Modality Translation for Object Detection Adaptation Without Forgetting Prior Knowledge

Heitor Rapela Medeiros*, Masih Aminbeidokhti, Fidel A. Guerrero Pena, David Latortue, Eric Granger, Marco Pedersoli

heitor.rapela-medeiros.1@ens.etsmtl.ca

Introduction

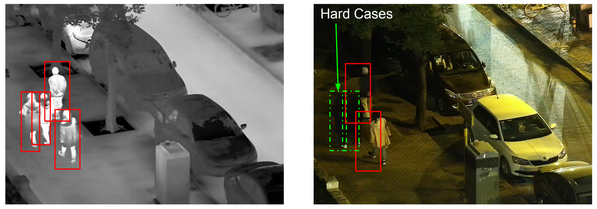

Figure 1. IR and RGB images (LLVIP dataset).

Our work investigates **modality translation** for **Object Detection**.

- On the web, there are **many pre-trained RGB detectors**.
- Our model adapts from **zero-shot RGB** detectors to I**R modality**.
- It **does not modify** the original detector weights.

ModTr

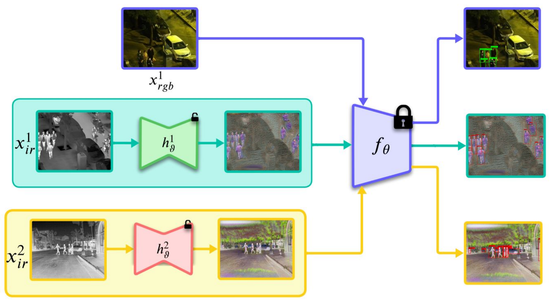

Figure 2. ModTr: Adapts the zero-shot RGB model in the input level.

$$\mathcal{L}_{\text{ModTr}}(x, Y; \vartheta) = \mathcal{L}_{det}(f_\theta\left(\Phi(h^d_\vartheta(x), x)\right), Y)$$

- We **fuse** the **translated modality** with the **input**.
- We **calculate** the **gradient** with the **detection output**.
- We **update translator parameters** with respect to the **detection loss**.

Adapting Without Forgetting

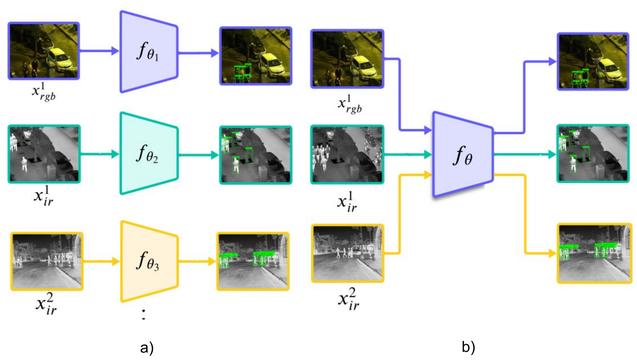

Figure 3. a) N-Detectors and b) 1-Detector models.

Detector	Dataset	N-Detectors	1-Detector	N-ModTr-1-Det.
FCOS	LLVIP	57.37 ± 2.19	**58.55 ± 0.89**	57.63 ± 0.66
	FLIR	27.97 ± 0.59	26.70 ± 0.48	**35.49 ± 0.94**
	COCO	**38.41 ± 0.00**	00.33 ± 0.04	**38.41 ± 0.00**
	AVG.	41.25 ± 0.92	28.52 ± 0.47	**43.84 ± 0.53**
RetinaNet	LLVIP	53.79 ± 1.79	53.26 ± 3.02	**54.83 ± 0.61**
	FLIR	28.46 ± 0.50	25.19 ± 0.72	**34.27 ± 0.27**
	COCO	**35.48 ± 0.00**	00.29 ± 0.01	**35.48 ± 0.00**
	AVG.	39.24 ± 0.76	26.24 ± 1.28	**41.52 ± 0.29**
Faster R-CNN	LLVIP	59.62 ± 1.23	**62.50 ± 1.29**	57.97 ± 0.85
	FLIR	30.93 ± 0.46	28.90 ± 0.33	**37.21 ± 0.46**
	COCO	**39.78 ± 0.00**	00.40 ± 0.00	**39.78 ± 0.00**
	AVG.	43.44 ± 0.56	30.60 ± 0.54	**44.98 ± 0.43**

Table 3. Detection performance (AP) of knowledge-preserving techniques.

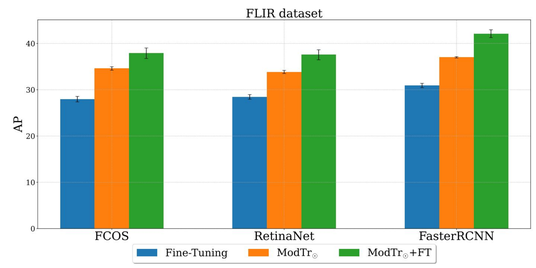

Figure 4. Performance of ModTr with and without FT for FLIR dataset.

Quantitative Analysis

Image translation	RGB	Box	Test Set IR (Dataset: LLVIP) FCOS	RetinaNet	Faster R-CNN
Histogram Equal. [14]			31.69 ± 0.00	33.16 ± 0.00	38.33 ± 0.02
CycleGAN [51]	✓		23.85 ± 0.76	23.34 ± 0.53	26.54 ± 1.20
CUT [37]	✓		14.30 ± 2.25	13.12 ± 2.07	14.78 ± 1.82
FastCUT [37]	✓		19.39 ± 1.52	18.11 ± 0.79	22.91 ± 1.68
HalluciDet [29]	✓	✓	28.00 ± 0.92	19.95 ± 2.01	57.78 ± 0.97
ModTr$_\odot$ (ours)		✓	**57.63 ± 0.66**	**54.83 ± 0.61**	**57.97 ± 0.85**

Table 1. IR detection AP performance with different image translation methods.

Method	Test Set IR (Dataset: LLVIP) FCOS	RetinaNet	Faster R-CNN
Fine-Tuning (FT)	57.37 ± 2.19	53.79 ± 1.79	**59.62 ± 1.23**
FT Head	49.11 ± 0.70	44.00 ± 0.28	59.33 ± 2.17
LoRA [18]	47.72 ± 0.58	-	54.83 ± 1.30
ModTr$_\odot$ (ours)	**57.63 ± 0.66**	**54.83 ± 0.61**	57.97 ± 0.85

Table 2. AP performance benchmark for different detection fine-tuning strategies.

Qualitative Results

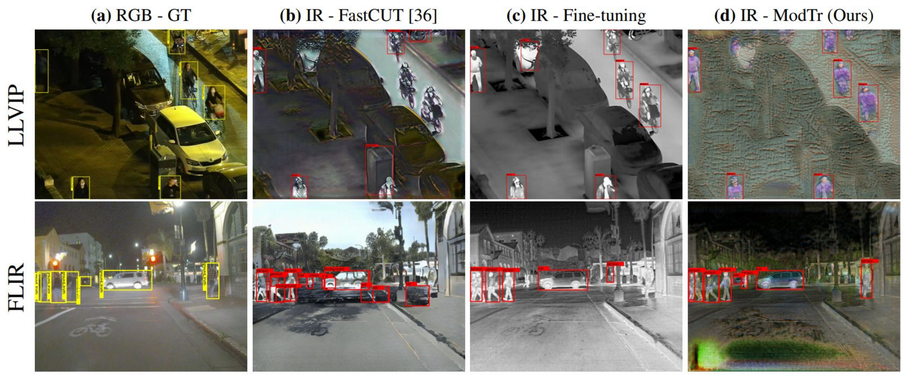

Figure 5. Bounding box predictions over different adaptations of the Faster R-CNN for IR images.

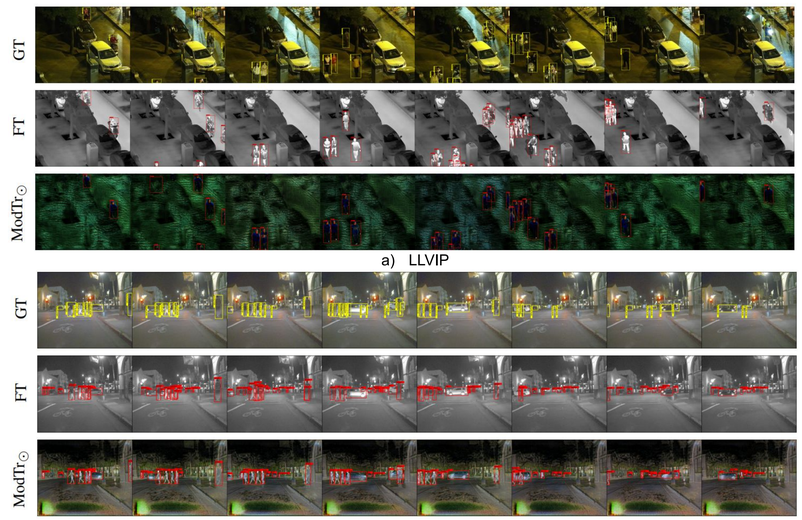

Figure 6. Illustration of a sequence of 8 images of a) LLVIP and b) FLIR for Faster R-CNN.

Conclusion

- ✓ We propose a **novel** approach: **ModTr**, which adapts **RGB object detectors** for **IR modality** without changing their parameters.
- ✓ It **preserves the full knowledge** of the detector, allowing the translation network to act as a node that **changes the modality** for an **unaltered detector**.
- ✓ Our technique demonstrated **good detection performance** on different traditional **IR benchmarks**.

arXiv https://arxiv.org/abs/2404.01492

GitHub https://github.com/heitorrapela/ModTr