Deep Active Learning from Multispectral Data Through Cross-Modality Prediction Inconsistency

Heng ZHANG[1,3] Elisa FROMONT[1,4] Sébastien LEFEVRE[2] Bruno AVIGNON[3]

[1]Univ Rennes, IRISA [2]Univ Bretagne Sud, IRISA [3]ATERMES Company [4] IUF, Inria

Abstract

Data from multiple sensors provide independent and complementary information, which may improve the robustness and reliability of scene analysis applications. While there exist many large-scale labelled benchmarks acquired by a single sensor, collecting labelled multi-sensor data is more expensive and time-consuming. In this work, we explore the construction of an accurate multispectral (here, visible & thermal cameras) scene analysis system with minimal annotation efforts via an active learning strategy based on the cross-modality prediction inconsistency. Experiments on multispectral datasets and vision tasks demonstrate the effectiveness of our method. In particular, with only 10% of labelled data on KAIST multispectral pedestrian detection dataset, we obtain comparable performance as other fully supervised State-of-the-Art methods.

Introduction

- Multispectral systems use two types of camera sensors (RGB and Thermal) to provide complementary information under various illumination conditions.
- Collecting labelled multi-sensor data is expensive and time-consuming, therefore active learning is particularly appealing.
- Multi-sensor **redundancy**: detection results from the two modalities are similar in most cases (Figure 1 left).
- Multi-sensor **complementarity**: at least one modality is wrong when the detections are contradictory (Figure 1 right).
- We rely on the Cross-Modality prediction inconsistency to adaptively select the most informative multispectral samples for annotation.

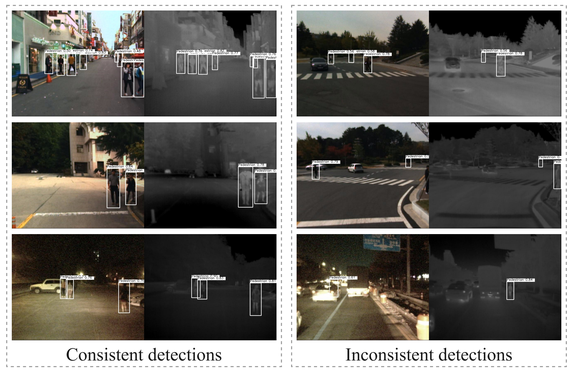


Figure 1:Exemplary multispectral image pairs and their corresponding mono-spectral pedestrian detection results.

Active learning

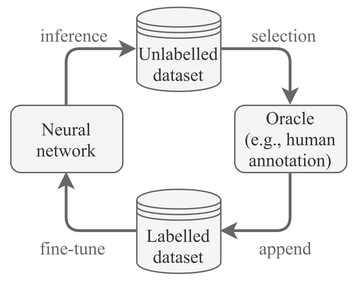


Figure 2:Active learning loop diagram.

The active learning loop starts by pre-training a model on a small subset of the labelled dataset. Then, several active learning cycles are repeated:

- The model inference is performed on the unlabelled dataset to select the most informative samples (i.e., multispectral image pairs).
- These selected samples are then sent to an external oracle for annotation and appended to the labelled dataset.
- The model is consequently fine-tuned on the labelled dataset.

In general, the most important component of an active learning cycle is the **scoring function**, that ranks the informativeness of unlabelled samples.

Network architecture

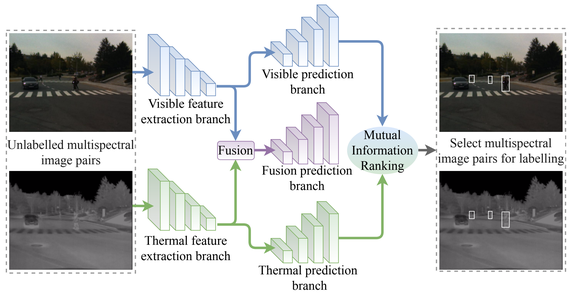


Figure 3:Overview of the proposed model for deep active multispectral scene analysis.

At the selection stage of each active learning cycle, we measure the relevance of labelling a particular image pair by comparing predictions from visible and thermal cameras, and we select image pairs with the highest prediction difference.

Cross-modality prediction inconsistency

For each prediction p, its inconsistency is defined as:

$$\mathcal{I} = \mathcal{H}(p) - \frac{1}{2} \sum_{m \in \{v,t\}} \mathcal{H}(p_m)$$

where p_v and p_t denote the prediction from visible and thermal prediction branches; p is the average of both predictions; $\mathcal{H}$ is the 2-set entropy function calculated as:

$$\mathcal{H}(p) = -p \log p - (1-p) \log (1-p)$$

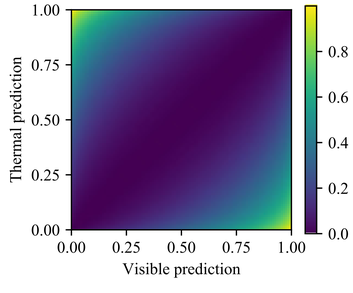


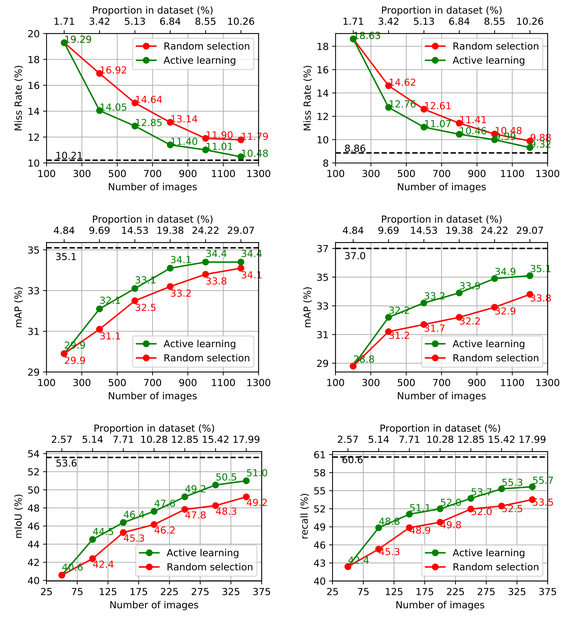
Figure 4:Cross-modality prediction inconsistency visualization.

Experiments

Multispectral datasets

- KAIST Dataset for pedestrian detection;
- FLIR Dataset for object detection;
- TOKYO Dataset for semantic segmentation.

Experimental results



(For more details, please refer to our paper.)

Conclusion

In this paper, we start from the observation of the **redundancy** and the **complementarity** of a multispectral system. We build upon these to suggest relying on the **cross-modality prediction inconsistency** as the criterion to select informative image pairs for labelling within active learning cycles.

Extensive experiments on three different multispectral scene analysis tasks demonstrate the effectiveness of the proposed method.