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

Srikar Varaganti[1], Asiegbu Miracle Kanu-Asiegbu[2], and Xiaoxiao Du[1]

I. Motivation

Two challenges exist for the object detection task. First, the majority of existing object detection methods are based on RGB camera data. However, RGB images can suffer from bad visibility at night and low-light conditions. Thermal cameras, on the other hand, can detect infrared radiation from objects (e.g., persons) and are robust to illumination changes and shadow effects while being less intrusive to privacy [1]. Thus, it is valuable to develop a multispectral object detector based on both RGB and thermal data to take advantage of the complementary and reinforcing information from these two sensor modalities. Second, existing cross-modality fusion methods for thermal and RGB data (e.g., [2], [3]) only consider image-level inputs. In other words, only one static RGB and thermal image pair is used as input for detection. However, traffic objects, such as pedestrians, move in continuous motion, and it is natural to consider temporal relationship when fusing RGB and thermal video data to enhance object detection performance. This abstract proposes *Temporal CFT*, a multi-temporal cross-modality fusion transformer for multispectral video object detection.

II. Problem Statement

Network Architecture: We extend the Cross-Modality Fusion Transformer (CFT) [3] for video thermal and RGB fusion to account for temporal information. The CFT is a highly effective multispectral object detector but only considers static thermal and RGB image pairs as input. In CFT, convolutional layers are used to extract thermal and RGB image features. Then, the flattened feature maps are passed through a transformer network [4], where the dependencies and global contextual information between modalities and between features are learned through the attention layers. A detector, such as YOLOv5, can then be used to detect objects given the feature output from the transformer module. In our proposed model, we adapt CFT by allowing multi-temporal video inputs. Denote RGB and thermal video inputs as $\mathbf{I}_R \in \mathbb{R}^{HW\times C\times N}$ and $\mathbf{I}_T \in \mathbb{R}^{HW\times C\times N}$, where N is the input image sequence length (in our experiments, $N=6$). The input images sequences are concatenated along the temporal dimension and passed through the convolution layers for feature extraction. The outputs of the network are object bounding boxes in the last frame of the sub-sequences.
Implementation Details: The models were trained using stochastic gradient descent for 250 epochs with a learning rate of $1\mathrm{e}{-2}$ and a momentum of 0.937. A focal loss with a gamma value of 2.0 was used with IoU threshold of 0.6.

III. Results

We evaluated the proposed temporal CFT method on the *CAMEL* dataset [5] with 17,574/2,009/877 RGB and thermal image pairs for training/validation/testing. We report the person detection results to demonstrate the model's performance. Table I shows quantitative comparisons between the single-frame CFT and our proposed Temporal CFT with multi-temporal, cross-modality video inputs. We report true positive rate (TPR) and false negatives (FN) as evaluation metrics for person detection. The single-frame CFT generated 54% TPR (i.e., correct detections), while our proposed temporal CFT model yielded a significant increase (24%) in correct detections, achieving 78% true positive rate. Figure 1 shows a visual comparison of the detection results from both models. As shown, the proposed temporal CFT model correctly detected many pedestrians on the sidewalk (shown in red circles) while the single-input CFT missed.

TABLE I: Person detection results on CAMEL dataset.

Model type	True Positive % ↑	False Negative % ↓
Single-frame CFT	54	46
Temporal CFT (Proposed)	**78**	**22**

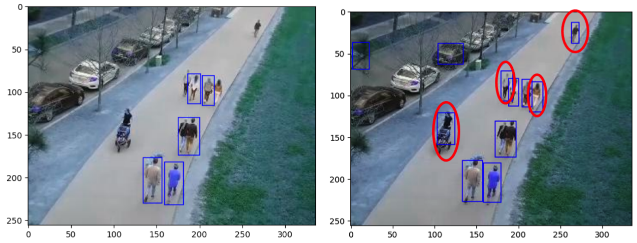

Fig. 1: Visual examples of person detection. Left: CFT with single-frame input. Right: Temporal CFT (proposed).

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.

*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..

[1]Department of Robotics and [2]Department of Mechanical Engineering, University of Michigan, Ann Arbor, USA 48109
{vsrikar, akanu, xiaodu}@umich.edu