OCTOBER 1 - 5, 2023
IEEE/RSJ International Conference
on Intelligent Robots and Systems

Temporal CFT: Multi-Temporal Cross-Modality Fusion Transformer for Multispectral Video Object Detection

Srikar Varaganti, Asiegbu Miracle Kanu-Asiegbu, Xiaoxiao Du

University of Michigan, Ann Arbor

Motivation

Challenges:

1. Fusion - RGB sensors can suffer in poor lighting and weather conditions, whereas thermal cameras operate well in poor lighting and weather conditions but can have a lower resolution.

2. Temporal - Single-frame models fail to consider temporal relationship in traffic object movement.

Goal: Multi-temporal Cross-Modality Fusion for Thermal + RGB sensors

Problem Statement

Input:

RGB video sequence $\mathrm{I}_R \in \mathbb{R}^{HW \times C \times N}$

+ Thermal video sequence $\mathrm{I}_T \in \mathbb{R}^{HW \times C \times N}$

Output:

Object bounding boxes at last frame

Network:

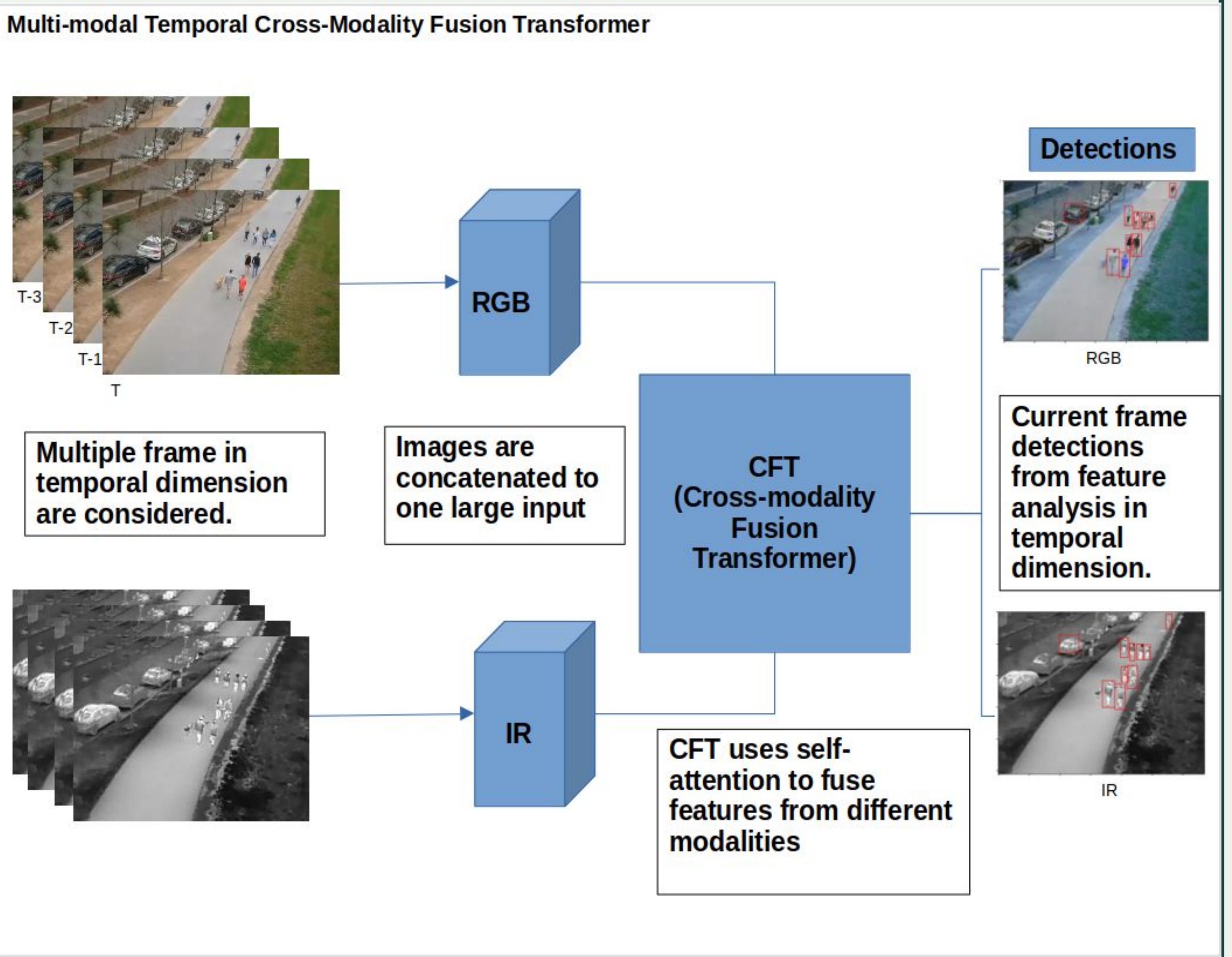
Cross-Modality Transformer [3] for fusion with temporal concatenation; YOLOv5 detector

References

1] A. Berg, J. Ahlberg, and M. Felsberg, “A thermal object tracking benchmark,” in Int. Conf. Advanced Video and Signal Based Surveillance, 2015, pp. 1–6.
[2] V. Vs, J. M. J. Valanarasu, P. Oza, and V. M. Patel, “Image fusion transformer,” in IEEE Int. Conf. Image Processing, 2022, pp. 3566–3570.
[3] F. Qingyun, H. Dapeng, and W. Zhaokui, “Cross-modality fusion transformer for multispectral object detection,” arXiv preprint arXiv:2111.00273, 2021.
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” NIPS, vol. 30, 2017.
[5] E. Gebhardt and M. Wolf, “Camel dataset for visual and thermal infrared multiple object detection and tracking,” in Int. Conf. Advanced Video and Signal-based Surveillance, 2018.

Method

Figure 1: Diagram of the proposed multi-modal Temporal CFT.


Multi-modal Temporal Cross-Modality Fusion Transformer
T-3
T-2
T-1
T
Multiple frame in temporal dimension are considered.
RGB
Images are concatenated to one large input
IR
CFT (Cross-modality Fusion Transformer)
CFT uses self-attention to fuse features from different modalities
Detections
RGB
Current frame detections from feature analysis in temporal dimension.
IR

Takeaways

1. Temporal model achieved 78% True positive rate compared to 54% from baseline: **24% improvement**
2. Temporal models showed **better performance in handling data imbalance** than baseline.
3. **In future**: Our research can be focused on designing and implementing fusion in the temporal dimension that is more memory efficient and better utilizes multi-modal features.
4. Temporal models trained with multi-modalities are better in **learning**, **inference** and **classification**.

Results

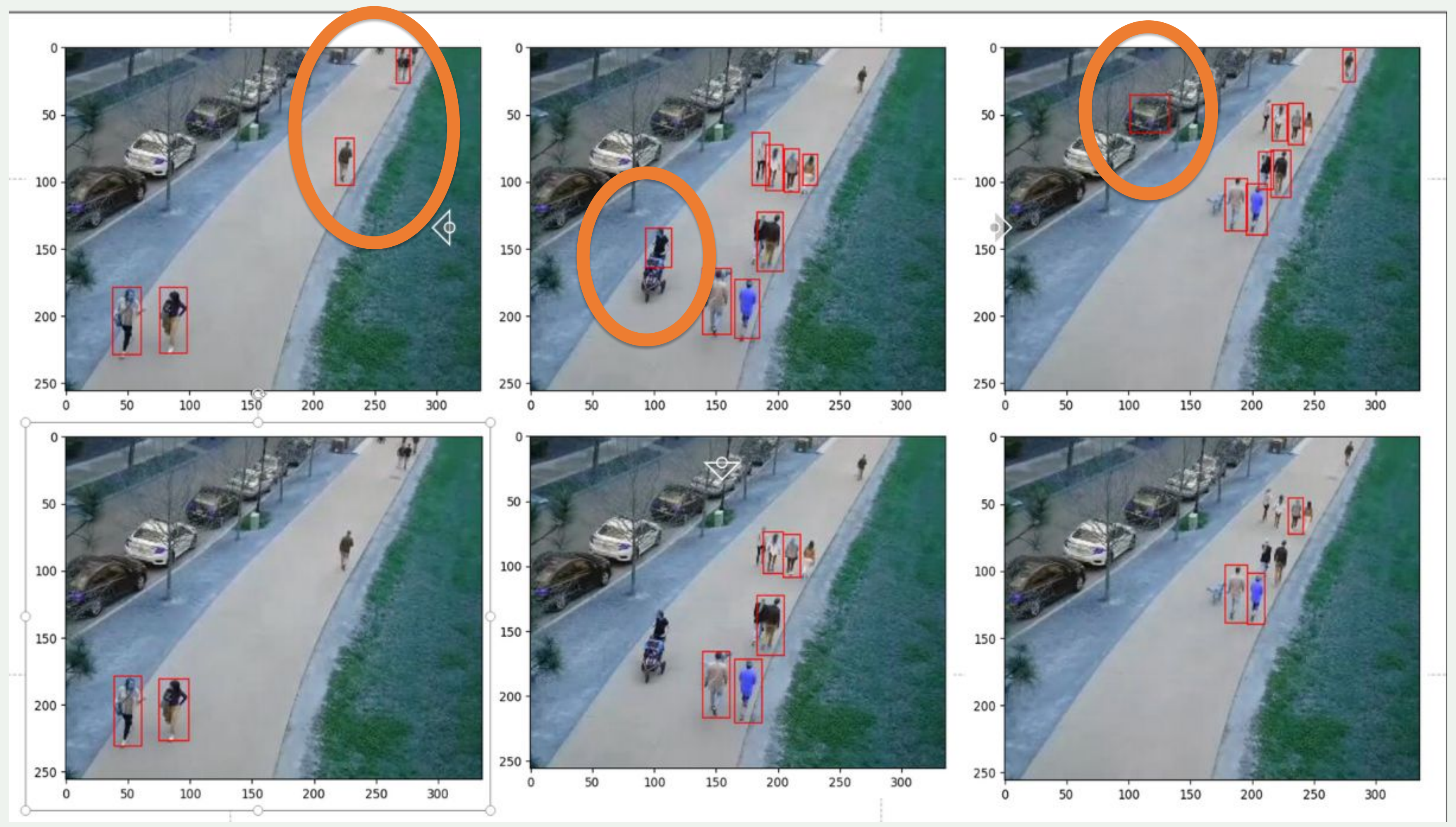


Figure 2: Object Detection Results. Top: Temporal-CFT model (proposed), Bottom: Single-frame model (Baseline). Orange indicates correct detections from the proposed Temporal CFT.

Table 1: Person detection results on the CAMEL dataset.

Model Type	True Positive ↑	False Negative ↓
Single-Frame CFT	54	46
Temporal CFT (Proposed)	**78**	**22**

Acknowledgements

*This work was supported by a grant from Ford Motor Company via the Ford-UM Alliance under award N028603 and the National Science Foundation under Grant IIS-2153171-CRII: III: Explainable Multi-Source Data Integration with Uncertainty.