The following publication Tianshan Liu, Rui Zhao, and Kin-Man Lam "Attention-based cross-modality interaction for multispectral pedestrian detection", Proc. SPIE 11766, International Workshop on Advanced Imaging Technology (IWAIT) 2021, 1176603 (13 March 2021) is available at https://doi.org/10.1117/12.2590661.

Attention-based Cross-modality Interaction for Multispectral Pedestrian Detection

Tianshan Liu, Rui Zhao and Kin-Man Lam

Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China

ABSTRACT

Multispectral pedestrian detection has attracted extensive attention, as paired RGB-thermal images can provide complementary patterns to deal with illumination changes in realistic scenarios. However, most of the existing deep-learning-based multispectral detectors extract features from RGB and thermal inputs separately, and fuse them by a simple concatenation operation. This fusion strategy is suboptimal, as undifferentiated concatenation for each region and feature channel may hamper the optimal selection of complementary features from different modalities. To address this limitation, in this paper, we propose an attention-based cross-modality interaction (ACI) module, which aims to adaptively highlight and aggregate the discriminative regions and channels of the feature maps from RGB and thermal images. The proposed ACI module is deployed into multiple layers of a two-branch-based deep architecture, to capture the cross-modal interactions from diverse semantic levels, for illumination-invariant pedestrian detection. Experimental results on the public KAIST multispectral pedestrian benchmark show that the proposed method achieves state-of-the-art detection performance.

Keywords: Multispectral pedestrian detection, attention mechanism, cross-modal representation

1. INTRODUCTION

Pedestrian detection plays a crucial role in the computer vision community, due to its widespread real-world applications,[1] such as autonomous driving vehicles, video surveillance, human activity understanding, etc. Over the past decade, with the success of deep-learning-based approaches in object detection,[2,3] remarkable progress has been made on the performance of pedestrian detectors. Nevertheless, most of the current pedestrian detection methods are trained using the visible images captured in good illumination conditions. Since real-world scenes usually involve a wide range of illumination variations,[4] designing an illumination-invariant pedestrian detection model is still a challenging task.

Recently, multispectral pedestrian detection has attracted more and more attention,[5] as the thermal images are insensitive to illumination changes and can provide additional information from the scenes complementary with RGB images. Thus, there have been numerous attempts to explore paired visible (RGB)-thermal images, for illumination-invariant pedestrian detection. Since RGB and thermal data exhibit different characteristics under various illumination conditions, adaptive fusion of these two modalities can facilitate robust performance in the pedestrian-detection task. However, the majority of existing multispectral pedestrian detectors[6–8] mainly investigate different fusion stages, and concatenate the features extracted from RGB and thermal images in a simple way, without exploring the interactions between them. This often restricts the detection performance of these methods.

Attention mechanism has been widely studied, and is embedded into deep-learning networks, to spotlight the features or regions, which are most discriminative for various visual tasks. Hu et al.[9] proposed a Squeeze-and-Excitation (SE) block to recalibrate channel-wise features by exploring the dependencies between different channels. Wang et al.[10] presented a non-local operation leveraging the relationships between each pair of points. The attention-based modules have also been investigated to learn context information for multimodal tasks. Liu

Corresponding author: Tianshan Liu
Tianshan Liu: E-mail: tianshan.liu@connect.polyu.hk
Rui Zhao: E-mail: rick10.zhao@connect.polyu.hk
Kin-Man Lam: E-mail: enkmlam@polyu.edu.hk

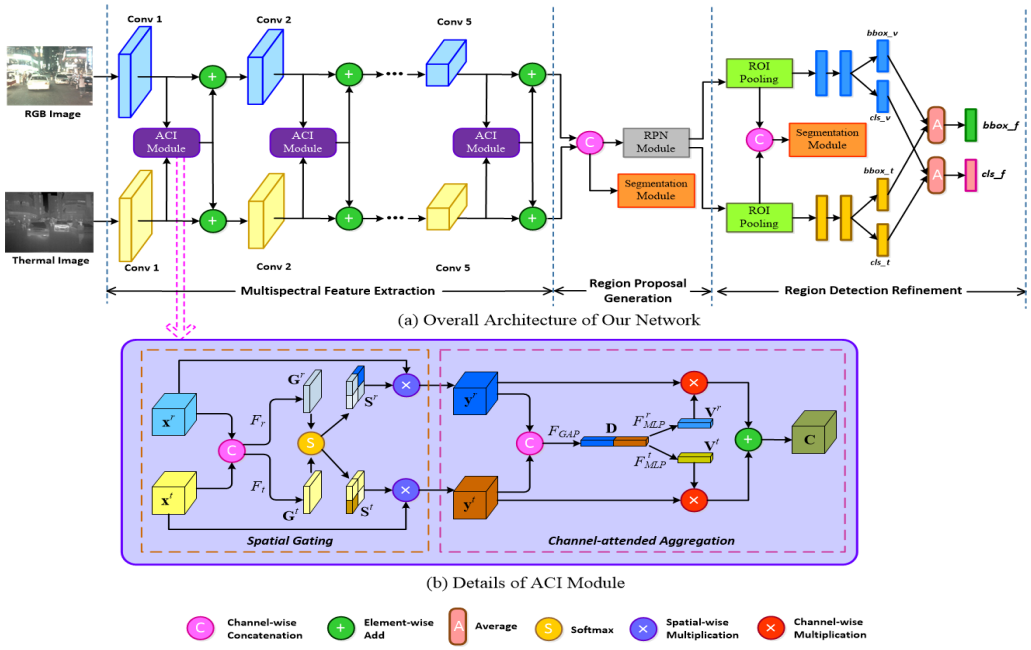


Figure 1. (a) The overall architecture of our multispectral pedestrian-detection network. (b) Schematic diagram of the proposed attention-based cross-modality interaction (ACI) module.

et al.[11] proposed a selective self-attention module by using multimodal information, and integrated the attention module into a two-branch Convolutional Neural Network (CNN), for RGB-D saliency detection. Chen et al.[12] employed the channel-wise attention mechanism to adaptively select features from RGB and depth modalities, and further fused them for salient object detection.

In order to highlight and aggregate the discriminative regions and channels of the feature maps from RGB and thermal modalities, we propose an Attention-based Cross-modality Interaction (ACI) module for multispectral pedestrian detection. The ACI module is designed based on a two-stage strategy. In the first stage, we employ a spatial-wise attention mechanism to generate modality-specific gates, which assign different importance weights for each modality at each position. In the second stage, we compute two channel-attended vectors via a channel-wise attention mechanism, and utilize them to aggregate the spatially gated feature maps obtained in the first stage. The proposed ACI module is deployed into multiple layers of the backbone CNN, to leverage the cross-modal representations from diverse semantic levels. The whole multispectral pedestrian detection network consists of two sub-branches, which take RGB images and thermal images as input, respectively. Each branch is built based on a Region Proposal Network (RPN)-based model, i.e., Faster R-CNN.[2] Motivated by Brazil et al.,[13] we also introduce a segmentation-based auxiliary task by leveraging box-based pedestrian masks. The overall network architecture is illustrated in Fig. 1 (a).

The main contributions of this paper are summarized as follows. First, we propose an Attention-based Cross-modality Interaction (ACI) module, which learns robust cross-modal representations by aggregating the discriminative regions and feature channels from RGB and thermal modalities. Second, we deploy the proposed ACI module into multiple layers of a two-branch-based architecture, which adequately captures the interactions between these two modalities in diverse semantic levels for pedestrian detection. Third, experimental results on the public KAIST multispectral pedestrian benchmark validate the effectiveness of the proposed method.

2. METHODOLOGY

2.1 Network Architecture

As illustrated in Fig. 1 (a), the overall network consists of two sub-branches, which take RGB images and thermal images as input, respectively. Each branch network is based on a two stage-based model, i.e., Faster

R-CNN. The VGG-16[14] is chosen as the backbone CNN to extract convolutional features for each branch. The proposed ACI module aims to highlight the most discriminative regions and channels of each modality, and fuses the weighted features to construct better cross-modal representations. We deploy the proposed ACI module after each convolutional layer in the backbone CNN, to fuse the feature maps from each modality in various semantic levels. The learned cross-modal representations can provide discriminative cues for the subsequent detection tasks. Moreover, to improve the learning capacity of the trunk network, we introduce an additional segmentation task in both of the two stages, i.e., region proposal generation (RPG) and region detection refinement (RDR). We leverage weakly-supervised box-based pedestrian masks rather than the pixel-wise annotations, to provide extra supervision signals to guide the learning of the network. The segmentation module is implemented as a shallow layer, with 1×1 convolutional kernels.

The overall objective function is formulated as follows:

$$\begin{aligned} L_{RPG} &= L_{cls}^{RPG} + \lambda L_{reg}^{RPG} + \gamma L_{seg}^{RPG}, \\ L_{RDR} &= L_{cls}^{RDR} + \lambda L_{reg}^{RDR} + \gamma L_{seg}^{RDR}, \\ L &= L_{RPG} + L_{RDR}, \end{aligned} \tag{1}$$

where L_{cls} denotes the classification loss function based on cross-entropy loss, L_{reg} is the bounding-box regression loss function based on smoothed L1 loss, and L_{reg} is the segmentation loss based on pixel-wise cross-entropy loss. The parameters λ and γ balance the importance between the different losses. We set them to be 1, throughout all the experiments.

2.2 Attention-based Cross-modality Interaction Module

The discrimination of the regions and channels of the feature maps from each modality may vary in different scenarios. For instance, in good illumination conditions, e.g., during daytime, the colour and texture cues in RGB images provide more informative patterns for pedestrian detection. In poor illumination conditions, e.g., during nighttime, the human silhouettes in thermal images are more distinct and reliable. Therefore, to localize the discriminative regions and feature channels of each modality, we propose the ACI module, by leveraging both spatial-attention and channel-attention mechanisms. As shown in Fig. 1 (b), given two feature maps extracted from RGB and thermal images, denoted as $\mathbf{x}^r \in \mathbb{R}^{C \times W \times H}$ and $\mathbf{x}^t \in \mathbb{R}^{C \times W \times H}$, respectively, where C represents the number of channels, and W and H are the width and height, respectively.

Spatial Gating. We weight the features of different modalities at each position according to their representational capabilities, by utilizing a spatial-attention strategy. Specifically, the feature maps from these two modalities are first concatenated along the channel dimension. Then, we map the concatenated feature to two modality-specific spatial-wise gates separately, as follows:

$$\mathbf{G}^r = F_r\left(\mathbf{x}^r || \mathbf{x}^t\right), \tag{2}$$

$$\mathbf{G}^t = F_t\left(\mathbf{x}^r || \mathbf{x}^t\right), \tag{3}$$

where $||$ denotes the channel-wise concatenation operation, $\mathbf{G}^r \in \mathbb{R}^{1 \times W \times H}$ and $\mathbf{G}^t \in \mathbb{R}^{1 \times W \times H}$ are the spatial-wise gates for RGB and thermal modalities, respectively. $F_r(\cdot)$ and $F_t(\cdot)$ are two individual mapping functions implemented with 1×1 convolutional kernels. A softmax operation is further applied on both of these gates, as follows:

$$\mathbf{S}_{(i,j)}^r = \frac{e^{\mathbf{G}_{(i,j)}^r}}{e^{\mathbf{G}_{(i,j)}^r} + e^{\mathbf{G}_{(i,j)}^t}}, \mathbf{S}_{(i,j)}^t = \frac{e^{\mathbf{G}_{(i,j)}^t}}{e^{\mathbf{G}_{(i,j)}^r} + e^{\mathbf{G}_{(i,j)}^t}}, \tag{4}$$

where $\mathbf{S}_{(i,j)}^r$ and $\mathbf{S}_{(i,j)}^t$ can be regarded as the weights assigned to each position in the RGB and thermal feature maps, respectively. The spatially gated feature maps are generated by applying the modality-specific gates to the original inputs, using element-wise multiplication for each channel, as follows:

$$\mathbf{y}_{(i,j)}^r = \mathbf{S}_{(i,j)}^r \odot \mathbf{x}_{(i,j)}^r, \tag{5}$$

$$\mathbf{y}_{(i,j)}^t = \mathbf{S}_{(i,j)}^t \odot \mathbf{x}_{(i,j)}^t. \tag{6}$$

Channel-attended Aggregation. We further assign importance weights to each channel of features from different modalities, by employing a channel-attention mechanism. Specifically, after obtaining the spatially gated feature maps $\mathbf{y}^r \in \mathbb{R}^{C\times W\times H}$ and $\mathbf{y}^t \in \mathbb{R}^{C\times W\times H}$, we first concatenate them along the channel dimension and utilize a global average pooling (GAP) to compute the cross-modality global statistics, as follows:

$$\mathbf{D} = F_{GAP}\left(\mathbf{y}^r||\mathbf{y}^t\right), \tag{7}$$

where $F_{GAP}(\cdot)$ indicates the GAP operation. Then, we learn two modality-specific channel-attended vectors based on the cross-modality global descriptor $\mathbf{D} \in \mathbb{R}^{2C\times 1}$, as follows:

$$\mathbf{V}^r = \sigma\left(F^r_{MLP}(\mathbf{D})\right), \tag{8}$$

$$\mathbf{V}^t = \sigma\left(F^t_{MLP}(\mathbf{D})\right), \tag{9}$$

where $F^r_{MLP}(\cdot)$ and $F^t_{MLP}(\cdot)$ are two multi-layer perceptron (MLP) networks, for RGB and thermal modalities, respectively. σ represents the sigmoid activation function, which normalizes the weight vectors. The channel-attended vectors, $\mathbf{V}^r \in \mathbb{R}^{C\times 1}$ and $\mathbf{V}^t \in \mathbb{R}^{C\times 1}$, can highlight the discriminative features and suppress the importance of the unreliable or noisy channels. Finally, we leverage the modality-specific channel-attended vectors to aggregate the spatially gated feature maps, as follows:

$$\mathbf{C} = \mathbf{V}^r \otimes \mathbf{y}^r + \mathbf{V}^t \otimes \mathbf{y}^t \tag{10}$$

where $\otimes$ denotes the channel-wise multiplication. The cross-modal representation $\mathbf{C} \in \mathbb{R}^{C\times W\times H}$, which is robust to di?erent illumination conditions, fuses the feature maps of RGB and thermal modalities by exploiting the most discriminative regions and channels.

Propagation Strategy. The proposed ACI module is inserted after each convolutional layer in the backbone CNN. For each layer l, we employ the cross-modal representation $\mathbf{C}_l$, generated by the l-th ACI module, to refine the original feature map of the l-th convolutional layer, as follows:

$$\mathbf{z}^r_l = \frac{\mathbf{x}^r + \mathbf{C}_l}{2}, \tag{11}$$

$$\mathbf{z}^t_l = \frac{\mathbf{x}^t + \mathbf{C}_l}{2}. \tag{12}$$

The refined outputs $\mathbf{z}^r_l$ and $\mathbf{z}^t_l$ are fed into the $l+1$-st layer in the RGB and thermal branches, respectively, for propagation. This is motivated by the residual learning.[15]

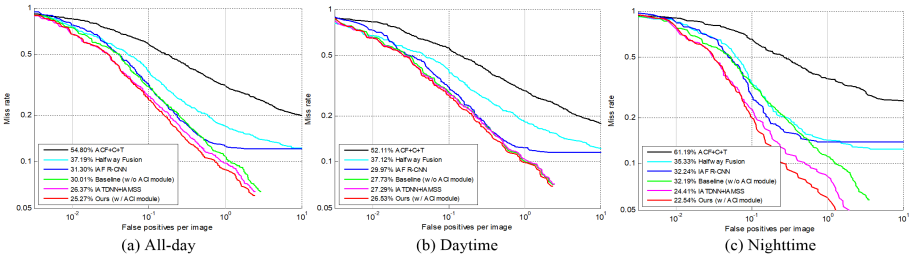
3. EXPERIMENTS

3.1 Data Set

The proposed method is evaluated on the public KAIST multispectral pedestrian data set.[5] The KAIST dataset consists of 50,172 aligned RGB-thermal image pairs, which are captured by visible and thermal cameras, under different illumination conditions. We use 25,086 multispectral image pairs for training, by following the works in.[6] The testing set includes 2,252 pairs of RGB-thermal images, in which 1,455 and 797 pairs are captured during daytime and nighttime, respectively. We follow the reasonable evaluation settings presented in.[5] We adopt the log-average miss rate (MR) to evaluate the performance of different pedestrian detection approaches. The miss rate is averaged over the false positives per image (FPPI) in the range of $[10^{-2}, 10^{0}]$.

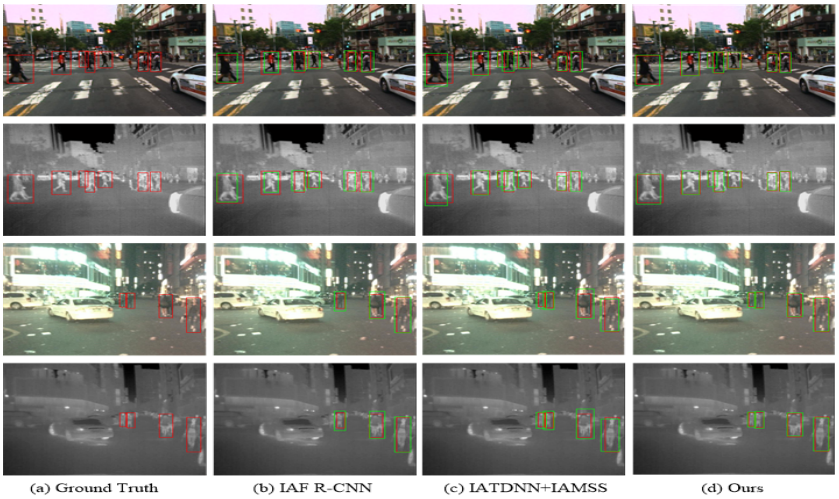
3.2 Implementation Details

We initialize the convolutional layers conv1-5 in the backbone-CNN VGG-16, with the weights pre-trained on ImageNet. The other layers and modules are initialized with random a Gaussian distribution. An anchor is considered to be pedestrian (positive), when its Intersection over Union (IoU) satisfies $IoU > 0.5$. The ratio of positive and negative anchors is 1:3. The batch size is set to be 64. For the training of the proposed multispectral pedestrian detection network, we adopt the ADAM optimization algorithm. The learning rate is initialized at 0.001, and decayed by 0.1 after 50 and 100 epochs, with a total of 150 epochs.



(a) All-day (b) Daytime (c) Nighttime

Figure 2. Miss rate curves of different methods on the KAIST data set under the reasonable evaluation settings.



(a) Ground Truth (b) IAF R-CNN (c) IATDNN+IAMSS (d) Ours

Figure 3. Sample detection results of the evaluated methods on the KAIST pedestrian data set.

3.3 Comparison with State-of-the-Art Methods

The proposed method is compared with several state-of-the-art multispectral pedestrian detectors. The curves of MR against FPPI are shown in Fig. 2. The comparison methods include hand-crafted descriptors, such as the ACF+C+T,[5] deep-learning-based architectures, such as Halfway Fusion,[16] IAF R-CNN[7] and IATDNN+IAMSS.[8] To evaluate the effectiveness of the proposed method, we also implement a baseline model without using the attention-based cross-modality interaction (ACI) module. We can find that the proposed model outperforms the other four state-of-the-art methods, in all of the all-day, daytime and nighttime evaluation settings. The state-of-the-art detectors listed in Fig. 2 fuse the features from different modalities in a holistic manner, without fully exploring the complementarity of RGB and thermal images. In contrast, the proposed ACI module can localize the most discriminative regions and channels of the feature maps from each modality, and adaptively fuse them to generate robust cross-modal representations. Moreover, we insert the ACI module into multiple layers of the network to leverage the cross-modal representations from diverse semantic levels. Compared with the baseline model, the performance gap (4.74%) in the all-day setting demonstrates the effectiveness of the proposed ACI module once again.

To further intuitively demonstrate the effectiveness of the proposed method, we visualize some detection results of the different methods in Fig. 3. The first two rows are RGB-thermal image pairs captured during daytime, and the other two rows are nighttime image pairs. The first column is the original input image pairs with ground-truth annotations, and the other three columns are the detection results generated by IAF R-CNN,[7] IATDNN+IAMSS[8] and our model, respectively. The red and green bounding boxes indicate the ground-truth annotations and predicted results, respectively. It can be observed that the proposed network achieves robust detection results under different light conditions.

4. CONCLUSION

In this paper, we propose an attention-based cross-modality interaction (ACI) module for multispectral pedestrian detection. By exploring both spatial-wise and channel-wise attention mechanisms, the proposed ACI module is capable of localizing and aggregating the discriminative regions and channels of the feature maps from RGB and thermal modalities. Moreover, we insert the proposed ACI module into multiple layers of a two-branch-based network, to learn the robust cross-modal representations from diverse semantic levels, for illumination-invariant pedestrian detection. Evaluation results on the public KAIST multispectral pedestrian data set demonstrate that our method can achieve better performance, compared with state-of-the-art detectors.

REFERENCES

1. Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**, pp. 2195–2211, Sep. 2020.
2. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**, pp. 1137–1149, June 2017.
3. N. Degawa, X. Lu, and A. Kimura, "A performance improvement of Mask R-CNN using region proposal expansion," in *International Workshop on Advanced Image Technology (IWAIT) 2019*, Q. Kemao, K. Hayase, P. Y. Lau, W.-N. Lie, Y.-L. Lee, S. Srisuk, and L. Yu, eds., **11049**, pp. 436 – 441, International Society for Optics and Photonics, SPIE, 2019.
4. L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5126–5136, Oct 2019.
5. S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1037–1045, June 2015.
6. D. Knig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 243–250, July 2017.
7. C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster r-cnn for robust multispectral pedestrian detection," *Pattern Recognition* **85**, pp. 161 – 171, 2019.
8. D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion* **50**, pp. 148 – 157, 2019.
9. J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**, pp. 2011–2023, Aug 2020.
10. X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, June 2018.
11. N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13753–13762, June 2020.
12. H. Chen and Y. Li, "Three-stream attention-aware network for rgb-d salient object detection," *IEEE Transactions on Image Processing* **28**, pp. 2825–2835, June 2019.
13. G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection and segmentation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4960–4969, Oct 2017.
14. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.
15. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.
16. J. Liu, S. Zhang, S. Wang, and D. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *British Machine Vision Conference (BMVC)*, 2016.