Tensor Factorization for Leveraging Cross-Modal Knowledge in Data-Constrained Infrared Object Detection





Manish Sharma[1] Moitreya Chatterjee[2] **Kuan-Chuan Peng**[2] Suhas Lohit[2] Michael J. Jones[2]

[1] RIT | Rochester Institute of Technology

[2] MITSUBISHI ELECTRIC Changes for the Better

ICCV23 PARIS

Representation learning with very limited images
— The potential of self-, synthetic- and formula-supervision —

Problem Statement

Most object detectors work well when provided with sufficient training data.

- Suffer overfitting due to over-parametrization in data scarce regime.
- RGB trained model does not generalize well to infrared/thermal due to significant domain shift.

Our task: Object detection in the data scarce infrared (IR) domain.

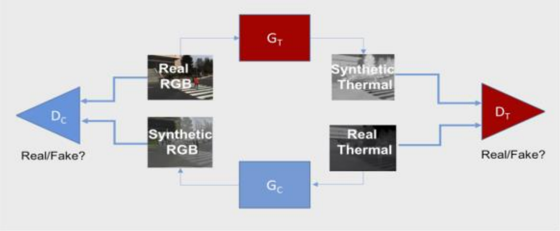
Given: Large amount of publicly available RGB training data.

Research questions:

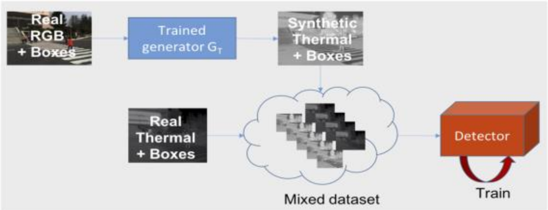
- How to achieve generalizability for object detection from few labelled IR training samples?
- Can we leverage the abundance of annotated RGB data for object detection, in the IR domain?

Related Works

- Domain-Adaptive Pedestrian Detection in Thermal Images [1].



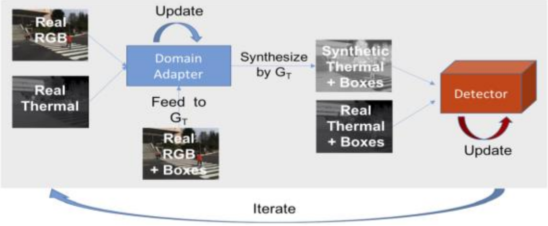
Training of the domain adapter.



Training of the detector with synthetic thermal images generated by a trained domain adapter.

Color Real Thermal Synthetic Thermal

Synthetic thermal image generated from color images in the KAIST test set.



Joint training of the domain adapter and the pedestrian detector in the thermal infrared domain.

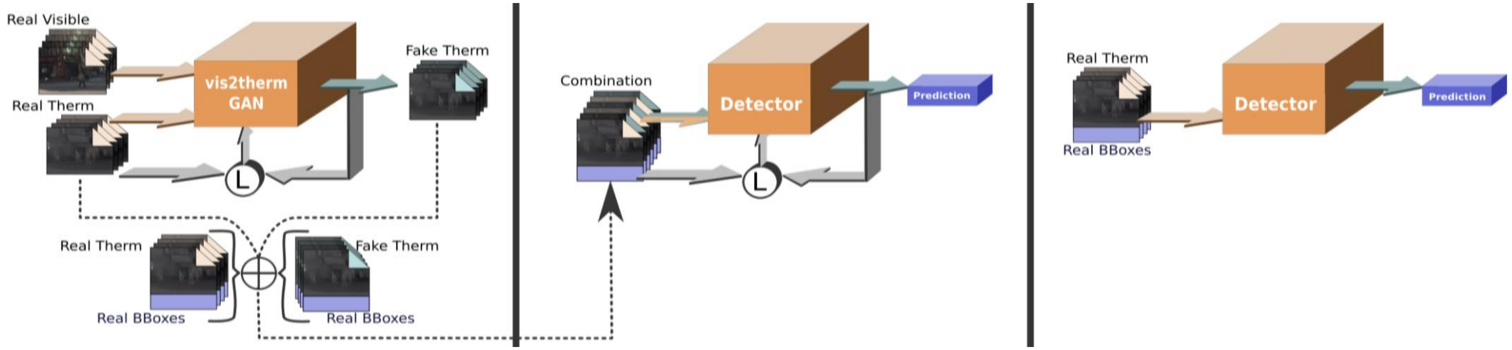
Color Synthetic Thermal

Sample synthetic thermal image transformed from the Caltech dataset.

The domain shift between RGB and IR is significant, so the synthesized IR images from RGB can be unrealistic and they may not capture the IR- specific information which is not in RGB.

- Robust Pedestrian Detection in Thermal Imagery using Synthesized Images [2].

Step 1: Generate Synthetic Data | Step 2: Train over Augmented Data | Step 3: Test over Real Data



1. T. Guo et al. Domain-Adaptive Pedestrian Detection in Thermal Images. ICIP19.
2. M. Kieu et al. Robust pedestrian detection in thermal imagery using synthesized images. ICPR20.

Motivation: Relatively Scarce IR Data

Challenges in acquiring IR data:

- Hardware cost and constraints (less ubiquitous than RGB cameras).
- Expensive and time-consuming data annotation process.
- Privacy concerns and export control regulation.

There exists common feature cues in both RGB and IR data.

- Exploit cross-modal cues at the model level.

Advantages of domain adaptation methods:

- Reduce data acquisition efforts.
- Reduce computational costs.

Contributions

- **TensorFact:** A novel tensor factorization method that can leverage both:
  - modality-specific cues.
  - cross-modal cues.

  for effective object detection in the IR data, where acquiring sufficient training data is a challenge.

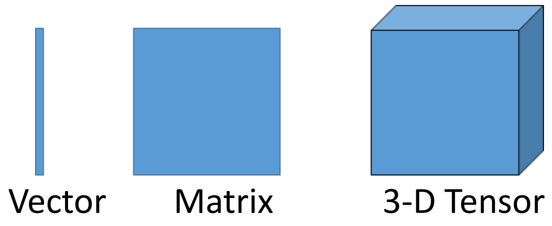
- **TensorFact** outperforms the competing state-of-the-art object detector trained directly on data scarce target IR domain while retaining source RGB domain performance.

Technical Background

Convolution layer: $\underbrace{X^{S\times H\times W}}_{\text{Input}} * \underbrace{K^{T\times S\times D_2\times D_1}}_{\substack{\text{Convolution Filter} \\ \text{(Trainable Parameters)}}} = \underbrace{Y^{T\times H'\times W'}}_{\text{Output}}$

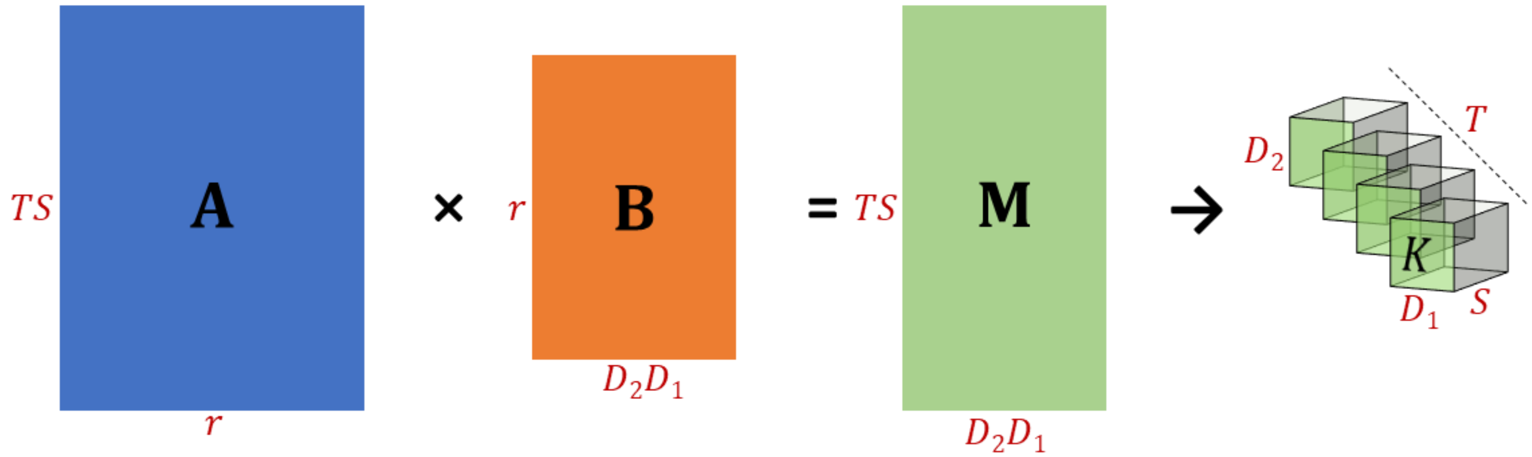
Tensor: n-D array

- 1-D array – Vector
- 2-D array – Matrix
- ≥3-D array – Tensor



Decomposed convolution filter:

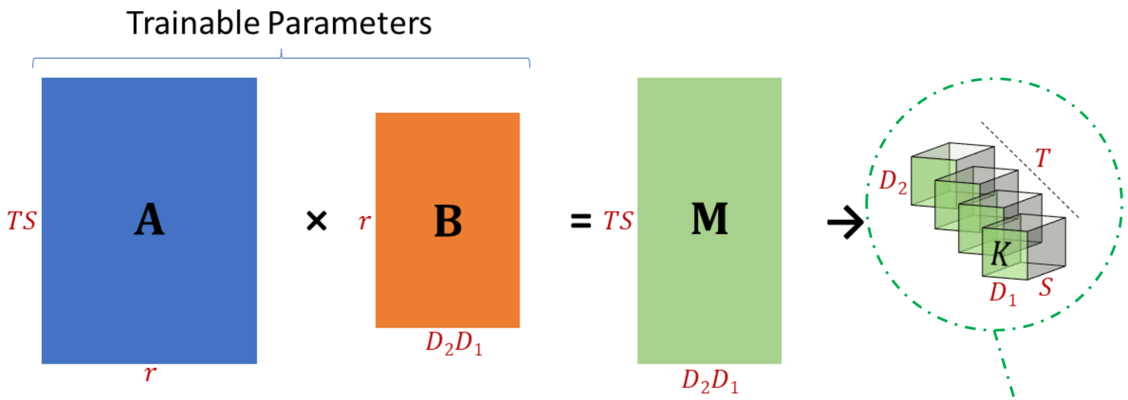
- $[\mathbf{M}]_{p,q} = \sum_{c=1}^{r}[\mathbf{A}]_{p,c}[\mathbf{B}]_{c,q}$
  - $p = 1,2,\dots,TS$
  - $q = 1,2,\dots,D_2D_1$
- $[K]_{t,s,d_2,d_1} = [\mathbf{M}]_{(t-1)S+s,(d_2-1)D_1+d_1}$
  - $t = 1,2,\dots,T$
  - $s = 1,2,\dots,S$
  - $d_2 = 1,2,\dots,D_2$
  - $d_1 = 1,2,\dots,D_1$



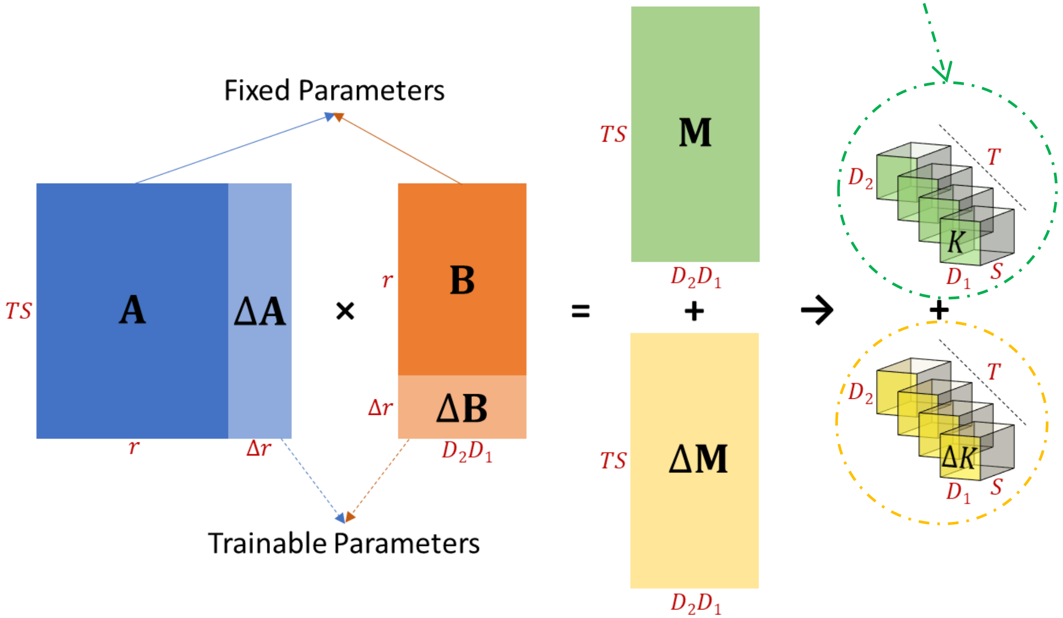
Proposed Method — TensorFact

TensorFact: Designed to tackle data scarcity in the IR data.

For RGB: Low-rank decomposed convolution filter.



For IR: Capacity augmentation



For Standard Convolution Filter:

- # trainable parameters $(P) = TSD_2D_1$

For RGB: $\mathbf{A}$ & $\mathbf{B}$ are SVD initialized

- # trainable parameters $\left(P_{fac}\right) = r(TS + D_2D_1)$

In general, $0 < r \leq r_{max}$, $r_{max} = \min(TS, D_2D_1)$.
For varying r across network layers using a single variable - Introduce α hyperparameter.

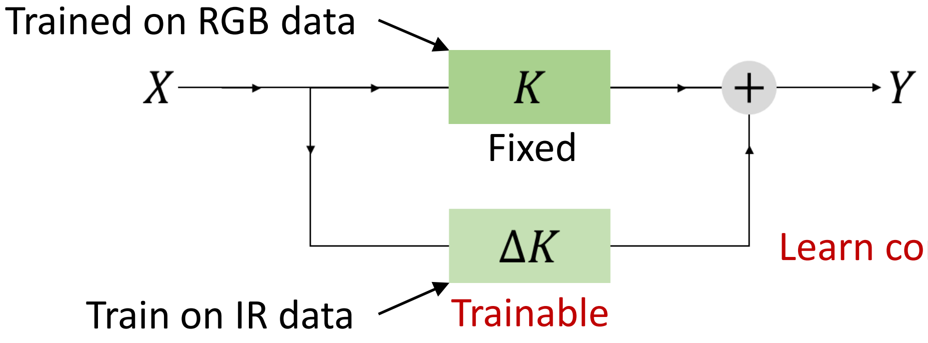
- $r = \alpha r_{max}$, $\alpha \in (0,1]$
- $P_{fac} = \alpha r_{max}(TS + D_2D_1)$

For IR:

- # trainable parameters $\left(\Delta P_{fac}\right) = \Delta r(TS + D_2D_1)$
  - $\Delta r = \Delta \alpha r_{max}$

Proposed Method — TensorFact (cont'd)

2-Branch Architecture:

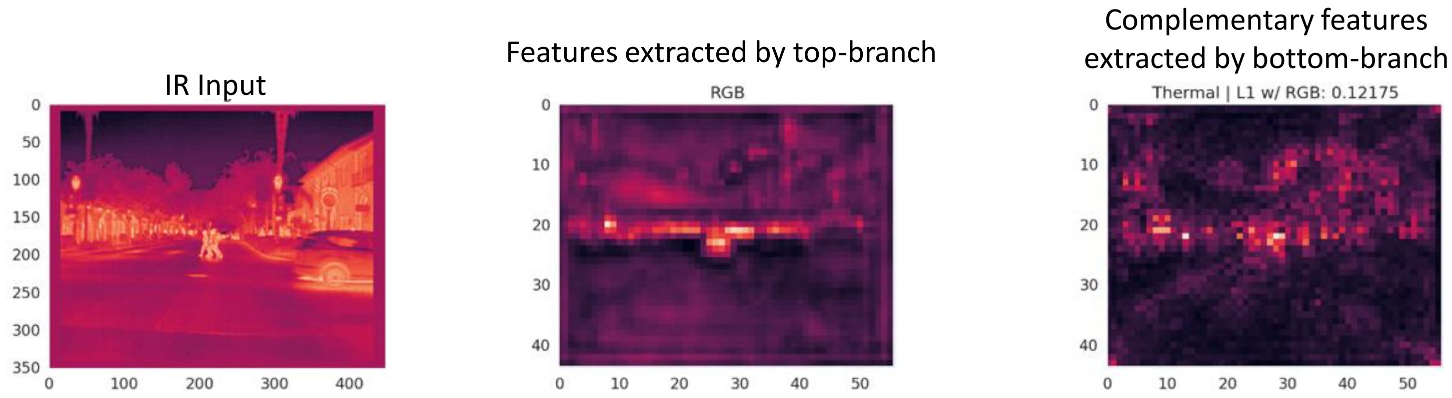


Learn complementary features: $\Delta K(X) = Y - K(X)$

To promote learning complementary features, we propose the following loss term:

- $\max(\|K(X) - \Delta K(X)\|_p), \quad p = \{1,2\}$

To increase the distance between feature maps



Experiment Setup

Baseline: YOLOv7 [1]

- # Trainable Parameters: about 37 M.

Datasets:

- FLIR Aligned RGB [2]
  - Classes: Person, Bicycle, and Car.
- FLIR ADAS v1 IR [3]
  - Classes: Person, Bicycle, and Car.
  - Dataset Configuration:
    - Data constrained (Use only 1% of training data).

Evaluation Metric:

- Mean Average Precision (mAP) $= \frac{1}{n_c}\sum_{i=1}^{n_c} \bar{P}_i$
  - mAP 50
  - mAP 50-95

$\bar{P}_i$: Average Precision for i^{th} class
n_c: number of classes

FLIR Aligned RGB dataset and instances distribution

Split	#Images
Train	4129
Val	1013
Total	5142

Class	#Train Instances	#Val Instances
Person	8987	4107
Bicycle	2566	360
Car	20608	4124
Total	32161	8591

FLIR ADAS v1 IR (1%) dataset and instances distribution

Split	#Images
Train	62
Val	1572
Total	1634

Class	#Train Instances	#Val Instances
Person	161	4611
Bicycle	24	842
Car	351	8472
Total	536	13925

1. Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. CVPR. https://arxiv.org/abs/2207.02696
2. FLIR aligned. FLIR Aligned Dataset, 2020. Accessed: August 20, 2022.
3. Teledyne Technologies Incorporated. FLIR ADAS v1 Dataset, 2020. Accessed: August 20, 2022.

Results

Model	# Parameters (M) ↓	Compression (%) ↑	mAP 50 (%) ↑	mAP 50-95 (%) ↑
YOLOv7	37.21	0	68.26	**31.73**
TensorFact ($\alpha = 0.9$)	35.40	4.85	**69.48**	31.62
TensorFact ($\alpha = 0.8$)	**33.59**	**9.71**	68.79	31.68

Results for FLIR Aligned RGB validation dataset

Pre-trained

Model	# Parameters ↓	Compression (%) ↑	mAP 50 (%) ↑	mAP 50-95 (%) ↑
YOLOv7	37.21	0	58.49	**28.07**
TensorFact ($\alpha = 0.1$)	**1.86**	**95.01**	62.05	**28.07**
TensorFact ($\alpha = 0.2$)	3.66	90.16	**62.13**	27.94

Pre-trained

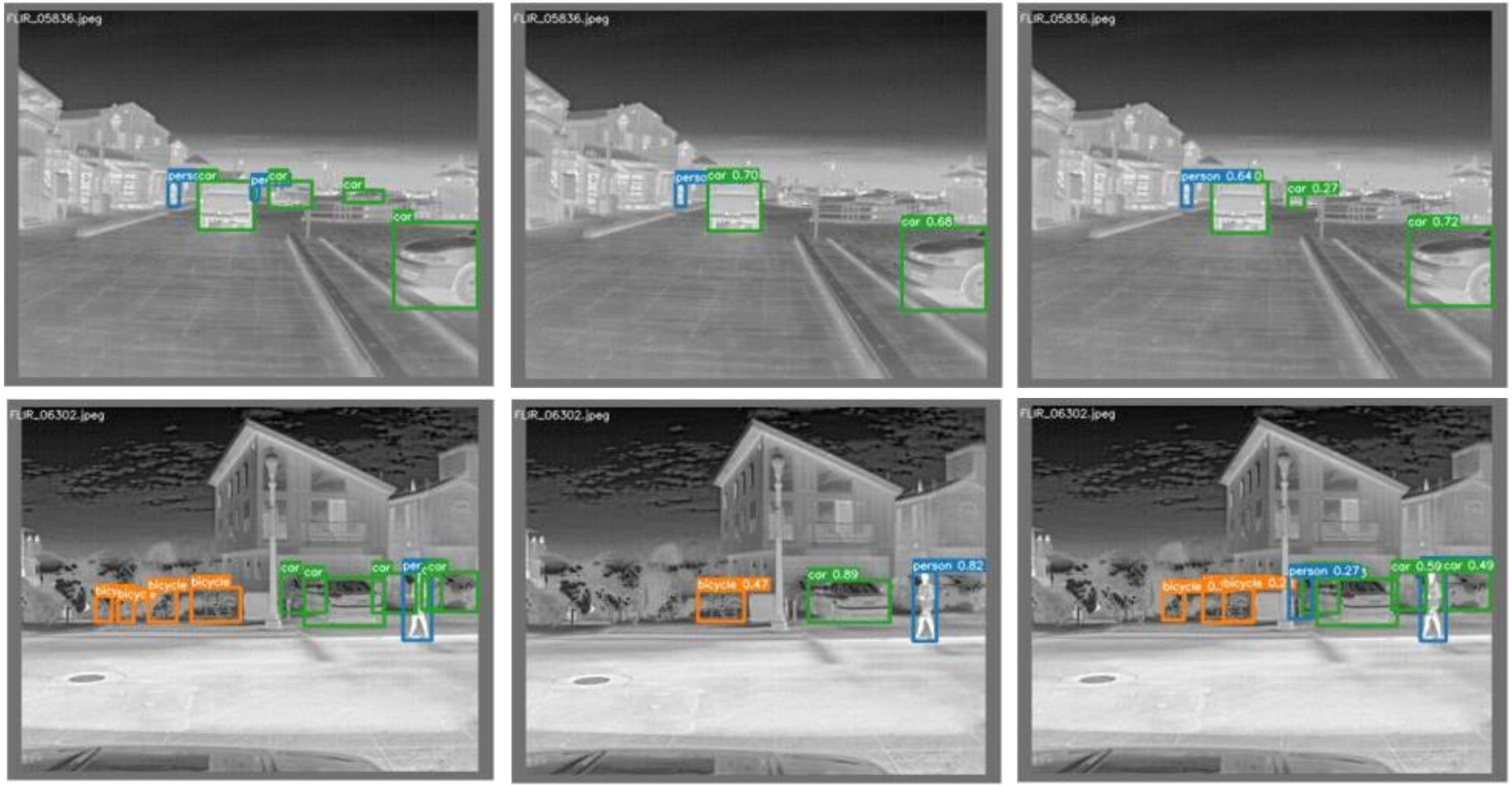
Results for FLIR ADAS v1 IR validation dataset

Regularization	mAP 50 (%) ↑	mAP 50-95 (%) ↑
N/A	62.05	28.07
L_1	**62.34**	**28.23**
L_2	62.22	28.15

Results with explicit complementary regularization for $\alpha = 0.1$ on FLIR ADAS v1 IR validation dataset

Qualitative Results

Ground Truth | YOLOv7 | TensorFact



YOLOv7 fails to detect small and distant objects, but TensorFact can detect them.

Conclusions & Future Work

Summary:

We propose TensorFact—a method to architecturally promote learning of cross-modal cues.

- Improve generalization for modalities with scarce training data (as low as 62 samples).
- Require only a fraction of trainable parameters (5% of total parameters).
- Empirically validated the efficacy of our method for object detection.

Future Work:

- Explore attention between RGB and IR branches during forward pass to reduce false detection.
- Extend to other applications (e.g. segmentation).

Thank you!

Questions?