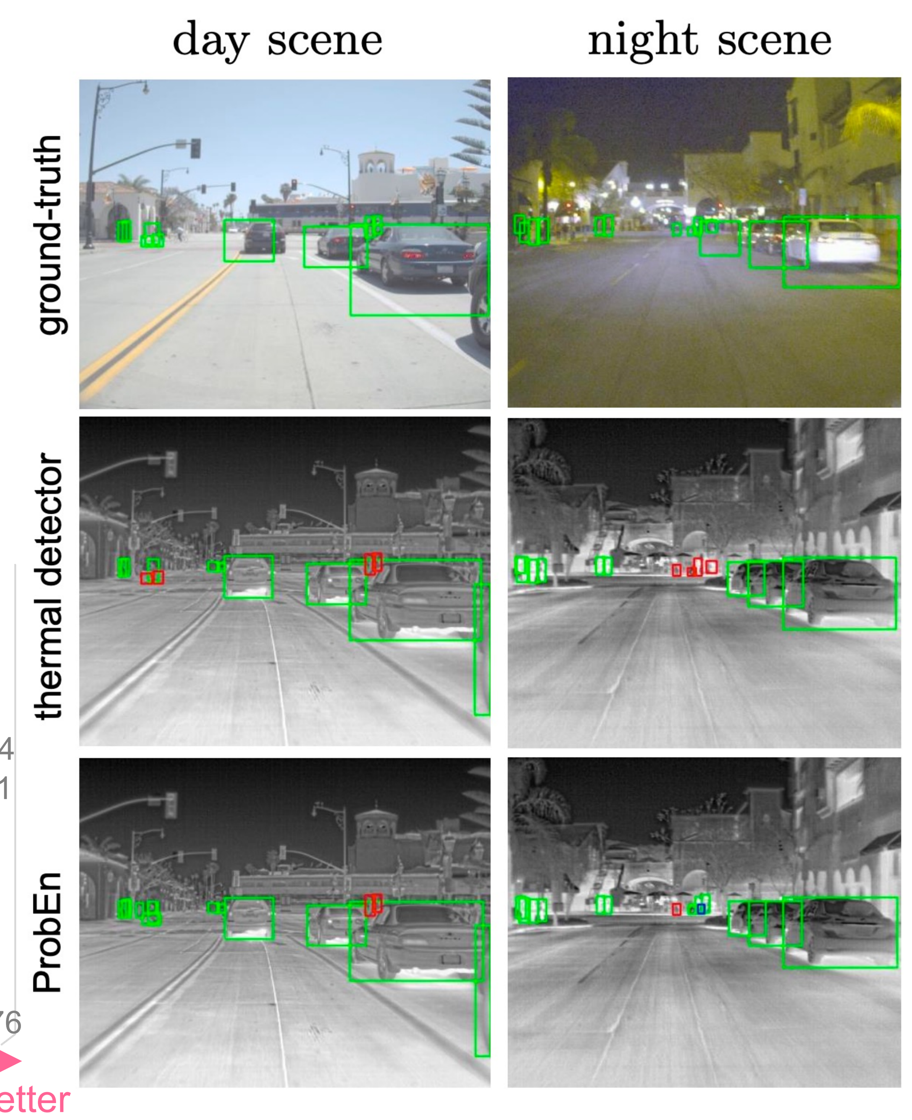
- FLIR (Unaligned / multiclass(3) / high resolution)



FLIR Dataset Comparison

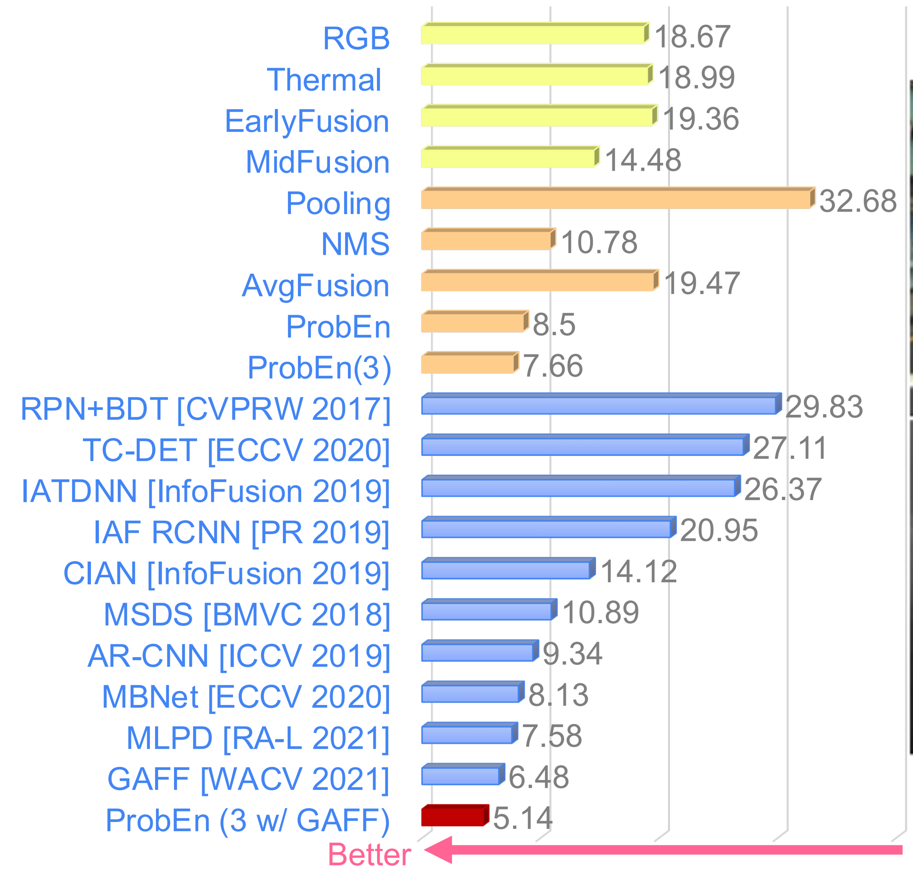


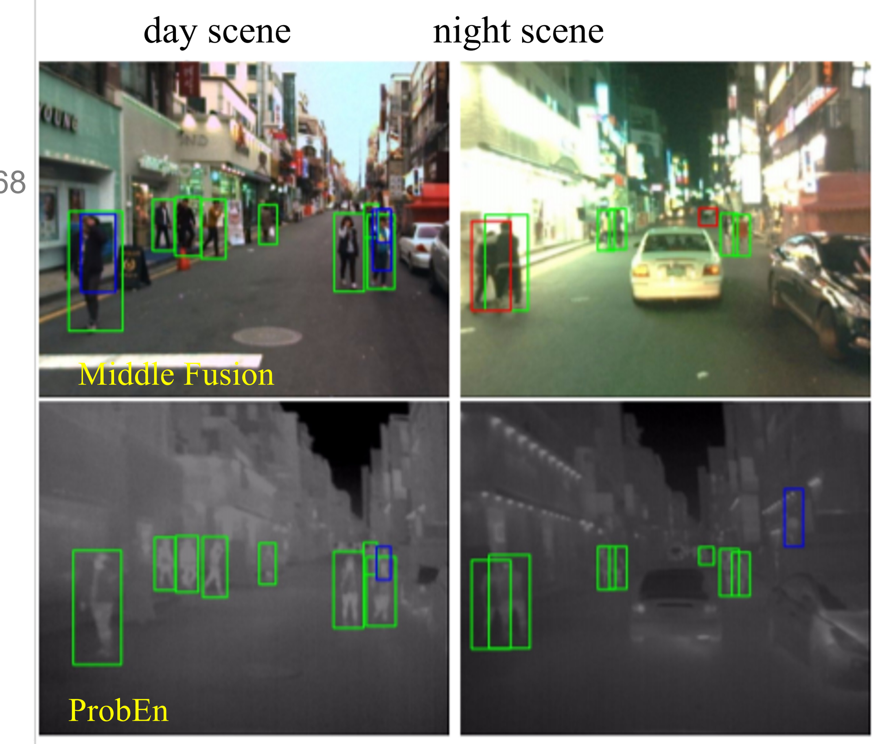
(*mAP*: *mean Average Precision*)



- KAIST (Aligned / one class / low resolution)

KAIST Dataset Comparison





(*LAMR*: *Log Average Miss Rate*)

Conclusion

- ProbEn is significantly better than heuristic methods (e.g., Avg fusion and NMS).
- ProbEn still improves when the conditional independence assumption does not hold.
- ProbEn outperforms significantly than other fusion methods.