Received September 1, 2020, accepted September 5, 2020, date of publication September 8, 2020,
date of current version September 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3022623

Attention Based Multi-Layer Fusion of Multispectral Images for Pedestrian Detection

YONGTAO ZHANG[1,2], ZHISHUAI YIN[1,2], LINZHEN NIE[1,2], AND SONG HUANG[1]
[1]School of Automotive Engineering, Wuhan University of Technology, Wuhan 430070, China
[2]Hubei Key Laboratory of Advanced Technology for Automotive Components, Wuhan University of Technology, Wuhan 430070, China

Corresponding author: Zhishuai Yin (zyin@whut.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 51805388.

ABSTRACT Multispectral images are increasingly used for pedestrian detection. Preliminary fusion strategies would fail to exploit informative features from cross-spectral images, or worse, may introduce additional interference. In this paper, we propose an attention based multi-layer fusion network in the triple-stream deep convolutional neural network architecture for multispectral pedestrian detection. The effectiveness of multi-layer fusion is examined and verified in this work. Furthermore, a channel-wise attention module (CAM) and a spatial-wise attention module (SAM) are developed and incorporated into the network aiming at more subtle adjustment to weights of multispectral features along both the channel and spatial dimensions respectively. Channel-wise attention is trained with self-supervision while spatial-wise attention is trained with external supervision as we remodel its learning process as saliency detection. Both attention-based weighting mechanisms are evaluated separately and then sequentially. Experimental results on the KAIST dataset show that the proposed multi-layer cross-spectral fusion R-CNN (CS-RCNN), with spatial-wise weighting applied alone, achieves state-of-the-art performance on all-day detection while outperforming compared methods at nighttime.

INDEX TERMS Convolutional neural networks, pedestrian detection, image fusion, deep learning.

I. INTRODUCTION

Detecting pedestrians in real-time with high accuracy is crucial to various cutting-edge applications including autonomous driving [10], [46]. Motivated by the emergence of deep learning over recent years, major advances have been made in pedestrian detection with computer vision-based techniques. Cross-spectral fusion of visible light and infrared thermal images has become a research focus [15], [21], [27], [40], [46] for all-day pedestrian detection since multi-modal information is intuitively considered to be complementary [14].

The main challenge of multispectral fusion is the design of the fusion strategy whereby two modalities of images could be effectively and dynamically fused to achieve accurate and robust detection of pedestrians at all times. Most proposals of fusion schemes act upon features instead of raw inputs to merge more expressive information [15], [47]. However, feature fusion is often either implemented at a fixed (or sometimes arbitrary) layer of the network without exploiting abundant features at other layers, or applied at multiple layers but with preliminary weighting mechanisms, which may introduce mutual interference rather than cross-spectral complementation. Recent research efforts propose more dynamic weighting mechanisms to fine-tune weights of each spectrum adaptively according to ambient conditions, such as illumination [17], [29].

Inspired by remarkable work done in attention mechanisms [19], [50] in convolutional neural networks (CNNs), we argue that the goal of fusing multispectral features is similar to what attention is designed for: to preserve useful or important features while suppressing interfering or undesired features. Therefore, attention based weighting mechanisms, which take into account far more complex and subtle factors than single or some decoupled evident characteristics such as illumination or temperature of the scene, would be more effective in enhancing feature expressiveness.

For multispectral pedestrian detection, we propose a triple-stream and multi-layer fusion network based on deep CNN in this study (Fig. 1). Both thermal and visible streams share the

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

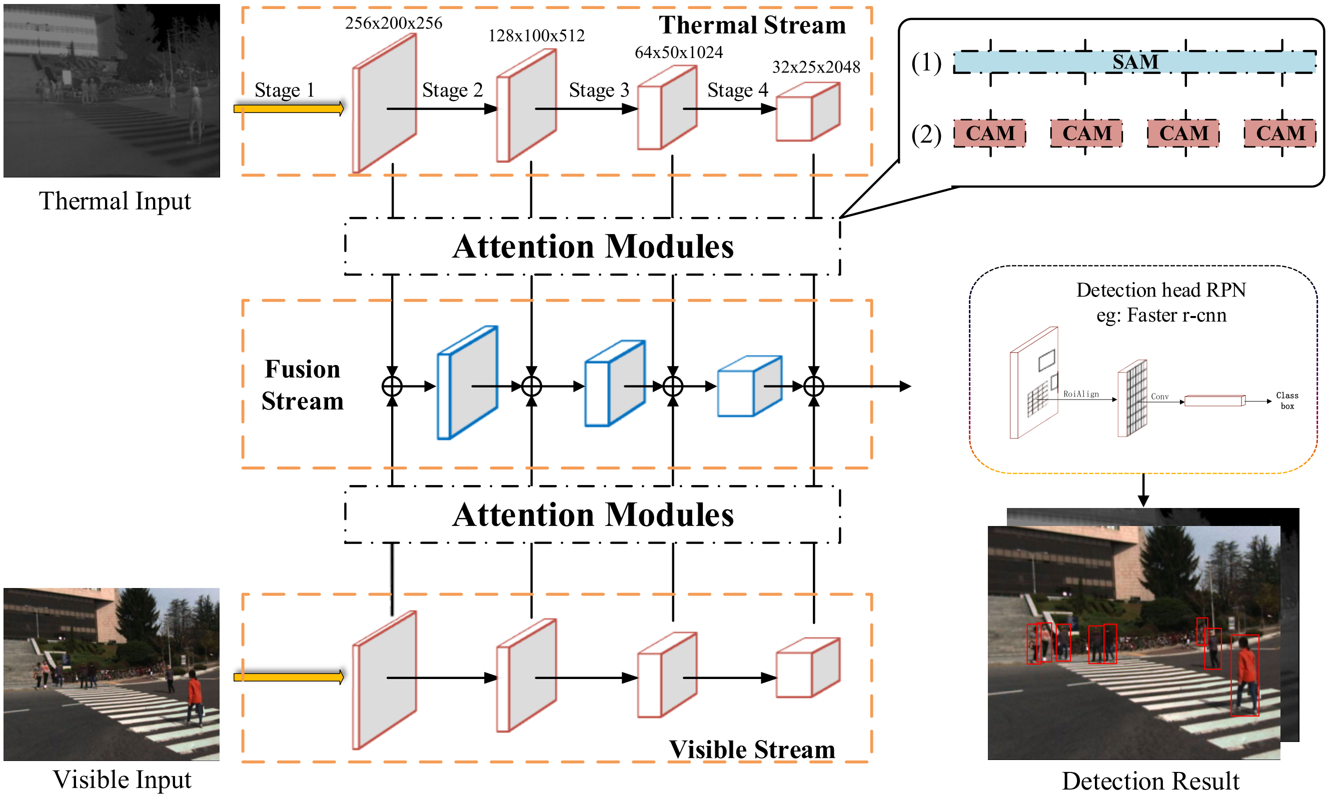


FIGURE 1. Overview of our framework. The ⊕ denotes pixel-wise addition. SAM and CAM are used to compute spatial-wise and channel-wise weights respectively. To reduce parameters, two feature extraction streams share the same parameters for CAMs and SAMs.

same backbone feature extraction network and each takes a spectrum of images as inputs. A fusion stream is introduced to fuse multiple layers of cross-spectral features. Furthermore, a channel-wise attention module (CAM) and a spatial-wise attention module (SAM), which generate attention values as weights to be applied at the channel and the pixel level respectively, are developed and incorporated into the detector. We apply one SAM for each stream of feature extraction, whose layers each are associated with a CAM. The CAM is trained in a self-supervised manner while the SAM is trained with external supervision, based on saliency detection.

We start the exploration of the optimal multi-layer fusion structure with single-layer fusions, and examine the effectiveness of multi-layer fusion by gradually increasing the number of fused layers. On the basis of the multi-layer fusion structure, we study the efficacy of two attention based weighting mechanisms by applying them separately and then sequentially.

All proposals are evaluated and validated on the KAIST benchmark [21]. The results show that the multi-layer fused network, when integrated with the SAM, is a state-of-the-art detector on all-day pedestrian detection and surpasses performance at nighttime.

Contributions of our work are as follows:

1. We have developed a novel multi-layer fusion network for multispectral pedestrian detection and validated the effectiveness of multi-layer fusion for learning cross-spectral features.

2. Both spatial and channel attention are innovatively introduced as weighting mechanisms for multispectral fusion and their significance in improving detection performance is validated. In addition, to obtain finer spatial attention, we model the process of learning spatial attention as saliency detection with external supervision.

3. Spatial-wise weighting is proved to be more effective, with which the multi-layer fused network is demonstrated to be a top-performance end-to-end pedestrian detector as compared to available competitors.

II. RELATED WORK

The release of large-scale multispectral benchmark datasets (e.g., KAIST [21], UTokyo [45], and CVC-14 [16]) and pioneer proposals of CNN-based models have prompted the research community to invest more into the field of multispectral pedestrian detection [8], [9], [24]. Faster R-CNN [42], which introduced a region proposal network (RPN) that shares full-image convolutional features with the detection network to generate region proposals, has become the de facto basis of most exsiting multispectral pedestrian detectors [26], [29], [32], [42]. Liu *et al.* [32] introduced a model based on Faster R-CNN [42] and compared four network architectures that fuse visible light and infrared thermal features from different stages. Li *et al.* [29] considered using a gate function which is designed based on illumination conditions to weight visible and infrared features before fusion. Konig *et al.* [26] utilized boosted decision trees to reselect the region proposals which are generated by RPN networks. Chen *et al.* [5] designed a multi-layer fused region proposal network, in which a summation fusion method was applied for integration of two convolutional layers. Guan *et al.* [17] presented a novel illumination-aware weighting mechanism which is incorporated into a two-stream deep convolutional neural network to learn multispectral human-related features under different illumination conditions. Park *et al.* [41] considered all detection probabilities from each modality in a unified three-branch CNN framework and selectively used them through a channel weighting fusion layer to maximize the detection performance. An accumulated probability fusion layer was also introduced to combine probabilities from different modalities at the proposal-level. Taking the position shift problem of multispectral data into consideration, Zhang *et al.* [54] proposed a region feature alignment module to capture position shifts and a confidence-aware fusion method to merge both modalities.

Although the fused network has yet to be explored for multispectral pedestrian detection, they are extensively studied in other fundamental vision tasks, including semantic segmentation [7], [37], action recognition [13], [23], [43], 3D object classification/detection [18], [44] and so on. Simonyan and Zisserman [43] proposed to recognize actions with a dual-stream CNN to process optical flow and visible images respectively. Karpathy *et al.* [23] explored a few methods to fuse frames at different speeds for classifying videos. For better semantic segmentation, Cheng *et al.* [7] designed a gated fusion layer for weighted fusion based on the varying contribution of color and depth information in detecting various categories of objects in different scenes. Liu *et al.* [34] proposed a gated multi-layer convolutional feature extraction method which could adaptively generate discriminative features for candidate pedestrian regions. Additionally, many researchers have also explored the use of multi-layer information for better detection and segmentation. HyperNet [25] and ParseNet [35] concatenate features from multiple layers and then make the final predictions. FPN [31] explores the top-down architecture to produce feature maps with high-level semantics at all scales. DenseNet [20] connects features at each layer to those at every other layer in a feed-forward fashion, so that the output feature comprises information at mutiple levels.

Attention is introduced over recent years to improve CNNs' performance. Spatial-wise attention, which investigates the spatial correlations is applied in vision-based studies including digits recognition [22], object recognition [1], image caption [51], object detection [2], and pose estimation [39], etc. As for channel-wise attention, Hu *et al.* [19] proposed the squeeze-and-excitation network to model the interdependence between feature channels to generate channel attention, Zhang *et al.* [55] explored different types of attention mechanisms at the channel level. Bello *et al.* [3] proposed to augment convolutional operators with the self-attention mechanism by concatenating convolutional feature maps with a set of feature maps produced via self-attention. In other studies [48], [50], attention along both dimensions are implemented at the same time, though with the spatial attention trained with self-supervision. In this paper, we propose to learn channel-wise attention with self-supervision and remodel the process of learning spatial-wise attention as saliency detection with external supervision.

III. OUR APPROACH

A. OVERVIEW OF OUR FRAMEWORK

We propose a multi-layer fused triple-stream CNN for multispectral pedestrian detection. Based upon Faster R-CNN that so far performs the best in pedestrian detection, we create two streams of feature extraction with ResNet50, one for each spectrum of images, and build a weighted fusion stream to fuse cross-spectral features at multiple layers.

Similar to [31], we say that the layers producing feature maps of the same size are at the same network *stage*. As is shown in Fig. 1, channel-wise attention modules are implemented at each stage to compute weights for each feature channel at that stage, and spatial-wise attention modules are implemented to generate a saliency map as spatial weights shared by all stages of a feature extraction stream. To facilitate implementation, we make all attention modules detachable so that ablation studies could be carried out.

Lastly, the fusion stream fuses multispectral features and feeds them into the RPN network to generate region proposals, and has the same subsequent steps as those in the original Faster R-CNN.

Since all attention modules are integrated into the network and are trained end-to-end, the loss function of the proposed pedestrian detector is designed to include the loss of training spatial attention, and is defined as below:

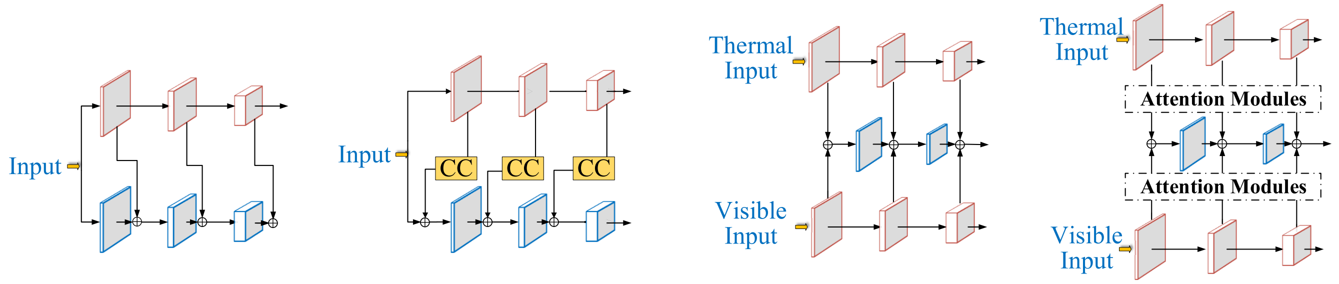
$$L_0 = \eta \times L_{spatial} + L_{rpn\_cls} + L_{rpn\_reg} + L_{cls} + L_{reg} \quad (1)$$

where: $L_{spatial}$ is the cross-entropy loss for the computation of spatial weights; η indicates whether spatial-wise weighting is applied. $\eta = 1$ means true, $\eta = 0$ false; L_{rpn_cls} and L_{cls} denotes the cross-entropy loss for classification in the RPN and the main classification network; L_{rpn_reg} and L_{reg} denotes the L1 loss for bounding box regression.

B. MULTI-LAYER FUSION

The fact that multiple layers of features extracted with deep CNNs embody richer semantics [30], [31] motivates us to propose a multi-layer fusion structure with the aim to merge cross-layer expressive features. Since both feature extraction streams share the same backbone network, they can be initialized with the same parameters. Both addition and concatenation are widely used as the basic operations for fusion. An in-depth analysis on these two operations can be found in [6]. Here, for simplicity, we adopt the pixel-wise addition instead of concatenation to fuse multi-modal features.

As authors pointed out in research [36], *adjacent high-level composition*(AHLC) (Fig. 2 (b)) tends to enrich the extracted features, while *same layer composition*(SLC) (Fig. 2 (a)), is non-beneficial since all streams share the same source



(a) Same Level Composition

(b) Adjacent Higher-Level Composition

(c) Multi-layer Fusion

(d) Multi-layer Weighted Fusion

FIGURE 2. Illustrations of 4 variants of multi-layer fusion structures, with the one in (d) being our proposal. '*CC*' (Composite Connection [36]) in (b) represents preparatory operations before feature fusion; '*Attention Modules*' in (d) represents either SAMs or CAMs, of which more details are presented in Section III-C.

of inputs and the same network settings, and thus similar features are produced by each stream, of which little improvement is obtainable via feature composition. However, we argue that in the case of multispectral fusion, composing features at the same layer is effective in enhancing feature expressiveness since each modality of inputs represents an entirely different set of physical information perceived.

Moreover, it's important to notice that neither *same layer composition* nor *adjacent high-level composition* should occurr within any feature extraction backbone network when designing the multi-layer fusion scheme. Removing the aforementioned compositions prevents one stream from continuously inserting its extracted features into another stream, thus intensifying the features at each level of fusion to a point where they become dominant in the output of fused features. Therefore, we build a standalone fusion stream by modifying the backbone network used in feature extraction so that features of each modality play their roles according to weights assigned specifically to them (Fig. 2 (c)).

C. ATTENTION BASED WEIGHTING MECHANISMS

We incorporate two attention based weighting mechanisms into the fusion process (Fig. 2 (d)) so that the network learns to emphasize meaningful features while suppressing interference at both the channel and the spatial level.

At the channel level, each feature channel is assigned an overall weight. The channel-wise weighting mechanism is described as in the following formula:

$$F^c = \lambda_t^c \times T^c + \lambda_v^c \times V^c \tag{2}$$

where: F^c denotes the fused feature from the corresponding channel c in the feature map from each spectrum; $\lambda \in (0, 1)$ is the overall weight for all the pixels in a channel; T, V represent the thermal and the visible light feature respectively.

The spatial-wise weighting mechanism that applies at the pixel level is expressed mathematically with the following formula:

$$F_{w,h} = \theta_{w,h}^t \times T_{w,h} + \theta_{w,h}^v \times V_{w,h} \tag{3}$$

where: F denotes the fused feature; (w, h) denotes the pixel position in the feature map; $\theta \in (0, 1)$ is the weight of the pixels at the same pixel position in all channels.

1) CHANNEL-WISE ATTENTION BASED WEIGHTING

As stated in previous studies, each feature channel could be considered as a feature detector [52], and the focus of *channel attention* is 'what' is necessary to learn in the input image. Therefore, we use the channel attention map as the channel-wise weights for feature fusion, and assign attention values as weights to feature map channels. As a result, useful or meaningful channels are preserved while interfering channels are suppressed before fusion.

We revise the SE [19] module to produce channel attention maps. To compute channel attention more efficiently, *global average pooling* (GAP) and *global max pooling* (GMP) are often adopted to squeeze the spatial dimension of the feature maps [19], [50], [56]. However, we argue that neither GAP nor GMP are ideal in our case where multiple small-scale targets are presented in background information-dominant scenes. By comparison, the feature descriptor will be overwhelmed by background features and become less informative if GAP is applied while a lack of an overall representation of all significant objects would be inevitable if GMP is applied. Additional considerations for the favoring of the most significant target in a scene and the possibility for generating defective maximum values with GMP have been considered.

Instead, we introduce a novel approach of global pooling, called *global attention average pooling*, to attend to all the significant regions that reflect the characteristics of all targets of interest. With a threshold of significancy determined, we define the boundaries between significant and non-significant regions based on the value of each pixel. The feature descriptor D in a particular channel c can be computed with global attention average pooling using the following formula:

$$D^c = \frac{1}{m^c} \sum_{w,h} \left(u_{w,h}^c \times p_{w,h}^c \right) \tag{4}$$

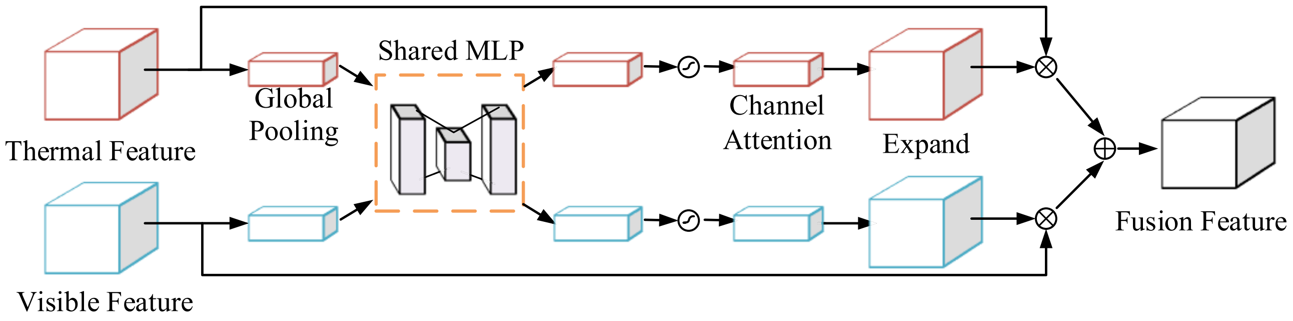


FIGURE 3. Diagram of channel-wise weighted fusion.

where: (w, h) denotes the pixel position in the feature map; p is the pixel value; $u \in \{0, 1\}$ is the significant label of a pixel, which equals to 1 if its p value is greater than the threshold, 0 otherwise; m is the number of all the significant pixels in a channel.

It should be noted that global attention average pooling encompasses both GAP and GMP. When all pixel values are greater than the threshold, they will be included to compute the global descriptor. When all pixel values are below the threshold, only the pixel with the maximum value will be retained in the channel. See Appendix A-A for more details about global attention average pooling.

The process of implementing the channel-wise weighted fusion is presented in Fig. 3. After the spatial context descriptor is generated with global average attention pooling, a shared multi-layer perceptron with a hidden layer where the activation size is set to $R^{C/r \times 1 \times 1}$ is used to generate channel attention. A reduction at the ratio r is adopted to reduce the computational cost. Once completed, we normalize the channel attention using a sigmoid function, and re-weight the original features before fusion.

2) SPATIAL-WISE ATTENTION BASED WEIGHTING

We implement the learning process of spatial-wise attention based on a salient object detection (SOD) network that generates saliency maps, which then serve as spatial attention maps in our work. Unlike channel-wise attention, the focus of spatial attention is 'where' in the input could provide more informative features, and hence deserves more attention.

SOD, which is in an effort to highlight the conspicuous objects in an image, has developed rapidly in recent years [49]. PiCANet [33] is a pixel-level contextual attention network that can selectively attend to local or global contexts and produce informative contextual features for each pixel. We incorporate PiCANets into CNNs hierarchically for joint training to accomplish saliency detection.

To generate spatial-wise weights for re-weighting features at different stages, we incorporate the PiCANets into the ResNet50 backbone network. For all the feature activation outputs of each stage's last residual block, a 1*1 convolutional layer is attached to perform a dimension reduction at the ratio of 1/4 to lower the computational complexity. We have also tried other reduction ratios (*i.e.* none or 1/2) and observed little improvement in detection performance with higher computation overhead. Besides, since PiCANet requires fixed-size inputs, we utilize a bilinear interpolation to scale the spatial resolution to meet the size requirements. Due to space constraints, we do not show operations mentioned above in Fig. 4.

After these preparatory steps, a global PiCANet is used to merge the features originally from the last two stages in the backbone, as shown in Fig. 4. Then another two local PiCANets are used hierarchically to fuse the features at the current stage and those from the previous stages. After that, a 1*1 Conv layer with sigmoid activation is implemented for generating the saliency map. Once again, a bilinear interpolation followed by a 1*1 Conv and a Batch Normalization layer is used to restore the saliency map to match the depth and the size of the feature maps at different stages. The resized and expanded attention maps are employed as spatial-wise weights for different stages.

For more accurate saliency predictions, we adopt box-level saliency annotations, as generated in [4], as external supervision and compute the cross-entropy loss which will be added to the total loss of the detector. It should be pointed out that although PiCANets are adopted in our work, other SOD networks are also worth trying. See Appendix A-B for more implementation details.

IV. EXPERIMENTS

A. SETUPS

1) DATASET AND METRICS

We evaluate all the proposed works on the KAIST multispectral pedestrian benchmark that contains a total of 50,172 well-aligned visible-infrared image pairs captured in all-day traffic scenes with 13,853 annotations of pedestrians. As suggested in [29], we remove all instances labelled as '*person?*' or '*people*', indicating that the class or the number of the instance is ambiguous even to human annotators. There are 2,252 image pairs in the testing dataset of KAIST, among which 797 pairs were captured during nighttime. The improved annotations provided in [32] is adopted for more accurate evaluations since there are some problematic annotations in the original testing dataset. As adopted in [12], we calculated the log miss

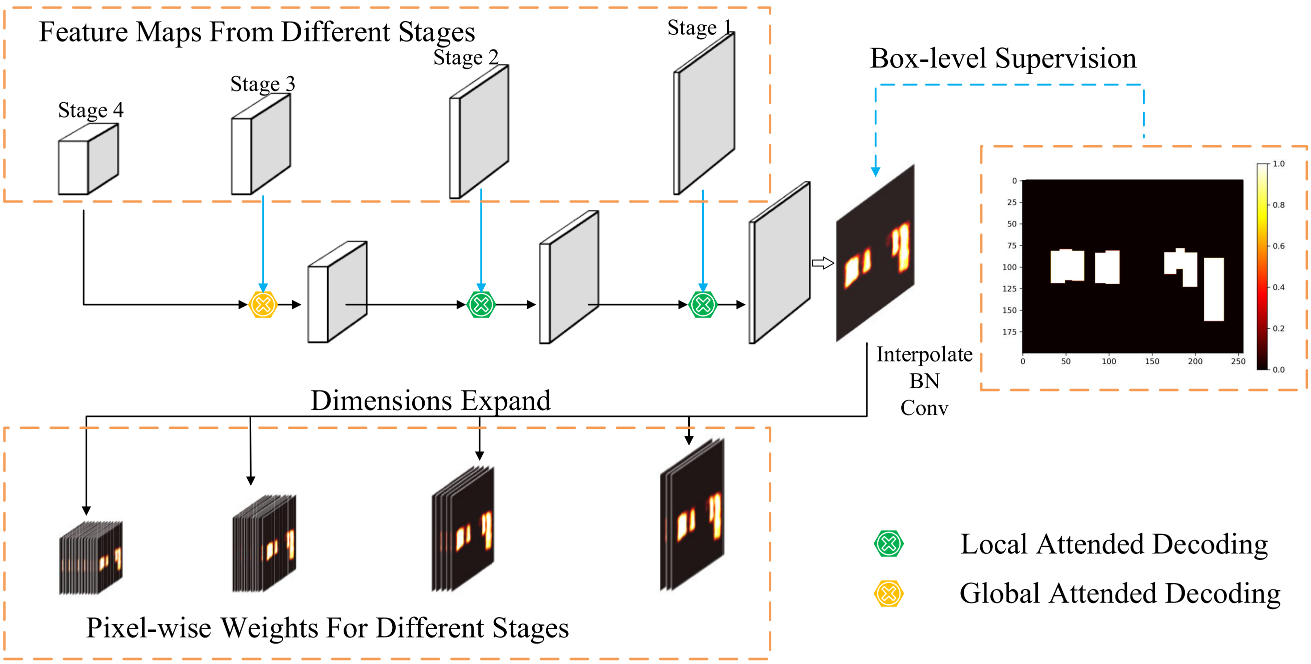


FIGURE 4. Diagram of spatial-wise attention module (SAM). For structures of *Global Pixel-wise Attending* and *Local Pixel-wise Attending*, refer to the research [33] for details.

rate averaged over the false positives per image (FPPI) range of [10^{-2}, 10^{0}] (MR) to measure the detection performance under reasonable configuration.

2) IMPLEMENTATION DETAILS

The pretrained ResNet50 on ImageNet [11] is used to initialize the backbones. The Kaiming initialization is applied to some model parameters that are not initialized with the pretrained model. The first two convolutional layers of the ResNet50 network are fixed and the rest are finetuned using Stochastic Gradient Descent (SGD) with momentum of 0.9, batch size of 3, learning rate of 0.075 which decreases at a rate of 0.1 at every 6000 iterations. The training process is terminated after 8000 iterations and early stopping is adopted. We implement our code based on the implementation of the maskrcnn-benchmark [38]. Other implementation details are as those in the original work of Faster R-CNN [42].

B. RESULTS

1) MULTI-LAYER FUSION

We examine the effectiveness of multi-layer fusion by increasing the number of fusion layers gradually. A series of experiments are carried out, including the single-layer fusion group labeled as: 1-fusion, 2-fusion, 3-fusion, the double-layer fusion group labeled as: 1-fusion, 1-fusion*, 2-3-fusion, 2-3-fusion*, and the triple-layer fusion group labeled as: 1-2-3-fusion, 1-2-3-fusion*. A few examples of the network structures designed for the experiment are illustrated in Fig. 5. Here, experiments are labeled in such a way that the number indicates the index of the fusion stage, while multiple numbers indicate that there are multiple layers of fusion, and the asterisk means that the proposed channel-wise weighting mechanism is applied.

We also validate the assumption that having a standalone fusion branch is of great importance by comparing the SLC and AHLC structures, as illustrated in Fig. 2 (a) and Fig. 2 (b) respectively, with our proposed multi-layer fusion structure, as illustrated in Fig. 2 (c). Here, both SLC and AHLC are implemented at all 3 layers.

Observations from the results given in Table 1 are five-fold: First, it's proved that the structure presented in Fig. 2 (a) and Fig. 2 (b) are unsuitable for multi-layer fusion. A standalone fusion branch is essential to prevent features from either stream from becoming dominant as the depth of fusion increases. Second, when single-layer fusion is applied, 2-fusion achieves the lowest MR since the intermediate features contain semantics while preserving fine visual details. Third, as long as the fusion starts at the same layer, a deeper fusion scheme seems to always contribute to improving detection performance. For example, 1-2-fusion and 1-2-3-fusion outperform 1-fusion, and 2-3-fusion outperforms 2-fusion. Fourth, the layer at which the multi-layer fusion starts matters. It's not surprising to see that the triple-layer scheme: 1-2-3-fusion results in a higher MR as compared to a double-layer design: 2-3-fusion, since features extracted at the first layer, which contains a lot of task irrelevant low-level features, would weigh in the fused features. Fifth, consistent improvements in detection performance are seen when the channel-wise weighting mechanism is applied. It's proved that attention along the channel dimension contributes to suppressing interfering features from both spectrums.

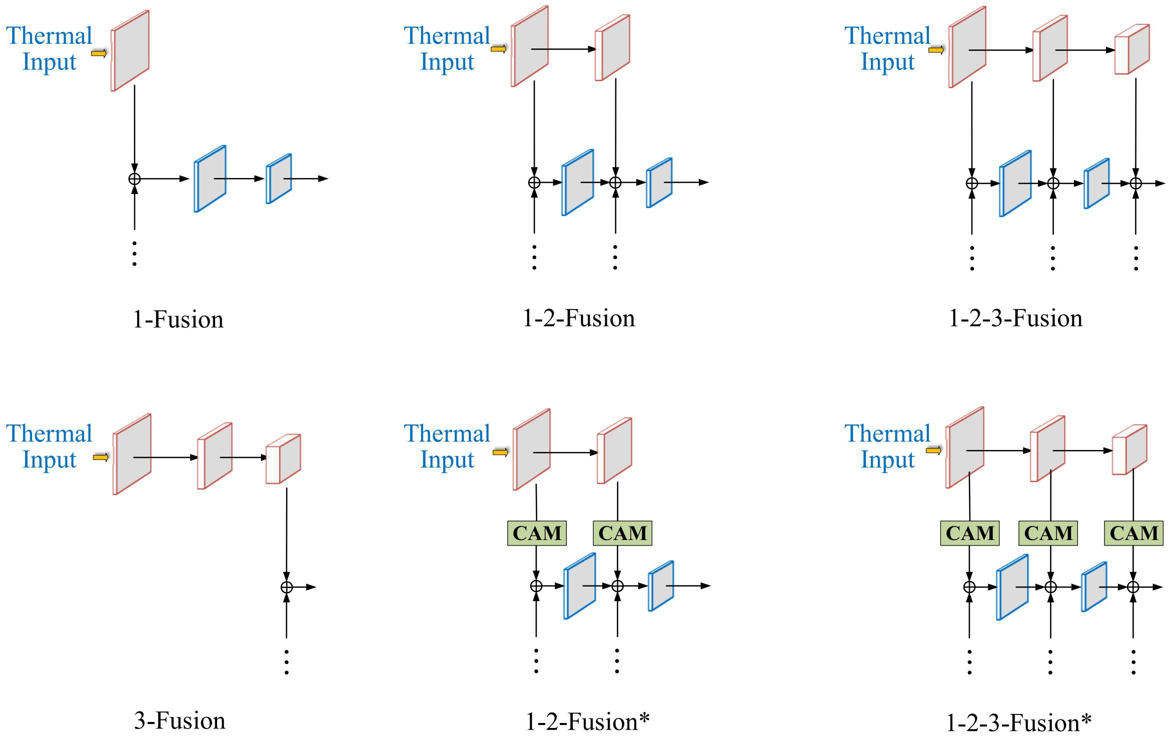


FIGURE 5. Selective structures of multi-layer fusion.

TABLE 1. Comparison of different fusion structures. Visible or thermal means only visible or thermal images are used as inputs. SLC and AHLC stands for 'same layer composition' and 'adjacent high-layer composition' respectively, (Visible base) means the composition occurs in the visible stream, while (Thermal base) means it occurs in the thermal stream. The numbers represent the layers of fusion, multiple numbers means multi-layer fusion, and *' indicates that channel-wise weighting is applied.

Fusion Type	All-day	Day	Night
Visible	31.42	23.49	48.68
Thermal	25.95	32.90	**10.65**
SLC (Visible base)	18.25	20.65	13.26
SLC (Thermal base)	19.95	19.56	20.79
AHLC (Visible base)	19.03	22.28	13.12
AHLC (Thermal base)	19.82	19.03	22.01
1-Fusion	22.30	20.77	24.93
2-Fusion	17.25	18.89	13.74
3-Fusion	18.01	19.23	15.22
1-2-Fusion	17.46	17.59	16.67
1-2-Fusion*	16.31	17.88	13.65
2-3-Fusion	15.20	16.47	12.66
2-3-Fusion*	14.66	16.50	11.57
1-2-3-Fusion	16.30	16.85	15.74
1-2-3-Fusion*	**14.45**	**15.78**	11.57

2) ATTENTION BASED WEIGHTING MECHANISMS

Table 2 shows our experimental results as well as the inference speed of multi-layer fusion with the same network structure but incorporated with different weighting mechanisms. For channel-wise weighted fusion, we implement the CAM based on the original SE [19] module and replace GAP with global attention average pooling with an empirical significancy threshold of 0.1. At the same time, the reduction ratio is fixed to 16.

The results prove that the proposed global attention average pooling contributes to extracting more accurate spatial information characteristics than the GAP or GMP. We observe that the difference in detection performance when using GAP or GMP alone is trivial. When GAP and GMP are applied in parallel, the improvement in precision is insignificant, while the computational complexity is doubled. With the global attention average pooling applied alone, the best performance is obtained at no extra computational cost.

TABLE 2. Effect of integrating different attention modules. '*STD*' means standard multi-layer fusion network. '*SE*' means channel-wise weighted fusion, and the subscript denotes the type of global pooling method applied, it's our proposed *global attention average pooling* if not specified. '*PiCA*' means spatial-wise weighted fusion, and the subscript denotes the type of annotations, and it's box-level if not specified.

Structure	All-day	Day	Night	Speed(s)
STD	16.34	16.76	15.74	0.11
STD+SE$_{avg}$	15.99	15.78	16.88	0.11
STD+SE$_{max}$	15.88	16.96	12.61	0.11
STD+SE$_{avg+max}$	14.66	16.50	11.57	0.12
STD+SE	**14.45**	**15.78**	**11.57**	0.11
STD+PiCA$_{pixel}$	**11.22**	**11.60**	8.95	0.22
STD+PiCA	11.43	11.86	**8.82**	0.22
STD+PiCA+SE	14.62	15.63	12.49	0.24
STD+CBAM [50]	18.75	21.24	14.07	0.15
STD+AA [3]	15.20	16.47	12.66	0.26

The best detection performance is achieved with the spatial-wise weighting mechanism applied alone. A decline instead of an increase in performance is seen when channel-wise attention modules are applied after spatial-wise attention modules. The underlying causes are discussed later in the next chapter. It's worth mentioning that although incorporating SAMs into the multi-layer fusion network leads to a reduction by half in detection speed, the performance gain is significant.

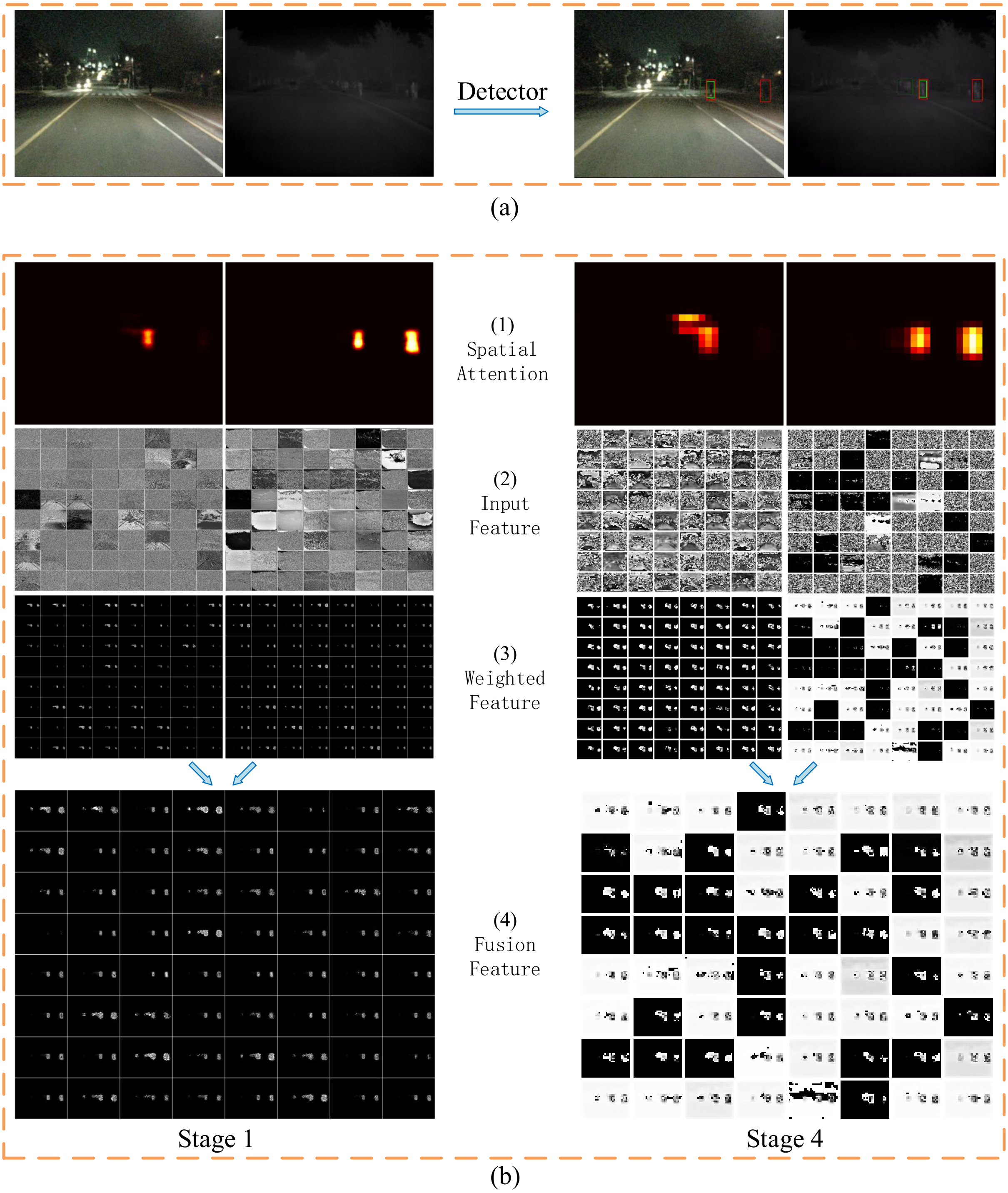


FIGURE 6. Process of the spatial-wise weighted fusion. (a) presents the original input images (left) and the detection results (right), and (b) presents the process of fusion at the initial stage in the left column, and that at the fourth stage in the right column. We only visualize the first 64 channels of the feature maps due to space constraints.

We also evaluated our proposals against two widely acknowledged attention mechanisms CBAM [50] and AACONV [3]. The results show that: First, our proposed channel-wise attention surpasses the performance of both CBAM and AAConv when global attention average pooling is adopted. Second, the proposed spatial-wise attention outperforms CBAM and AAConv by a large margin. It's also worth noticing that as compared to the standard network with no attention mechanism applied, the one with CBAM shows some strength in the nighttime, but performs much worse in the daytime.

TABLE 3. Comparisons with other multispectral detectors on the KAIST dataset.

Methods	All-day	Day	Night	Speed(s)
ACF+T+THOG [21]	47.24	42.44	56.17	-
Halfway Fusion [32]	25.75	24.88	26.75	0.43
Fusion RPN [26]	20.67	19.55	22.12	-
Fusion RPN+BF [26]	15.91	16.49	15.15	0.80
IAF-RCNN [29]	15.73	14.55	18.26	0.21
IATDNN+IAMSS [17]	14.95	14.67	15.72	0.25
CIAN [53]	14.12	14.77	11.13	0.07
MSDS-RCNN [28]	11.63	**10.06**	13.73	0.23
CS-RCNN(ours)	**11.43**	11.86	**8.82**	0.22

In Fig. 6, we visualize the first and last stage of the spatial-wise weighted fusion process to intuitively demonstrate the effectiveness of the spatial-wise weighting mechanism.

Fig. 6 (a) presents the original input images (left side) and the detection results (right side), Fig. 6 (b) presents the process of fusion at the first stage (left side) and the fourth stage (right side). In Fig. 6 (b), we have selected the first 64 channels of each feature map, due to length limitations. As shown in row (1) in Fig. 6 (b), the visible light attention map only highlights 1 region that a pedestrian presents, while the thermal attention map highlights 2 regions with the right one being undetectable via the visible light spectrum. It's also clearly shown that the fused feature map at row (4) highlights both the left and right pedestrian regions, with non-salient regions suppressed.

3) PERFORMANCE COMPARISON AGAINST THE STATE-OF-THE-ART DETECTORS

Our proposed spatial-wise weighted fusion method is compared with other available competitors, as illustrated in Table 3. For the sake of fairness, studies using pruned and unpublished dataset are not included. With a MR of 11.43 for pedestrian detection at all-day, our proposal outperforms most methods and is evenly matched with MSDS-RCNN [28], by showing significant strength in the nighttime.

Here, CIAN [53] runs much faster as compared to all other detectors including ours because the input size it adopts is much smaller. It's worth mentioning that the way that MSDS-RCNN uses box-level segmentation as supervision during the training phase is quite different from our method since we remodel the process as saliency detection and use the predicted saliency map as the spatial weights.

Additionally, we plot the MR against FPPI (using log-log plots) of our work and selected competitors with varying thresholds of detection confidence, as shown in Fig. 7.

V. DISCUSSION

In this section, we share more findings that are supportive to our work.

A. ARE HIGH-QUALITY SALIENCY ANNOTATIONS NECESSARY?

High-quality pixel-level annotations for SOD networks are quite labor intensive and sometimes even impossible, especially for multispectral datasets. As discovered in our study, high-quality pixel-level annotations aren't necessary because it's the regional characteristics of pedestrians we wish to preserve. Box-level annotations suffice.

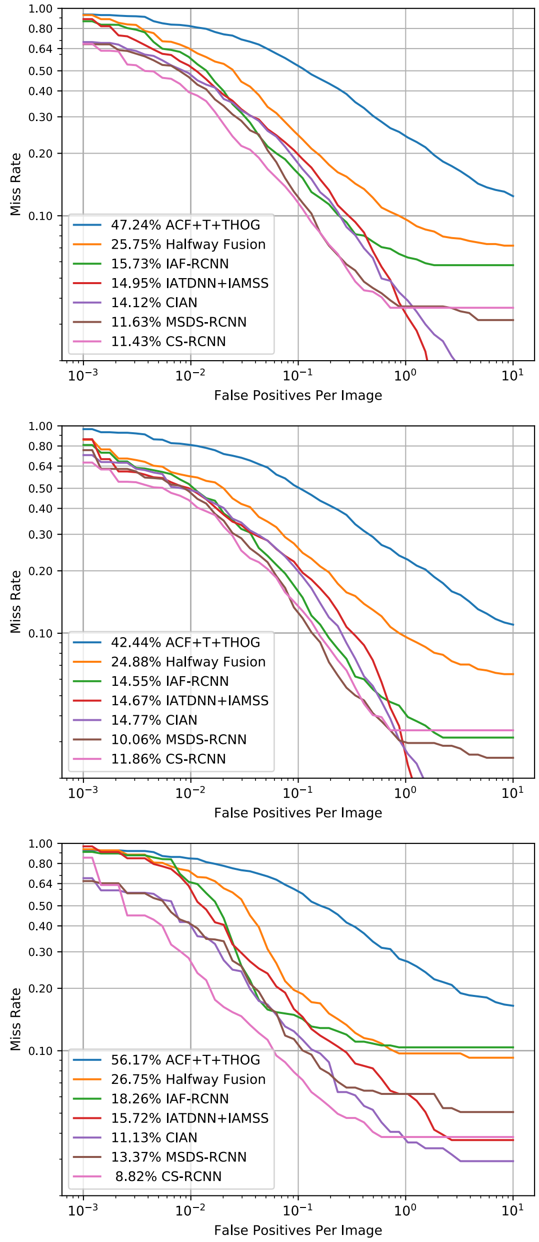


FIGURE 7. Comparison of detection performances reported on the improved KAIST multispectral pedestrian test dataset, in all-day (top), daytime (middle), and nighttime scenes (bottom).

As a comparison, we conducted experiments with pixel-level saliency annotations as the supervision, which were generated using deep SOD networks, in our case the R^3Net. We used pixel-level annotations, provided by [14] for partial KAIST training data, to train R^3Net independently, and then used the trained R^3Net to predict saliency maps for the rest of the training data to produce supervision for the PiCANets embedded in our model.

As suggested by the results in Table 2, there is little improvement in performance with pixel-level annotations at a much higher cost of data preparation.

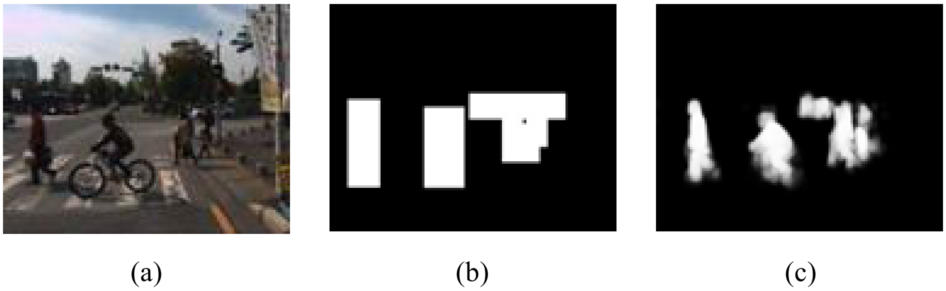


FIGURE 8. Illustrations of box-level annotations and pixel-level annotations. (a) is the input image, (b) shows box-level annotations. (c) shows pixel-level annotations.

TABLE 4. Experimental results from decreasing number of RoIs.

Number of RoIs	All-day	Day	Night
1000	11.43	11.86	8.82
500	11.43	11.86	8.82
300	11.55	12.03	9.01
200	11.76	12.14	9.06
100	12.61	14.03	10.01
50	12.83	14.38	9.90
20	13.82	15.39	9.90
10	16.34	18.23	11.80
5	20.95	23.21	15.33

B. CAN WEIGHTING MECHANISM BASED ON SPATIAL-WISE ATTENTION REALLY HELP TO SUPPRESS INTERFERING FEATURES?

To verify the effectiveness of spatial-wise weighting mechanism in suppressing interfering features, we carried out a set of experiments with pertinence and the results are presented in Table 4. With the number of RoIs input to the subsequent main classification network drops dramatically from 1000 to 50, the performance of detection stays almost unaffected, proving that the proposed spatial-wise weighting mechanism effectively eliminates interfering features.

C. WHY IS SPATIAL-WISE WEIGHTING SUPERIOR TO CHANNEL-WISE WEIGHTING, AND ALSO THE COMBINATION OF BOTH?

As stated earlier in this paper, each feature channel could be considered as a feature detector. In the case of classification, where targets are conspicuous, assigning a single attention value to an entire channel is effective since the channel is dominated with either useful or meaningless features.

However, in the case of pedestrian detection, where multiple instances of targets presented in a background information dominant scene, it becomes a tricky task to decide whether to emphasize or suppress a channel since it could contain both useful and interfering features. We argue that the inter-channel relationship would become so complicated that a self-supervised attention generation module would easily fail to learn fine attention values for each channel. To address this issue to some extent, the global attention average pooling method is proposed, so that more accurate channel-wise attention is obtainable through attending to all salient regions.

As for spatial-wise attention, it doesn't suffer from the same problem since attention values are generated at the pixel level. Moreover, we introduced external supervision to refine the learning scheme in our work. Also, spatial-wise attention is learned on the basis of merging semantics from all layers, while channel-wise attention is learned by exploiting features at each single layer.

As a result, the improvement made by channel-wise attention is shadowed by gains achieved through spatial-wise attention, and a set back in performance would be resulted in when applying spatial-wise and channel-wise weighting mechanisms sequentially.

VI. CONCLUSION

For multispectral pedestrian detection, we propose a triple-stream multi-layer weighted fusion network by first exploring the optimal structure of layers to fuse for preserving cross-layer informative features. Experimental results suggest that a deeper fusion structure is beneficial for improving the detection performance as long as the layer to start is chosen carefully. Channel attention and spatial attention based weighting mechanisms are developed and incorporated into the fused network for re-weighting multispectral features at the channel and the pixel level before fusion. Both weighting mechanisms contribute significantly to enhancing detection performance, while the spatial-wise weighting proves to be the most effective when applied alone. Experimental results on KAIST show that our multi-layer fusion network incorporated with the spatial-wise weighting mechanism achieves the state-of-the-art performance on all-day pedestrian detection and outperforms at nighttime.

In future work, we will improve the method of global attention average pooling so that a dynamic significancy threshold is learned and set for each layer. Also, more work will be done to optimize the spatial-wise attention module to reduce the computational overhead.

APPENDIX A ADDITIONAL IMPLEMENTATION DETAILS

A. BACKPROPAGATION FOR GLOBAL ATTENTION AVERAGE POOLING

Here, we show some more details of backpropagation for the proposed global attention average pooling. Based on the

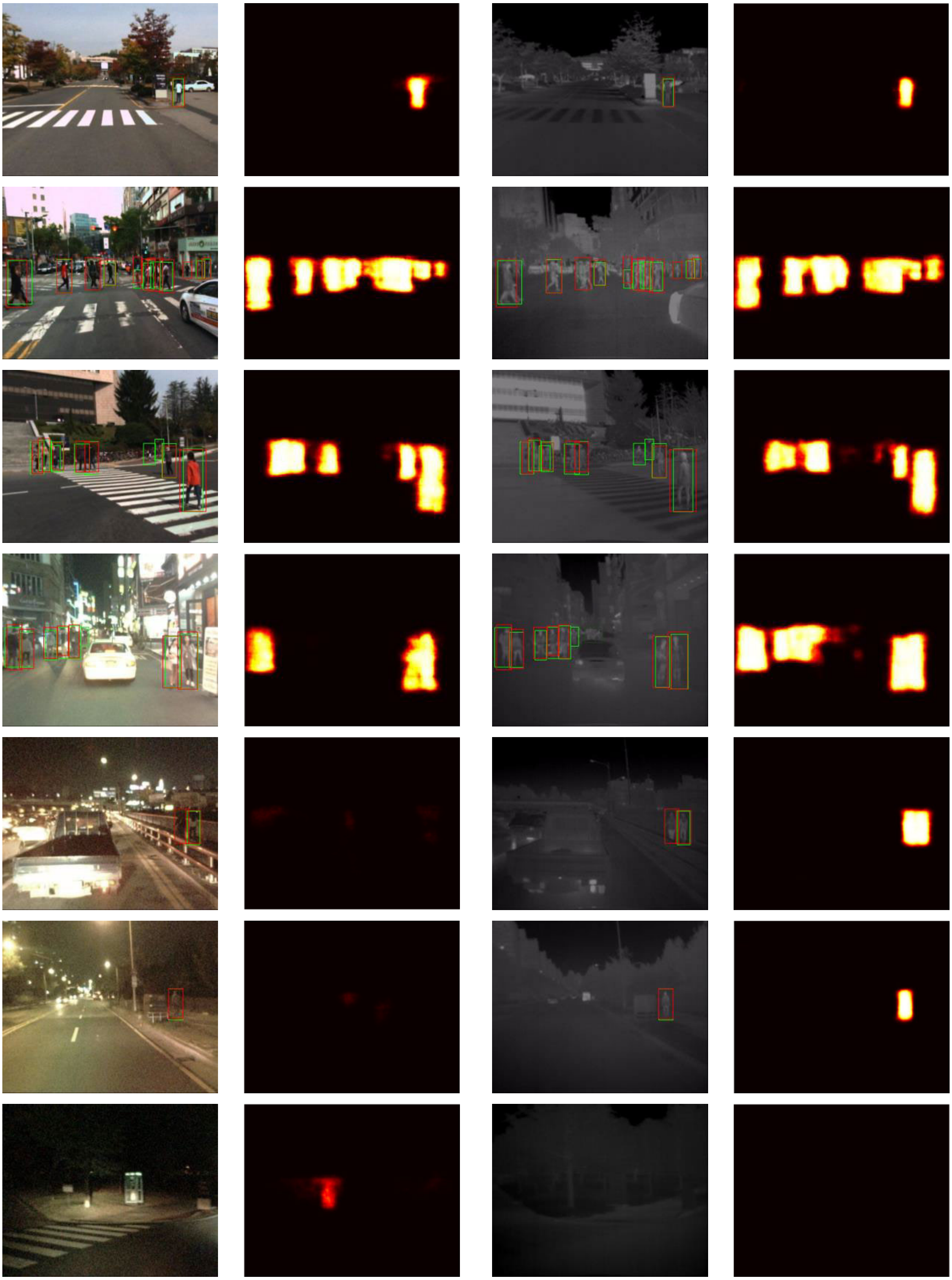


FIGURE 9. Examples of the learned attention maps of the spatial-wise weighting mechanism. The 1 to 3 and 4 to 7 rows show daytime and nighttime scenes respectively. The first and the third column show the visible and thermal input respectively, the second and the fourth column show their saliency predictions respectively generated by the PiCANets embedded in the backbones. Note that the red and the green bounding boxes (BBs) represent the BB prediction and BB ground truth, and the dark green boxes with thinner line represent the ignored BB ground truth.

gradient $D^{R\times1\times1}$ input to global attention average pooling, we use the following formula to compute the gradients in a channel (labelled as c) of the previous layer:

$$d^c_{w,h} = \mu^c_{w,h} \times \frac{D^c}{m^c} \tag{5}$$

where: (w, h) is the pixel position; $u \in \{0, 1\}$ is the significant label of a pixel, 1 means above the significancy threshold, 0 otherwise; m is the number of significant pixels in a channel.

As mentioned in the submitted manuscript, when the value of all pixels is lower than the significancy threshold, we retain the pixel with the maximum value in that channel. In this case, $m = 1$.

B. DETAILS IN LEARNING SPATIAL-WISE ATTENTION

In the training phase, we use the box-level annotations generated based on the ground truth bounding boxes (manually annotated) as the supervision of PiCANets, as shown in the Fig. 8. According to [33], we set the weights of the losses of all 4 stages in a feature extraction stream to 0.5, 0.5, 0.8, and 1, respectively. We then use the sum of the weighted loss of each stage as the saliency loss of the very stream.

We use box-level annotations to supervise the visible and the thermal stream for saliency predictions. The tolerance on the loss of saliency detection in each stream is different since one spectrum of inputs is superior to another depending on varying ambient conditions. With that taken into account, for re-weighting the saliency losses which will be added to the total loss, we introduce another two hyperparameters α, β and empirically set (α, β) to (1,0.5) in the daytime and (0.8, 1) in the nighttime to obtain optimal performance.

$$L_{spatial} = \alpha \times L_{spatial\_V} + \beta \times L_{spatial\_T} \tag{6}$$

where: $L_{spatial\_V}$ and $L_{spatial\_T}$ are the cross-entropy loss computed in the visible and thermal streams respectively.

APPENDIX B
OTHER EXPERIMENTAL RESULTS
A. MORE EXAMPLES OF THE LEARNED ATTENTION MAPS

We present more examples of learned attention maps in Fig. 9 used for spatial-wise weighting. As shown in the figure, the visible stream is trained to produce accurate saliency maps in the daytime, while the thermal stream also performs fairly well. However, at nighttime, the thermal stream generates much more accurate saliency maps than the visible stream does.

B. EXPERIMENTAL RESULTS WITH THE SANITIZED ANNOTATIONS

We note that MSDS-RCNN [28] created a sanitized version of KAIST training annotations and improved the detection performance significantly. For a fair comparison, we use the sanitized annotations provided by them to conduct experiments with our spatial-wise weighted fusion detector. The detection performance of ours improves from 11.43% MR to 7.38% MR (all-day: 7.38% MR, daytime: 8.22% MR, nighttime: 5.34% MR), which is comparable to that of MSDS-RCNN (all-day: 7.49% MR, daytime: 8.09% MR, nighttime: 5.92% MR).

REFERENCES

[1] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," in *Proc. IEEE Int. Conf. Learn. Represent.*, 2015, pp. 1–10.

[2] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2874–2883.

[3] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3286–3295.

[4] Y. Cao, D. Guan, Y. Wu, J. Yang, Y. Cao, and M. Y. Yang, "Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 70–79, Apr. 2019.

[5] Y. Chen, H. Xie, and H. Shin, "Multi-layer fusion techniques using a CNN for multispectral pedestrian detection," *IET Comput. Vis.*, vol. 12, no. 8, pp. 1179–1187, Dec. 2018.

[6] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4467–4475.

[7] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3029–3037.

[8] Y. Choi, N. Kim, S. Hwang, and I. S. Kweon, "Thermal image enhancement using convolutional neural network," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 223–230.

[9] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "KAIST multi-spectral day/night data set for autonomous and assisted driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 934–948, Mar. 2018.

[10] J. W. Davis and V. Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *Comput. Vis. Image Understand.*, vol. 106, nos. 2–3, pp. 162–182, May 2007.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[12] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[13] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[14] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, "Pedestrian detection in thermal images using saliency maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 988–997.

[15] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, and A. López, "Pedestrian detection at day/night time with visible and FIR cameras: A comparison," *Sensors*, vol. 16, no. 6, p. 820, Jun. 2016.

[16] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vázquez, J. Xu, and A. López, "Pedestrian detection at Day/Night time with visible and FIR cameras: A comparison," *Sensors*, vol. 16, no. 6, p. 820, Jun. 2016.

[17] 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," *Inf. Fusion*, vol. 50, pp. 148–157, Oct. 2019.

[18] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 345–360.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[21] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.

[22] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[23] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[24] N. Kim, Y. Choi, S. Hwang, and I. S. Kweon, "Multispectral transfer network: Unsupervised depth estimation for all-day vision," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6983–6991.

[25] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 845–853.

[26] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 49–56.

[27] S. J. Krotosky and M. M. Trivedi, "Person surveillance using visual and infrared imagery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1096–1105, Aug. 2008.

[28] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," 2018, *arXiv:1808.04818*. [Online]. Available: http://arxiv.org/abs/1808.04818

[29] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, pp. 161–171, Jan. 2019.

[30] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[31] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[32] J. Liu, S. Zhang, S. Wang, and D. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 73.1–73.13

[33] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[34] T. Liu, J.-J. Huang, T. Dai, G. Ren, and T. Stathaki, "Gated multi-layer convolutional feature extraction network for robust pedestrian detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3867–3871.

[35] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*. [Online]. Available: http://arxiv.org/abs/1506.04579

[36] Y. Liu, Y. Wang, S. Wang, T. Liang, Q. Zhao, Z. Tang, and H. Ling, "CBNet: A novel composite backbone network architecture for object detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 11653–11660.

[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[38] F. Massa and R. Girshick. (2018). *MaskrCNN-Benchmark: Fast, Modular Reference Implementation of Instance Segmentation and Object Detection Algorithms in PyTorch*. [Online]. Available: https://github.com/facebookresearch/maskrcnn-benchmark

[39] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.

[40] M. Oliveira, V. Santos, and A. D. Sappa, "Multimodal inverse perspective mapping," *Inf. Fusion*, vol. 24, pp. 108–121, Jul. 2015.

[41] K. Park, S. Kim, and K. Sohn, "Unified multi-spectral pedestrian detection based on probabilistic fusion networks," *Pattern Recognit.*, vol. 80, pp. 143–155, Aug. 2018.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[43] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[44] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3D object classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 656–664.

[45] K. Takumi, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada, "Multispectral object detection for autonomous vehicles," in *Proc. Thematic Workshops ACM Multimedia-Thematic Workshops*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 35–43.

[46] A. Torabi, G. Massé, and G.-A. Bilodeau, "An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications," *Comput. Vis. Image Understand.*, vol. 116, no. 2, pp. 210–221, Feb. 2012.

[47] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *Proc. ESANN*, 2016, pp. 509–514.

[48] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[49] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," 2019, *arXiv:1904.09146*. [Online]. Available: http://arxiv.org/abs/1904.09146

[50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[51] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[52] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.

[53] L. Zhang, Z. Liu, S. Zhang, X. Yang, H. Qiao, K. Huang, and A. Hussain, "Cross-modality interactive attention network for multispectral pedestrian detection," *Inf. Fusion*, vol. 50, pp. 20–29, Oct. 2019.

[54] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5127–5137.

[55] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.

[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

YONGTAO ZHANG received the B.S.E. degree in vehicle engineering from the Wuhan University of Technology, China, in 2018, where he is currently pursuing the M.S. degree in vehicle engineering with the School of Automotive Engineering.

His research interests include computer vision and intelligent connected vehicles.

ZHISHUAI YIN received the B.S.E. degree in vehicle engineering from Tsinghua University, China, in 2006, and the Ph.D. degree in interdisciplinary engineering from Northeastern University, USA, in 2013.

He is currently an Associate Professor with the School of Automotive Engineering, Wuhan University of Technology, China. His research interests include intelligent connected vehicles and driving behavior analysis.

LINZHEN NIE received the B.S.E. degree in control science and engineering from the Huazhong University of Science and Technology, China, in 2008, and the Ph.D. degree in interdisciplinary engineering from Northeastern University, USA, in 2013.

She is currently an Associate Professor with the School of Automotive Engineering, Wuhan University of Technology, China. Her research interests include intelligent connected vehicles and driving behavior analysis.

SONG HUANG received the B.S.E. degree in vehicle engineering from the Wuhan University of Technology, China, in 1978.

He is currently a Professor with the School of Automotive Engineering, Wuhan University of Technology. His research interests include intelligent connected vehicles and automotive design.