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

Heng Zhang, Elisa Fromont, Sébastien Lefevre, Bruno Avignon

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 multiple 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.

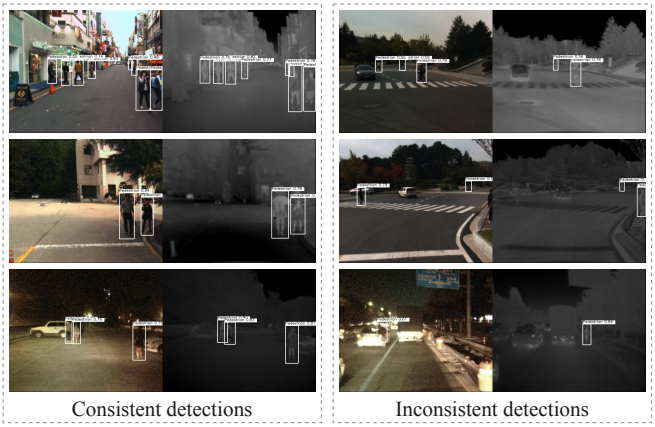


Fig. 1. Exemplary multispectral image pairs and their corresponding mono-spectral pedestrian detection results.

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. 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.

2.2. Multispectral semantic segmentation

MFNet [3] employs two identical backbone networks for visible and thermal feature extraction and a short-cut block to concatenate the extracted features. Based on that, RTFNet [13] integrates residual layers into this network architecture to further boost the performance. FuseNet [14] adopts a similar double feature extraction network for RGB-D semantic segmentation. In this paper, we replace its RGB-D input images by multispectral images for comparison.

2.3. Active learning

Labelled data are critical for today's supervised deep learning applications. Active learning, which aims to relieve human labelling efforts, is thus particularly appealing. The active learning protocol usually starts by pre-training a model on a small subset of the labelled dataset D_l. Then, several active learning cycles are repeated. Fig. 2(a) illustrates a typical active learning cycle. The model inference is performed on the unlabelled dataset D_u 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 D_l, where the model is consequently fine-tuned on. The most important component of an active learning cycle is the scoring function which ranks the informativeness of unlabelled samples.

Most studies on deep active learning in computer vision are based on mono-modal RGB images, including the most recent ones in deep active learning for object detection [15, 16] and semantic segmentation [17]. *Conversely to these existing works that score the informativeness of a single image, we aim to score a pair of multispectral images according to their relationships.* Our work can be seen as complementary to existing methods, and coupling intra-modality (as done with existing methods) and inter-modality (as proposed here) informativeness scoring could lead to further improvements than what is presented in this paper.

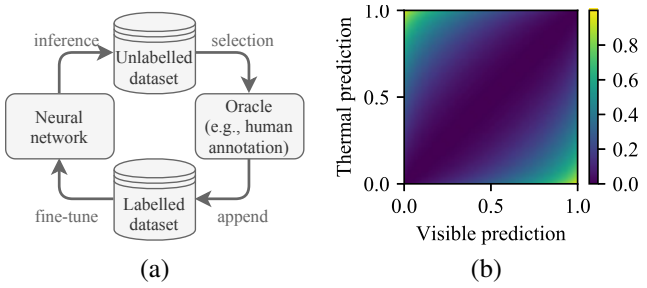


Fig. 2. Active learning loop diagram (a) and cross-modality prediction inconsistency visualization (b).

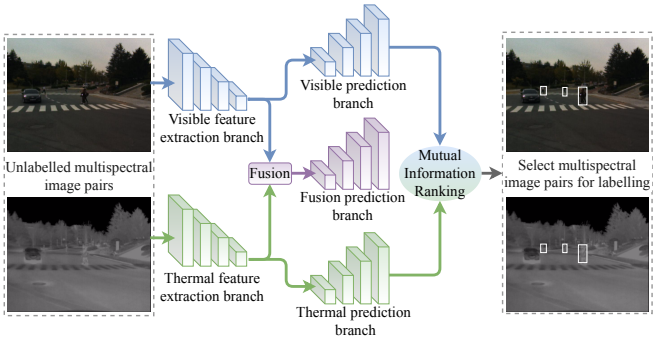


Fig. 3. Overview of the proposed model for deep active multispectral scene analysis. The blue and green mono-modal branches are used for data informativeness ranking while the purple one provides the final detection results.

3. APPROACH

3.1. Overview

An overview of our network architecture is given in Fig. 3. It takes a spatially-aligned multispectral image pair as input, then visible and thermal features are extracted independently via the modality-specific feature extraction networks. Afterwards, three prediction branches are used: one based on visible features, one based on thermal features, and the last one based on fused features. These three prediction branches are jointly optimized during the model training phase. Note that in Fig. 3 the prediction networks are used for a pedestrian detection task but can be adapted to other vision tasks such as general object detection or semantic segmentation.

3.2. Cross-modality prediction inconsistency

At the selection stage of each active learning cycle, we measure the relevance of labelling a particular image pair by ranking the aforementioned *cross-modality prediction inconsistency*, i.e., we compare predictions from visible and thermal cameras, then select for labelling the image pairs with the highest prediction difference. More specifically, for each pre-

diction p, its inconsistency 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; $\mathcal{H}$ is the 2-set entropy function calculated as:

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

For a better understanding of this inconsistency calculation, we plot in Fig. 2(b) the visualization of the inconsistency score with different visible (x-axis) and thermal (y-axis) prediction scores. It can be observed that this inconsistency score varies from 0 (very consistent) to 1 (very different).

3.3. Scale-balanced inconsistency aggregation

After obtaining the inconsistency for one prediction (i.e. classification of an anchor box in object detection or classification of a pixel in semantic segmentation), we adopt the scale-balanced strategy for full-images inconsistency aggregation. This is justified because recent deep learning approaches apply feature pyramid for multi-scale prediction thus, if we directly average all predictions for a given image pair, the inconsistency estimation will be dominated by the scale with the most predictions (i.e., the largest feature map in a feature pyramid). Therefore, we first separately average the inconsistency for each pyramid scale, then average the averaged inconsistency across all scales.

4. EXPERIMENTS

4.1. Datasets

KAIST Dataset [1]. This well-known multispectral dataset is built for the *pedestrian detection* task. In order to tackle the misalignment problem between visible-thermal image pairs, [18] proposes the "paired" annotations by separately relabelling pedestrians for each modality. We remove unpaired images according to the matching of visible and thermal annotations, thus keeping 11,695 images for training. For a fair comparison with other State-of-the-Art methods, we evaluate our model with the Miss Rate metric (lower is better) under the "reasonable" setting, i.e., a test set that does not contain heavily/partially occluded pedestrians or pedestrians smaller than 55 pixels.

FLIR Dataset [2]. This thermal dataset is released for *general object detection* from thermal images within the Advanced Driver Assistance Systems (ADAS) context. Three categories are involved: car, pedestrian and bicycle. [11] proposes the multispectral version of FLIR dataset [1] by manually aligning corresponding colour-thermal image pairs, resulting in 4,128 multispectral pairs for training. The usual mean Average Precision (mAP) metric is applied for evaluation.

TOKYO Dataset [3]. This dataset provides labelled multispectral image pairs for *semantic segmentation* within the ADAS context. It contains nine hand-labelled classes. It includes 2,338 multispectral pairs in total. Visible and thermal images are again well aligned. The mean Intersection over Union (mIoU) metric is adopted for evaluation.

[1]This aligned dataset can be downloaded here: http://shorturl.at/ahAY4

4.2. Implementation details

Network architecture. We adopt VGG16 [19] as the feature extraction network, GAFF [8] as the multispectral feature fusion network and SSD [20] as the prediction network for the object detection tasks. For the semantic segmentation task, the prediction branch is simply one layer of convolution whose number of output channels is equal to the number of classes. In order not to change the aspect ratio of the original images, input images are resized to 480×384 or 640×512 for KAIST and FLIR datasets (object detection) and 640×480 for TOKYO dataset (semantic segmentation). Random cropping, expanding, flipping are adopted for data augmentation.

Active learning setting. For each active learning experiment, we first randomly initialize a labelled dataset D_l with b images and pretrain the model on D_l; then we actively select b images from an unlabelled dataset D_u with the most cross-modality prediction inconsistency $\mathcal{I}$ for annotation and add these newly labelled images into D_l; afterwards we fine-tune the model with the new D_l; we repeat the previous two steps until the annotation budget B is exhausted. Since semantic segmentation annotations are more difficult to acquire, we set b to 200 and B to 1200 for the object detection tasks, b to 50 and B to 350 for the semantic segmentation task.

4.3. Results

Active vs Random. Fig. 4 plots the performance evolutions along all learning cycles for KAIST Dataset (subfigure a and b), FLIR Dataset (c and d) and TOKYO Dataset (e and f). For all multispectral datasets, all tasks, all evaluation metrics and all input resolutions, our active strategy (green lines in the figure) achieves statistically significant better performance than the random strategy (red lines).

Active vs SotA. We list in Tables 1, 2 and 3 the comparisons between our active learning results and other State-of-the-Art methods for each multispectral dataset. With a small quantity of labelled data, our active models achieve comparable results with fully supervised SotA methods, which demonstrates the effectiveness of the proposed method.

4.4. Visualization

We show in Fig. 5 some image pairs selected by our active method. For each dataset, we plot the separate predictions from the visible or thermal cameras, and their cross-modality

inconsistency map: our strategy does select some difficult cases where at least one modality makes mistakes. We believe that adding these informative examples into the labelled dataset for fine-tuning is the main reason for improvements.

5. 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 public multispectral datasets and two scene analysis tasks demonstrate the effectiveness of the proposed method. To the best of our knowledge, our work is the first applying deep active learning for multispectral scene analysis. We hope that our method could help reduce manual labelling efforts when setting up multispectral or multi-sensor datasets.

Methods	Miss Rate (lower, better)		
	All	Day	Night
ACF [1]	47.32%	42.57%	56.17%
Halfway Fusion [21]	25.75%	24.88%	26.59%
Fusion RPN+BF [5]	18.29%	19.57%	16.27%
IAF R-CNN [10]	15.73%	14.55%	18.26%
IATDNN+IASS [9]	14.95%	14.67%	15.72%
CIAN [7]	14.12%	14.77%	11.13%
MSDS-RCNN [6]	11.34%	10.53%	12.94%
AR-CNN [18]	9.34%	9.94%	8.38%
MBNet [12]	8.13%	8.28%	7.86%
Ours (full dataset)	8.86%	10.01%	6.77%
Ours (10.26% of data)	9.32%	10.13%	7.70%

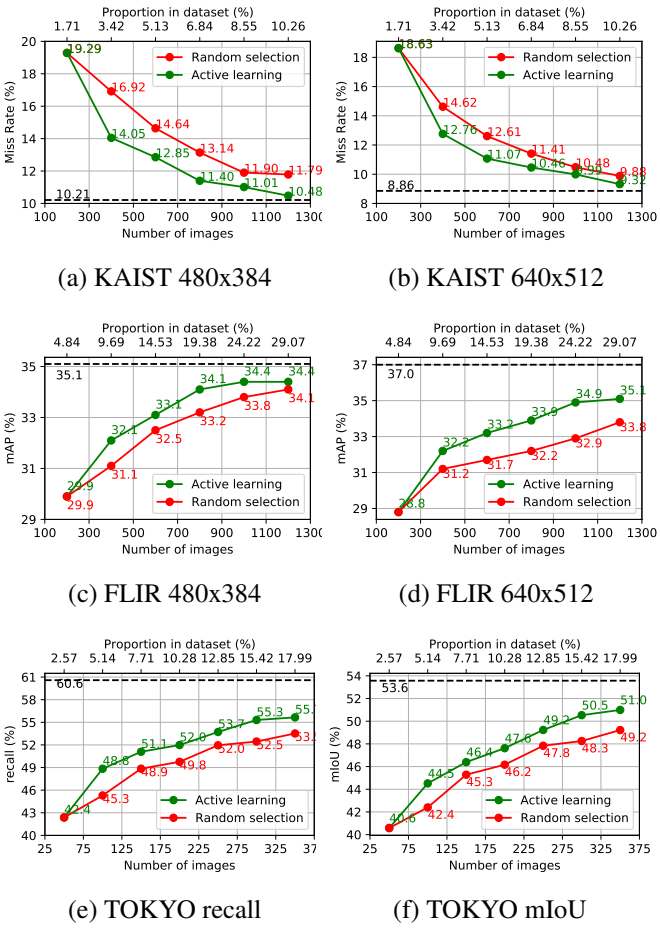
Table 1. Miss Rate comparisons on KAIST Dataset.

Methods	mAP	AP50	AP75
CFR [11]	-	72.4%	-
GAFF [8]	37.3%	72.7%	30.9%
Ours (full dataset)	37.0%	72.1%	31.2%
Ours (29.07% of data)	35.1%	71.0%	30.6%

Table 2. mAP comparisons on FLIR Dataset.

Methods	mIoU (higher, better)		
	All	Day	Night
MFNet [3]	39.7%	36.1%	36.8%
FuseNet [14]	45.6%	41.0%	43.9%
RTFNet [13]	53.2%	45.8%	54.8%
Ours (full dataset)	53.6%	46.8%	53.3%
Ours (17.99% of data)	51.0%	46.6%	48.9%

Table 3. mIoU comparisons on TOKYO Dataset.



(a) KAIST 480x384 (b) KAIST 640x512

(c) FLIR 480x384 (d) FLIR 640x512

(e) TOKYO recall (f) TOKYO mIoU

Fig. 4. Experimental results of models trained by the proposed active learning strategy (green lines) and random selection strategy (red lines) on KAIST Dataset (a, b), FLIR Dataset (c, d) and TOKYO Dataset (e, f). Black dotted lines indicate fully supervised results.

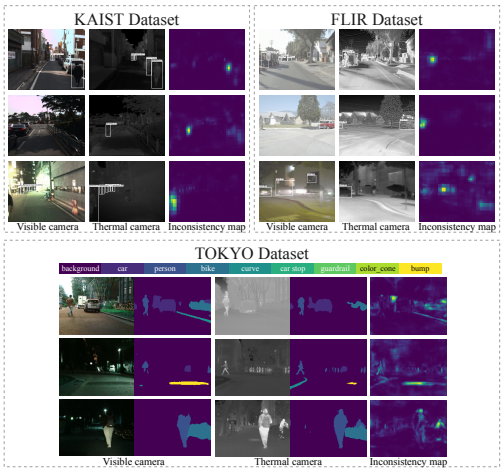


Fig. 5. Examples of selected image pairs for labelling by the proposed method. Zoom in to see details.

6. REFERENCES

[1] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baselines," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] "Free flir thermal dataset for algorithm training," `https://www.flir.com/oem/adas/adas-dataset-form/`.

[3] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[4] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *24th European Symposium on Artificial Neural Networks, (ESANN)*, 2016.

[5] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, 2017.

[6] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *British Machine Vision Conference (BMVC)*, 2018.

[7] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain, "Cross-modality interactive attention network for multispectral pedestrian detection," *Information Fusion*, 2019.

[8] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021.

[9] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, 2019.

[10] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognition*, 2019.

[11] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon, "Multispectral Fusion for Object Detection with Cyclic Fuse-and-Refine Blocks," in *ICIP 2020 - IEEE International Conference on Image Processing*, 2020.

[12] Kailai Zhou, Linsen Chen, and Xun Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *European Conference on Computer Vision (ECCV)*, 2020.

[13] Yuxiang Sun, Weixun Zuo, and Ming Liu, "RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, July 2019.

[14] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian Conference on Computer Vision*, November 2016.

[15] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu, "Localization-aware active learning for object detection," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 506–522.

[16] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López, "Active learning for deep detection neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[17] Arantxa Casanova, Pedro O. Pinheiro, Negar Rostamzadeh, and Christopher J. Pal, "Reinforced active learning for image segmentation," in *International Conference on Learning Representations*, 2020.

[18] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *International Conference on Computer Vision*, 2019.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg, "SSD: single shot multibox detector," in *European Conference on Computer Vision (ECCV)*, 2016.

[21] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.