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

Heng ZHANG, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

IRISA Laboratory, ATERMES Company
{heng.zhang, elisa.fromont, sebastien.lefevre}@irisa.fr bavignon@atermes.fr

2021 IEEE International Conference on Image Processing

Paper ID: 1463
Keywords: active learning, multispectral pedestrian detection, semantic segmentation, multiple sensor fusion

Multispectral scene analysis

Day

Night

❖ The performance of scene analysis applications using only RGB images may be compromised in many real life situations (such as nighttime or shaded areas).

❖ Multispectral systems use two types of camera sensors (**RGB** and **Thermal**) to provide complementary information under various illumination conditions.

❖ However, collecting labelled multi-sensor data is **expensive and time-consuming**.

Proposed solution: Active learning

Consistent detections

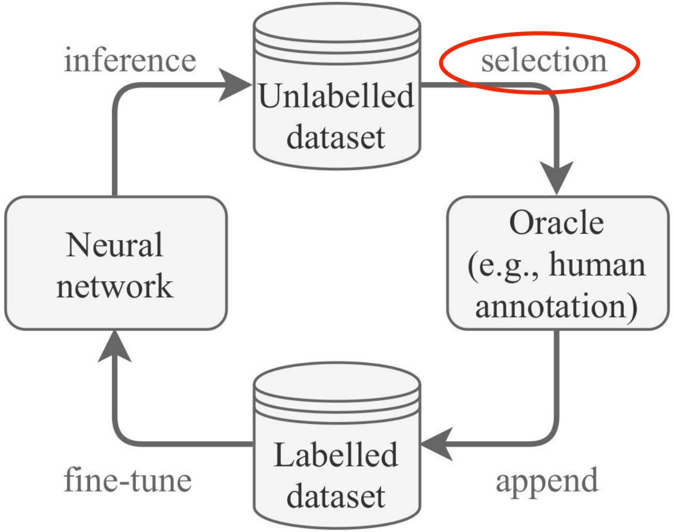
Inconsistent detections

Multi-sensor redundancy: detection results from the two modalities are similar in most cases

Multi-sensor complementarity: at least one modality is wrong when the detections are contradictory

We rely on the **cross-modal predictions' inconsistency** to adaptively select the multispectral samples to be annotated.

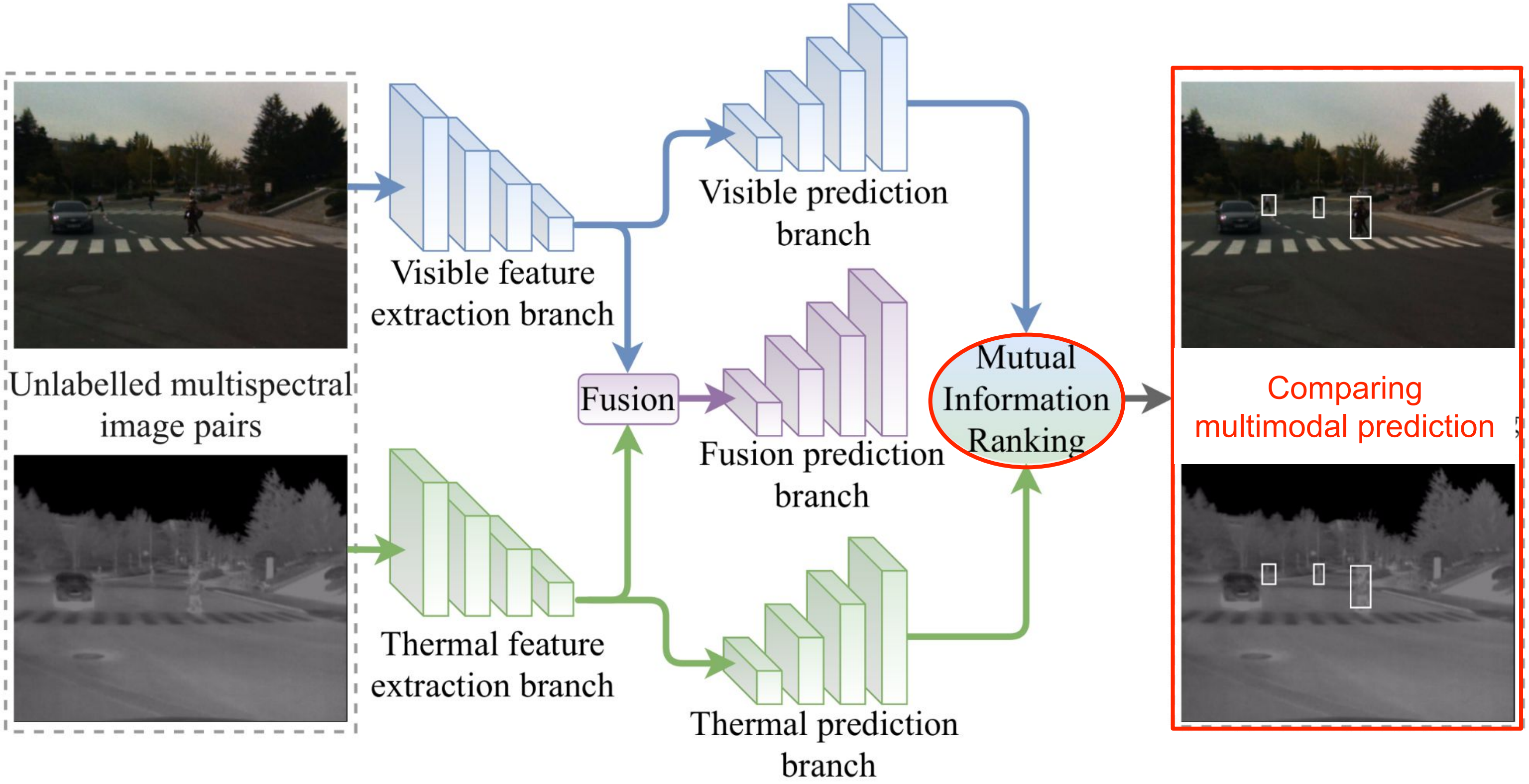
An introduction to Active learning



Active learning loop diagram

- The model inference is performed on the unlabelled dataset to select the most informative samples (i.e., multispectral image pairs in our work).
- 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.

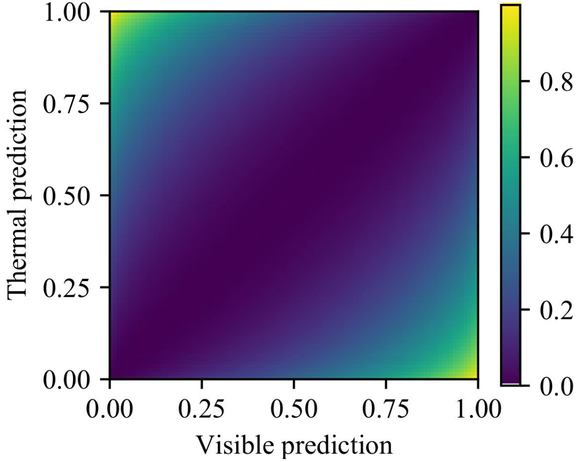
Overview of the proposed model



Cross-modality prediction inconsistency

For each prediction p, its **inconsistency score** is defined as:

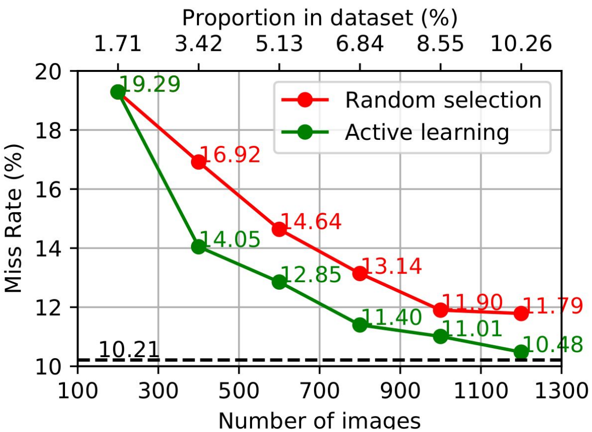
$$\mathcal{I} = \mathcal{H}\left(\overline{p}\right) - \frac{1}{2} \sum_{m \in \{v,t\}} \mathcal{H}\left(p_m\right)$$



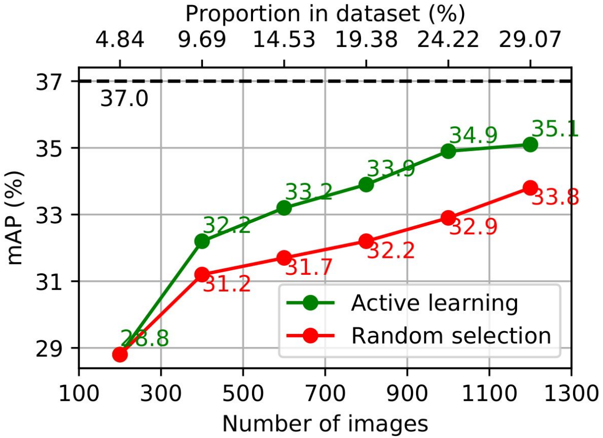
where p_v and p_t denote the prediction from visible and thermal prediction branches; $\overline{p}$ is the average of both predictions; is the $\mathcal{H}$ set entropy function calculated as:

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

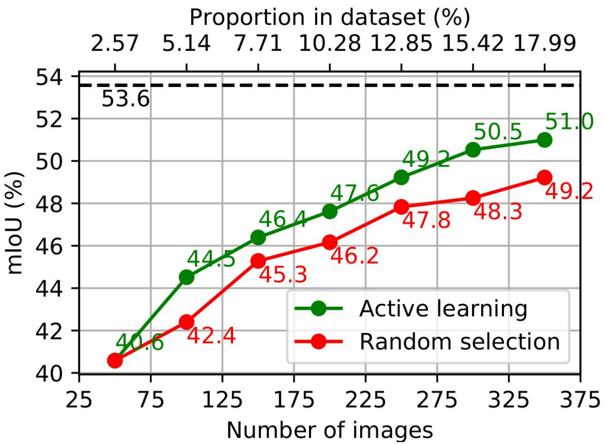
Experiments (Active VS Random)



KAIST dataset for
multispectral pedestrian detection



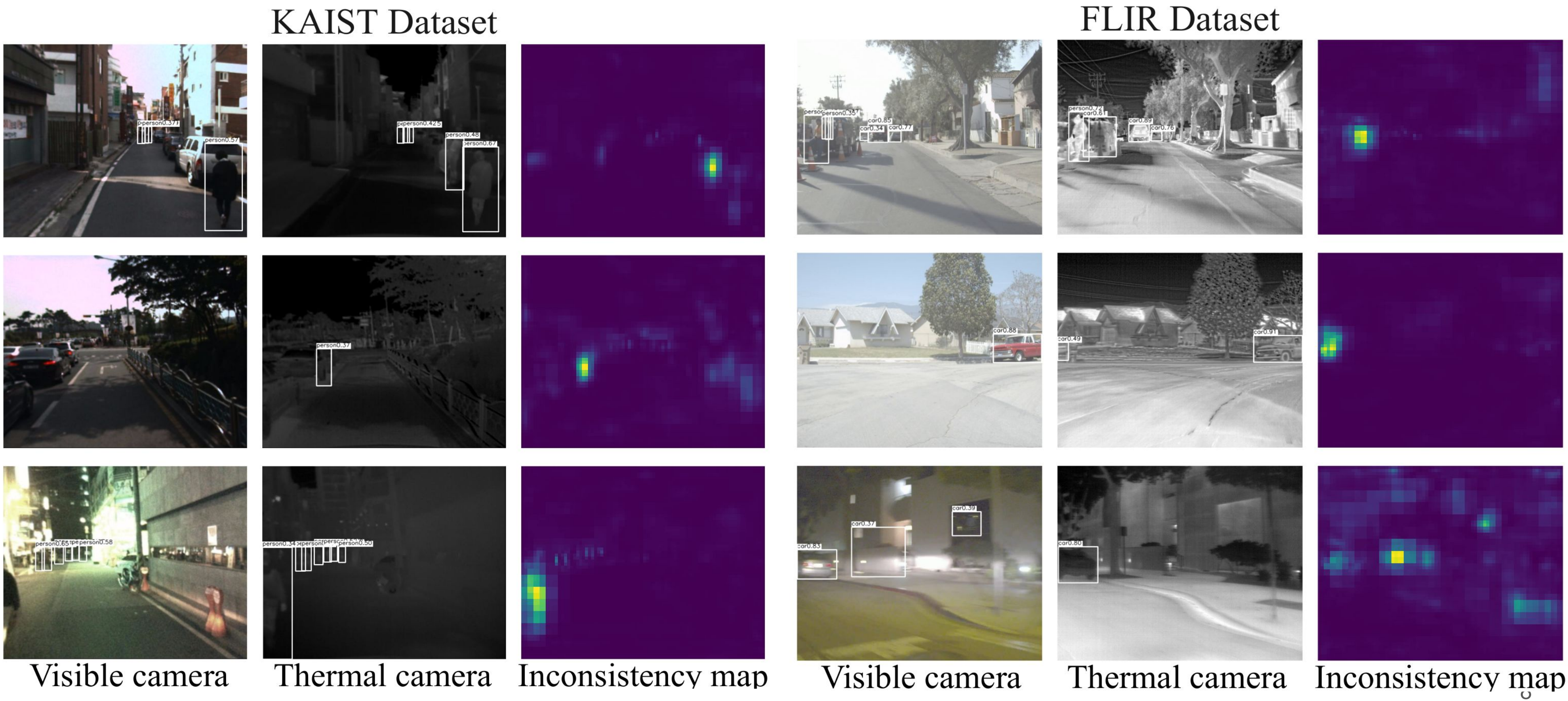
FLIR dataset for
multispectral object detection



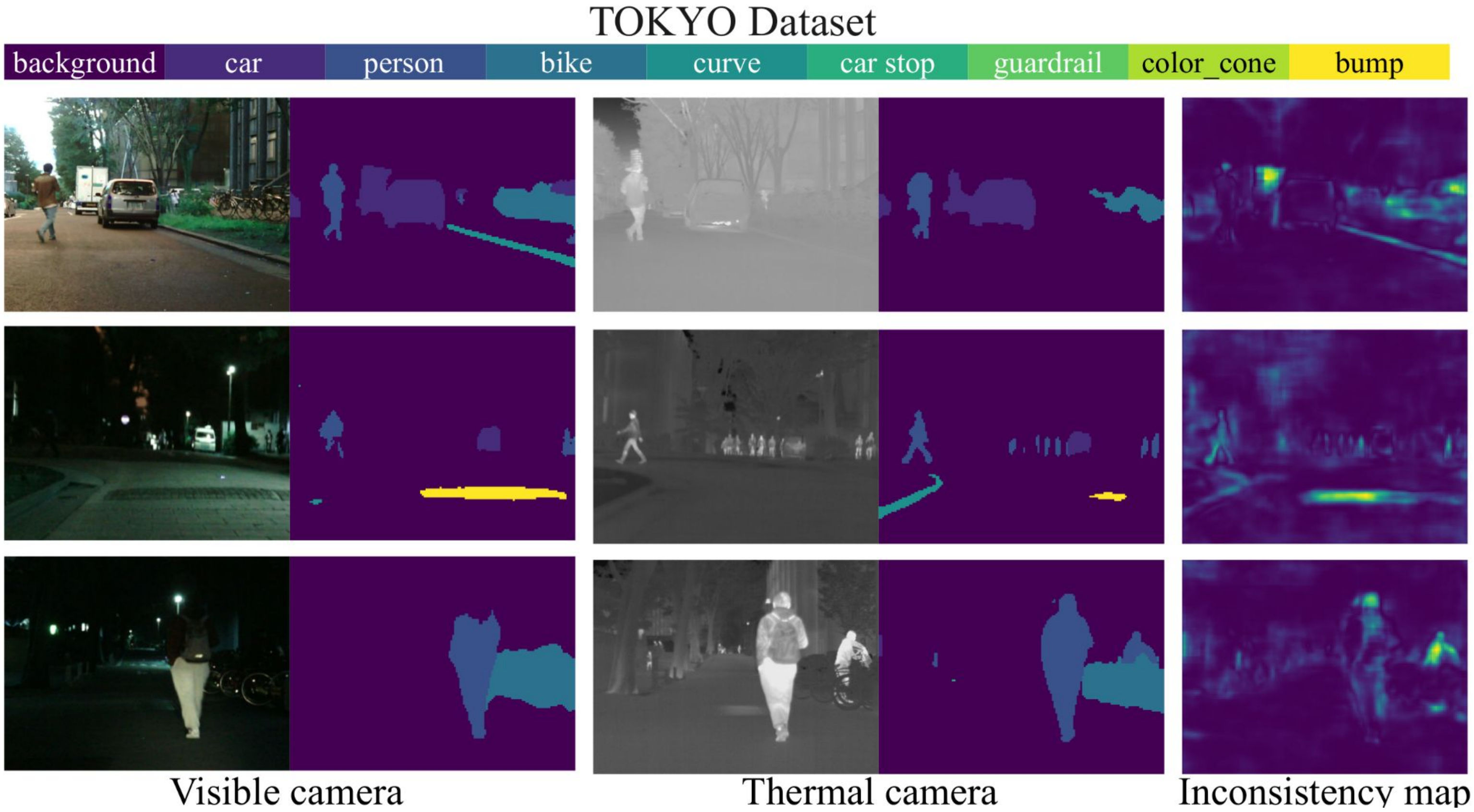
TOKYO dataset for
multispectral semantic segmentation

Observation: our **active strategy** achieves statistically significant better performance than the **random strategy** for all multispectral scene analysis tasks.

Inconsistency visualization (I)



Inconsistency visualization (II)

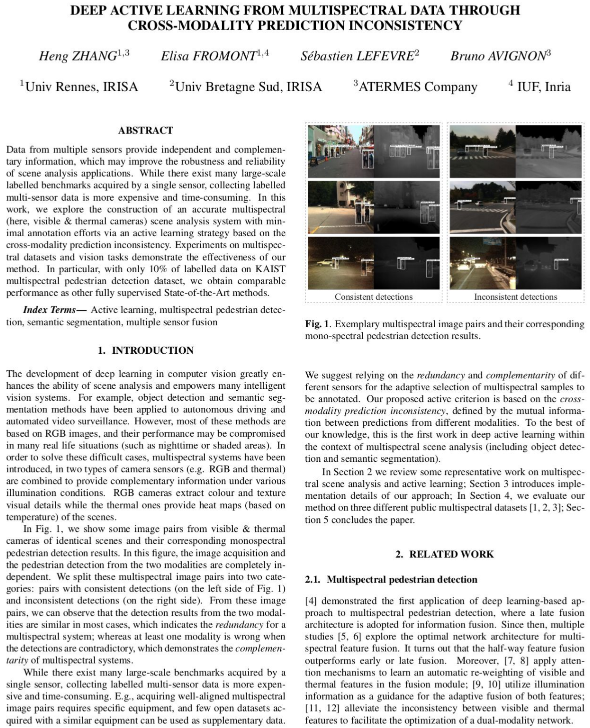


Thanks for your attention

Heng ZHANG, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

IRISA Laboratory, ATERMES Company

{heng.zhang, elisa.fromont, sebastien.lefevre}@irisa.fr bavignon@atermes.fr



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.

Index Terms— Active learning, multispectral pedestrian detection, semantic segmentation, multiple sensor fusion

1. INTRODUCTION

The development of deep learning in computer vision greatly enhances the ability of scene analysis and empowers many intelligent vision systems. For example, object detection and semantic segmentation methods have been applied to autonomous driving and automated video surveillance. However, most of these methods are based on RGB images, and their performance may be compromised in many real life situations (such as nighttime or shaded areas). In order to solve these difficult cases, multispectral systems have been introduced, in two types of camera sensors (e.g. RGB and thermal) are combined to provide complementary information under various illumination conditions. RGB cameras extract colour and texture visual details while the thermal ones provide heat maps (based on temperature) of the scenes.

In Fig. 1, we show some image pairs from visible & thermal cameras of identical scenes and their corresponding monospectral pedestrian detection results. In this figure, the image acquisition and the pedestrian detection from the two modalities are completely independent. We split these multispectral image pairs into two categories: pairs with consistent detections (on the left side of Fig. 1) and inconsistent detections (on the right side). From these image pairs, we can observe that the detection results from the two modalities are similar in most cases, which indicates the *redundancy* for a multispectral system; whereas at least one modality is wrong when the detections are contradictory, which demonstrates the *complementarity* of multispectral systems.

While there exist many large-scale benchmarks acquired by a single sensor, collecting labelled multi-sensor data is more expensive and time-consuming. E.g., acquiring well-aligned multispectral image pairs requires specific equipment, and few open datasets acquired with a similar equipment can be used as supplementary data.

Consistent detections Inconsistent detections

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

We suggest relying on the *redundancy* and *complementarity* of different sensors for the adaptive selection of multispectral samples to be annotated. Our proposed active criterion is based on the *cross-modality prediction inconsistency*, defined by the mutual information between predictions from different modalities. To the best of our knowledge, this is the first work in deep active learning within the context of multispectral scene analysis (including object detection and semantic segmentation).

In Section 2 we review some representative work on multispectral scene analysis and active learning; Section 3 introduces implementation details of our approach; In Section 4, we evaluate our method on three different public multispectral datasets [1, 2, 3]; Section 5 concludes the paper.

2. RELATED WORK

2.1. Multispectral pedestrian detection

[4] demonstrated the first application of deep learning-based approach to multispectral pedestrian detection, where a late fusion architecture is adopted for information fusion. Since then, multiple studies [5, 6] explore the optimal network architecture for multispectral feature fusion. It turns out that the half-way feature fusion outperforms early or late fusion. Moreover, [7, 8] apply attention mechanisms to learn an automatic re-weighting of visible and thermal features in the fusion module; [9, 10] utilize illumination information as a guidance for the adaptive fusion of both features; [11, 12] alleviate the inconsistency between visible and thermal features to facilitate the optimization of a dual-modality network.