EasyChair Preprint

№ 6672

Pedestrian Detection by Fusion of RGB and Infrared Images in Low-Light Environment

Qing Deng, Wei Tian, Yuyao Huang, Lu Xiong and Xin Bi

September 23, 2021

Pedestrian Detection by Fusion of RGB and Infrared Images in Low-Light Environment

Qing Deng, Wei Tian*, Yuyao Huang, Lu Xiong, Xin Bi
Institute of Intelligent Vehicles, School of Automotive Studies
Tongji University, Shanghai, China
{dengqing, tian_wei, huangyuyao, xiong_lu, bixin}@tongji.edu.cn

***Abstract*—Pedestrian detection in low-light environment is an essential part for autonomous driving in all-day and all-weather situations. A current trend is utilizing multispectral information such as RGB and infrared images to detect pedestrians. Despite its efficacy, such an approach suffers from underperformance in dealing with varied object scales due to its limited feature fusion on semantic levels. To address the above problem, we propose a novel multi-layer fusion network called as MLF-FRCNN. In this network, multi-scale feature maps are created from RGB and infrared channels from each backbone block. A feature pyramid network module is further introduced to facilitate predictions on multi-layer feature maps. The experimental results on the KAIST Dataset reveal that our method achieves a runtime performance of 0.14s per frame and an average precision of 91.2% which outperforms state-of-the-art multispectral fusion methods. The effectiveness of our approach in dealing with scaled objects in low-light environment is further proven by ablation studies.**

***Index Terms*—pedestrian detection, multispectral multi-layer fusion, low light condition**

I. Introduction

With its wide application in autonomous driving, vision-based pedestrian detection has become a research focus in recent years. Given images captured in real-world traffic scenarios, the task of pedestrian detection is to distinguish between pedestrians and background and to locate individual pedestrian instances with bounding boxes. Though significant progress has been made over the past few decades, it's still a challenging task to design a robust pedestrian detector especially adapted to all-day and all-weather situations.

Current pedestrian detection approaches mostly utilize RGB images, which may not work well under low-light conditions, such as in the nighttime. Infrared images are regarded as solutions to overcome the above limitations. Since a long-wavelength infrared camera captures radiated heat from objects, infrared images can show clear human bodies even in low-light environment. However, they lost visual details which can be provided by RGB images [1]. Thus, RGB and infrared images provide complementary information about the target of interest. By fusing the information effectively, the precision of pedestrian detection can be enhanced.

Project supported by the National Natural Science Foundation of China (No.52002285), the Shanghai Pujiang Program (No.2020PJD075), the Natural Science Foundation of Shanghai (No.21ZR1467400) and the Key Special Projects of the National Key R&D Program of China (No.2018AAA0102800).
*Corresponding author.

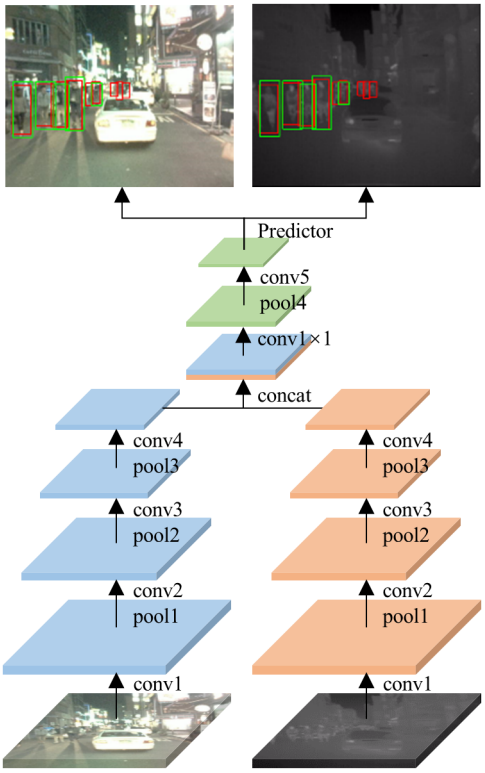


Fig. 1. Halfway Fusion model [3] for multispectral pedestrian detection. Detection results are shown at the top with green bounding boxes, while red bounding boxes denote the ground truth. Obviously, it manifests difficulty in detecting small pedestrians.

Hence, conventional detection approaches such as Faster R-CNN [2] have been adapted to multispectral inputs for pedestrian detection. However, existing fusion methods are commonly based on Halfway Fusion [3], which concatenates separate feature extractor for RGB and infrared images at fixed level and makes predictions on single-layer feature maps, as shown in Fig. 1 [3]. Due to the limited fusion level, this kind of methods manifests difficulty in detecting pedestrians with greatly varied scales.

To address the above problem, in this work, we propose a novel multi-layer fusion network based on Faster R-CNN for pedestrian detection in low-light environment, denoted as MLF-FRCNN. Our contributions are summarized as follows:

- To deal with pedestrian instances with greatly varied

sizes, we create multi-scale feature maps by fusing each block of feature extractor network for both RGB and infrared channels. Additionally, a feature pyramid network (FPN) [4] is introduced to facilitate predictions on multi-layer feature maps.
- We train and test our MLF-FRCNN on the KAIST Multispectral Pedestrian Dataset [5]. Experimental results under both daytime and nighttime light conditions reveal that an average precision (AP) of 91.2% is achieved by our MLF-FRCNN, which outperforms state-of-the-art fusion methods with a similar framework. Moreover, our method performs particularly well in the reasonable night and the small-scale settings, with an AP gain of 5.1% and 28.8%, respectively.
- We conduct further experiments to demonstrate the effectiveness of our proposed fusion method in comparison with single-modal approaches. Moreover, by training MLF-FRCNN only on the day subset and test its transferring ability during nighttime, we further reveal the fact that MLF-FRCNN is able to learn the adaptive fusion of RGB and infrared information according to illumination by the mixed training under different light conditions.

II. Related Works

A. Pedestrian Detection Based on RGB Images

Inspired by the success of convolutional neural networks (CNNs) in object detection tasks, pedestrian detection methods based on CNNs have been proposed in the past few years. The early studies to apply CNNs to pedestrian detection mostly utilized two-stage detectors of R-CNN series due to its robustness. Though Faster R-CNN has become a general network for object detection, the original version performed poorly when directly applied to pedestrian detection. Zhang et al. [6] indicated that the reason of the under-performance was low resolution of pedestrian instances and lack of guiding strategy. Appel et al. [7] achieved better performance by adding boosted forest on the top of feature maps extracted by region proposal network (RPN), represented as RPN+BF. It was proven in [8] that the performance of Faster R-CNN for pedestrian detection can be greatly improved by appropriate adjustments, such as designing a specific RPN scale for pedestrians. Li et al. [9] designed two sub-networks adapted to large-scale and small-scale pedestrians respectively, in order to suppress the decrease of precision caused by large variance of scales. Zhang et al. [10] studied different kinds of attention models, including self-attention, attention based on bounding boxes or on body parts. They revealed that the greatest effectiveness can be obtained by attention on body parts.

There are also a few researches on pedestrian detection using one-stage detectors. Liu et al. [11] proposed Asymptotic Localization Fitting (ALF), which improves detection results by gradually stacking a series of predictors in SSD [12]. Lin et al. [13] proposed a graininess-aware deep feature learning method, which introduced fine-grained information and utilized an attention mechanism to better distinguish human bodies. An anchor-free method was proposed in [14] to determine central points and scales of the pedestrian points at a high-level semantic abstraction. However, one-stage approaches are generally inferior to two-stage approaches on pedestrian detection precision.

Related researches also improved the training mode of networks. Mao et al. [15] proposed the joint learning of pedestrian targets and additional features. This multi-task training mode made use of prior information of given features to improve the performance during prediction stage without additional inputs. Brazil et al. [16] tried to enhance the detection precision through the joint supervision of pedestrian detection and semantic segmentation. The experimental results revealed that weak annotations of semantic segmentation can boost the precision improvement. Xu et al. [17] utilized infrared images as supervision for RGB input during training process while used only RGB images to extract cross-modal representations in prediction stage. However, the aforementioned methods using separate RGB images for detection suffer from significant underperformance in low-light environment.

B. Pedestrian Detection Based on Fusion of RGB and Infrared Images

Due to the robustness of infrared images in low-light environment, there is a growing research interest in pedestrian detection utilizing RGB and infrared images. The release of multispectral pedestrian datasets such as KAIST [5], UTokyo [18], CVC-14 [19], etc. boosted researches on fusion of RGB and infrared images. ACF+T+THOG [5] is the initial baseline of the KAIST Dataset, which was extended from traditional aggregated channel features (ACF) [20] to infrared channels. Based on two-stage R-CNN series, Choi et al. [21] generated region proposals using RPN on RGB and infrared images separately and obtained final results using support vector regression (SVR). Liu et al. [3] studied four different fusion models based on Faster R-CNN, in which features or detection results from separate RGB and infrared channels were fused in different stages. Their experiment proved that the Halfway Fuison model achieved the best performance. König et al. [22] extended RPN+BF [7] to multispectral pedestrian detection and proposed fusion RPN+BF. Li et al. [1] proposed Multispectral Simultaneous Detection and Segmentation R-CNN (MSDS-RCNN), which combined the detection task and the semantic segmentation task in both region proposal stage and detection head stage. As for one-stage detectors, the Halfway Fusion method was transferred to YOLOv3 [23] framework by Geng et al. [24]. Their comparative experiment showed that dual-modal Faster R-CNN outperformed dual-modal YOLOv3 in low-observable pedestrian detection.

Considering the advantage of RGB images during daytime and infrared images during nighttime, illumination information was introduced to guide the fusion. Guan et al. [25] proposed an illumination-aware network for pedestrian detection. On the one hand, the illumination was learned from feature maps. On the other hand, the whole network was divided into two sub-networks to learn pedestrian features during daytime

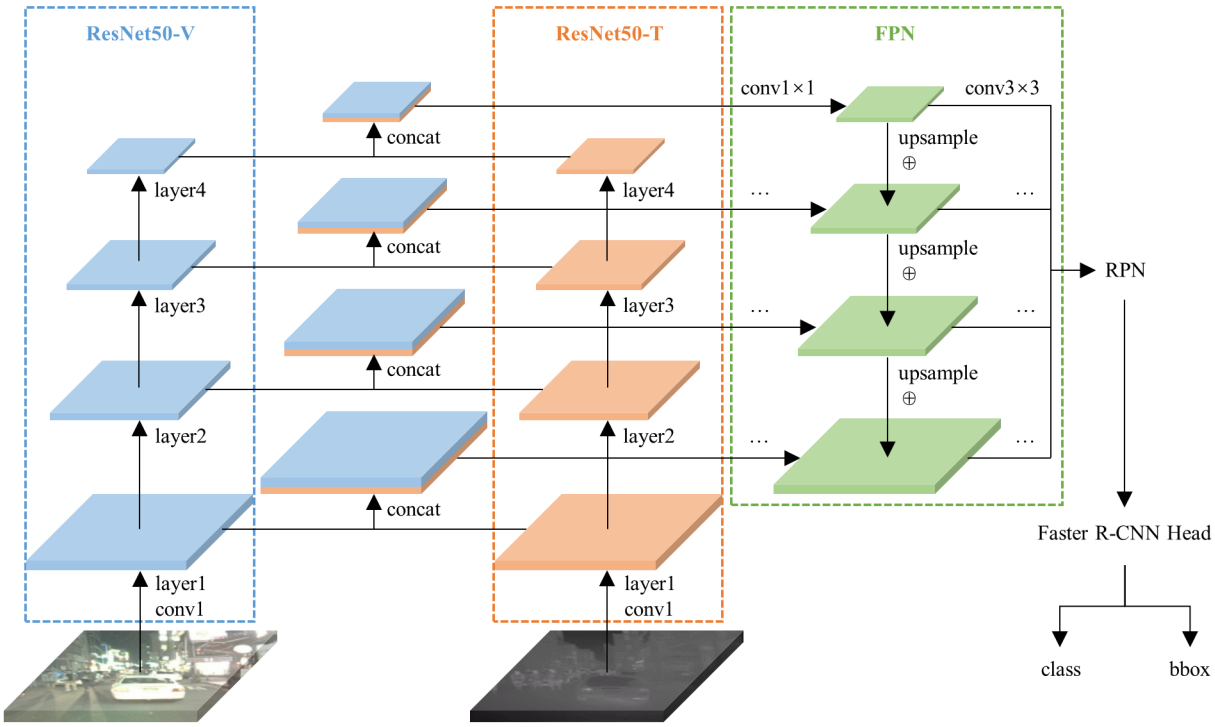


Fig. 2. Architecture of the proposed MLF-FRCNN.

and nighttime separately. The final result was obtained by weighting the output of the sub-networks according to the learned illumination. Li et al. [26] proposed Score Fusion II by adding novel fusion models to the work [3], which performs with minor gap to Halfway Fusion. Based on Score Fusion II they proposed the further improved network IAF R-CNN with an illumination-aware weighting mechanism. Since most multispectral pedestrian datasets only provide hard 0-1 labels of day/night, there is a lack of illumination details for training in this kind of methods.

Inspired by the feature pyramid method in object detection networks, fusion on multi-layer feature maps was proposed to adapt to varied scales of pedestrians. Chen et al. [27] designed a fused deconvolutional single-shot detector (MFDSSD), which extracted features from RGB and infrared images separately and fused them to generate strong representations for pedestrians. Another multispectral feature fusion network (MSFFN) is proposed by Song et al. [28], in which dual-modal features were fused after separately passing through the backbone and FPN module. However, their MSFFN can not integrate dual-modal information between multi-scale feature layers and becomes more complex. Our method is similar to MSFFN but differs from it by fusing features in backbone before sending them to the shared FPN module. By doing this, we achieve a comparative approach with reduced complexity.

III. Proposed Method

In this work, we boost the performance of pedestrian detection in low-light environment by designing a novel multi-layer fusion network based on Faster R-CNN, called as MLF-FRCNN. As shown in Fig. 2, MLF-FRCNN is composed of two parallel feature extractors, an FPN, an RPN and a detection head. The fusion network accepts RGB and infrared images as inputs and yields pedestrian detection results on both modalities.

A. Backbone

The backbone of MLF-FRCNN consists of two parallel ResNet50s (i.e., the ResNet50-V and ResNet50-T) [29], which extract features from RGB and infrared (thermal) images separately. We make a few adjustments on ResNet50 by reducing the feature stride and removing the max-pooling layer to better detect small pedestrian instances.

To process multispectral information, features from RGB and infrared channels are concatenated after each of the 4 blocks in ResNet50 (left side of Fig. 2). Then we obtain fused features on 4 layers of different scales, which are later sent to the FPN module. Such multi-layer fusion overcomes the limitation of fusion in a single stage, in which features only at a fixed semantic level are extracted. Multi-layer fusion can help the network learn to adaptively fuse complementary information from RGB and infrared features at different semantic levels. With the full use of information from RGB and infrared images, the network should be more robust to illumination variations, which is verified in experiments.

B. Fusion Module

The multi-scale fusion learning in MLF-FRCNN is done by an FPN module (right side of Fig. 2). In this module, it first reduces the channel number of fused features extracted from the backbone to 256 through a 1×1 convolutional layer on each feature layer. Then the fused features on the upper layer are upsampled and added to the lower layer stage by stage. After that, a 3×3 convolutional layer is introduced to extract features again to avoid the aliasing effect [4].

The FPN module is introduced into MLF-FRCNN due to two reasons. Firstly, the FPN can reduce the sensitivity of the network to different scales of pedestrians by combining detailed location information on the lower layers with abundant semantic information on the upper layers. Secondly, by utilizing the 1×1 convolutional layer in the beginning, FPN reduces the number of parameters and enhances the computational efficiency.

C. Detection Module

The MLF-FRCNN consists of a region proposal network (RPN) and a detection head, which are similar with Faster R-CNN. The RPN generates region proposals based on anchors on each feature layer. The detection head further makes predictions of class labels and bounding boxes on these proposals. Considering the aspect ratio of typical pedestrians, we discard the anchor ratio of 0.5 to facilitate the training and predicting speed [3].

The network is trained with a joint loss defined as:

$$L_{\text{total}} = L_{\text{rpn}} + L_{\text{det}}, \tag{1}$$

where L_{total} denotes the total loss; L_{rpn} is RPN loss; L_{det} is the detection head loss.

The RPN loss is defined as:

$$L(p_i, \boldsymbol{t}_i) = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{\text{reg}}(\boldsymbol{t}_i, \boldsymbol{t}_i^*), \tag{2}$$

where N_{cls} denotes the total sample number of a batch and N_{reg} corresponds to the number of positive samples; The predicted object probability on anchor i is indicated by p_i while p_i^* denotes its binary objectness label; $\boldsymbol{t}_i = [t_{ix}, t_{iy}, t_{iw}, t_{ih}]$ denotes the offset of the predicted bounding box relative to anchor i with center coordinates (x, y) and size (w, h) while $\boldsymbol{t}_i^*$ represents its ground truth; The binary cross entropy loss of classification L_{cls} is calculated by (3) while L_{reg} is the smooth L1 loss of localization calculated by (4); Weight λ is set to 1 to balance above two losses.

$$L_{\text{cls}}(p_i, p_i^*) = -[p_i^* \log p_i + (1 - p_i^*) \log(1 - p_i)] \tag{3}$$

$$L_{\text{reg}}(\boldsymbol{t}_i, \boldsymbol{t}_i^*) = \sum_{j \in \{x,y,w,h\}} \text{smooth}_{L_1}(t_{ij} - t_{ij}^*) \tag{4}$$

The detection head loss L_{det} is defined in a similar way as L_{rpn}.

IV. Experiments

A. Experimental Setup

1) Dataset: We conduct experiments on the KAIST Multispectral Pedestrian Dataset [5] because it provides a great number of aligned RGB-infrared image pairs with a resolution of 640×512 captured in various traffic scenarios during daytime and nighttime. The KAIST Dataset was divided into training and test videos. Following Liu et al. [3], we sample image pairs from training videos in every 2 frames and exclude the ones where no proper pedestrian exists as the training set, among which 3698 were captured during daytime and 2328 during nighttime. 2252 image pairs are sampled from test videos with a sampling rate of 1:20 as the test set, among which 1455 were captured during daytime and 797 during nighttime. We utilize the reasonable day/night, the small-scale and the all-scale settings on the test set for experiments. The reasonable setting contains fully visible or partially occluded pedestrians whose heights are larger than 55 pixels as defined in [5], while the small-scale and the all-scale settings contain pedestrians whose heights are smaller than 55 pixels and pedestrians of all scales, respectively. Since the original annotations contain some problematic bounding boxes, we utilize the improved annotations of the test set provided by [3] as well as the sanitized annotations of the training set provided by [1].

2) Evaluation Metric: As a common metric in pedestrian detection, we use AP50 to evaluate the detection performance in our experiments. In AP50 metric, a predicted bounding box is judged as true positive (TP) if its Intersection over Union (IoU) with a ground truth is over 50%. Unmatched predicted bounding boxes and unmatched ground truths are judged as false positives (FP) and false negatives (FN), respectively. The precision and the recall are calculated by $\text{TP}/(\text{TP} + \text{FP})$ and $\text{TP}/(\text{TP} + \text{FN})$, respectively. Finally, AP50 is calculated by averaging the precision at equally spaced recalls between 0 and 1, which can be obtained by changing the threshold of classification scores.

3) Training Details: During the training process, in RPN, an anchor is set as positive if it has the maximum IoU or an IoU over 0.7 with a ground truth. Conversely, an anchor is set as negative if its IoU with all ground truths are under 0.3. While in the detection head, a region proposal is set as positive if its IoUs with a ground truth is over 0.5, and otherwise negative. We randomly sample 256 anchors per image with a ratio of 1:1 between positive and negative samples in RPN to compute the loss while for the detection head we use 512 proposals per image with a positive-to-negative ratio of 1:3. The proposed MLF-FRCNN model is initialized by a Faster R-CNN model with ResNet50+FPN pre-trained on the MS COCO dataset [30]. All other layers are initialized by the "Kaiming Uniform" method [31]. The entire model is implemented by PyTorch [32]. The batch size for training is set to 4. The optimizer is stochastic gradient descent (SGD) [33] with a momentum of 0.9 and a weight decay of 0.0005. The initial learning rate is set to 0.005 with a warm-up strategy.

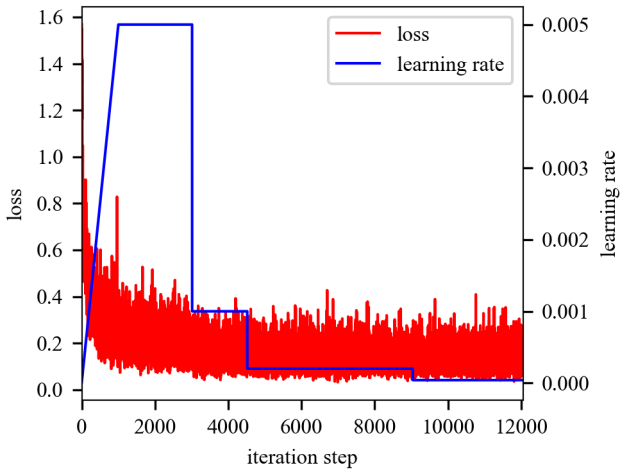


Fig. 3. The variations of the loss and the learning rate during the training process.

TABLE I
EXPERIMENTAL RESULTS OF MLF-FRCNN WITH DIFFERENT BACKBONES ON THE KAIST TEST SET IN THE REASONABLE SETTING.

Backbone	AP50	Prediction Time
VGG16 [34]	88.0%	0.10s/frame
ResNet18 [29]	77.5%	0.09s/frame
ResNet34 [29]	78.8%	0.10s/frame
ResNet50 [29]	91.2%	0.14s/frame

To avoid gradient explosion, we divide the learning rate by 5 after the second, the third as well as the sixth epoch and finish training after 8 epochs. The variations of the loss and the learning rate during the training process are shown in Fig. 3.

B. Comparison of Different Backbones

We first conduct experiments on MLF-FRCNN with different backbones. The experimental results on the KAIST test set in the reasonable setting are shown in Table I, from which we can find out that ResNet50 [29] achieves the highest AP at the cost of only 0.04s extra prediction time per frame compared with the second highest VGG16 [34]. We attribute this to the residual structure which allows deep layers to capture more complex features while avoids vanishing gradient. Therefore, we select ResNet50 as the backbone of MLF-FRCNN in following experiments.

C. Comparison with State-of-the-arts

Here we evaluate the proposed MLF-FRCNN on the KAIST test set mentioned in section IV-A1, in comparison with state-of-the-art fusion methods based on Faster R-CNN, as shown in Table II. We can find out that our MLF-FRCNN achieves the highest AP of 91.2% in the reasonable setting and also outperforms other fusion methods in three of the following four settings. Although our method performs with 0.9% inferior to the MSDS-RCNN [1] in the reasonable day setting, it performs particularly well in the reasonable night and the small-scale settings, with an AP gain of 5.1% and 28.8%, respectively. This indicates that the proposed MLF-FRCNN yields much better adaptability to light conditions and pedestrian scales. Moreover, MLF-FRCNN shows a fast prediction speed with 0.14s per frame and much less parameters with 251MB while the MSDS-RCNN has a much larger model of 3482MB. It's also worth mentioning that IAF R-CNN and MSDS-RCNN add illumination information and semantic segmentation as additional supervision, respectively. Our approach only conducts detection task without additional supervision and already outperforms above methods in most situations. A further improvement in AP is also conceivable by integrating illumination or semantic information in our model.

D. Ablation Studies

To prove the effectiveness of our multi-layer fusion method, we train and test single-modal (only with RGB or infrared images as input) Faster R-CNN with FPN module in the reasonable day/night settings. We also train MLF-FRCNN only on the day subset and test its transferring ability. The test results are shown in Table III and the precision-recall curves are plotted in Fig. 4.

We can find out that the model only with RGB input performs relatively better during daytime while the model only with infrared input performs relatively better during nighttime. By fusion of RGB and infrared images, we achieve the best performance during both daytime and nighttime. In the reasonable night setting, our MLF-FRCNN outperforms RGB-based Faster R-CNN (with FPN) by 36.4% in AP and infrared-based one by 3.1% in AP, respectively. The effectiveness of proposed fusion method in low-light environment is demonstrated.

We can also see that MLF-FRCNN trained only on the day subset performs even worse than infrared-based Faster R-CNN (with FPN) by 12.8% in AP in the reasonable night setting. We believe that it's because MLF-FRCNN tends to learn a fusion mode dominated by RGB features without training images captured in low-light environment. When transferred to the night scene, the model can't take great advantage of infrared features. From this we further conclude that MLF-FRCNN is able to learn adaptive fusion of RGB and infrared information according to illumination by the mixed training under different light conditions.

E. Qualitative Detection Results

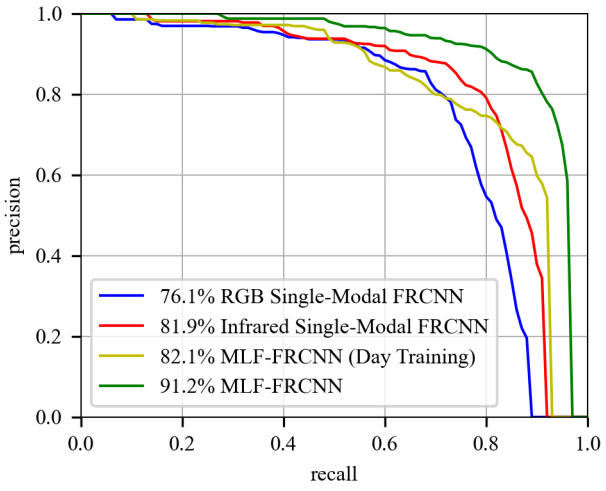
Fig. 5 shows the comparison of detection results on three RGB-infrared image pairs under different light conditions sampled from the KAIST test set. The first column shows ground truths. The other four columns show detection results of RGB-based Faster R-CNN (with FPN), infrared-based one, Halfway Fusion FRCNN [3] and our MLF-FRCNN, respectively. Green bounding boxes denote detection results while red bounding boxes denote ground truths. Obviously, it can be seen that the proposed MLF-FRCNN achieves more accurate detection especially in low-light environment.

TABLE II
EXPERIMENTAL RESULTS OF MLF-FRCNN COMPARED WITH STATE-OF-THE-ART FUSION METHODS BASED ON FASTER R-CNN ON THE KAIST TEST SET.

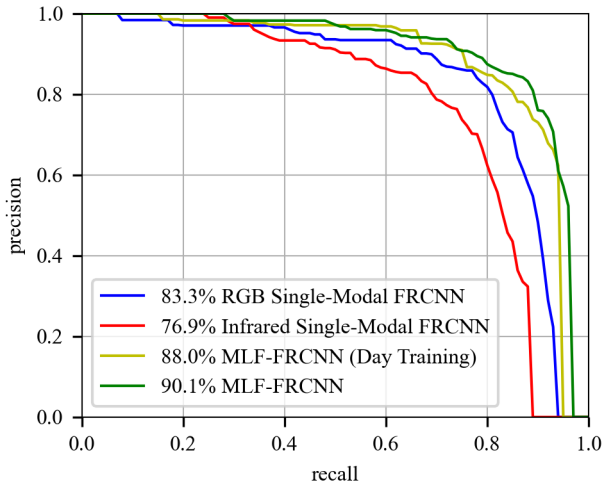
Methods	AP50					Prediction Time
	Reasonable	*Reasonable Day*	*Reasonable Night*	*Small-Scale*	*All-Scale*	
ACF+T+THOG [5] (baseline)	64.2%	68.9%	54.9%	14.9%	38.8%	0.13s/frame
Halfway Fusion FRCNN [3]	83.2%	83.2%	83.8%	38.5%	59.1%	0.16s/frame
Fusion RPN+BF [22]	86.2%	86.3%	85.9%	33.1%	54.3%	0.80s/frame
IAF R-CNN [26]	87.2%	89.0%	83.7%	33.2%	56.2%	0.21s/frame
MSDS-RCNN [1]	90.6%	**91.0%**	89.9%	46.2%	63.8%	0.15s/frame
MLF-FRCNN (ours)	**91.2%**	90.1%	**95.0%**	**75.0%**	**78.0%**	0.14s/frame

TABLE III
ABLATION STUDIES OF MLF-FRCNN.

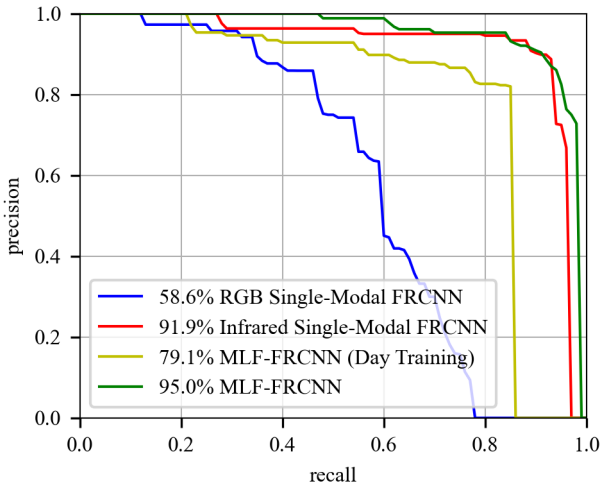
Input		Training		AP50		
RGB	*Infrared*	*Day*	*Night*	*Reasonable*	*Reasonable Day*	*Reasonable Night*
✓		✓	✓	76.1%	83.3%	58.6%
	✓	✓	✓	81.9%	76.9%	91.9%
✓	✓	✓		82.1%	88.0%	79.1%
✓	✓	✓	✓	91.2%	90.1%	95.0%



(a) Reasonable

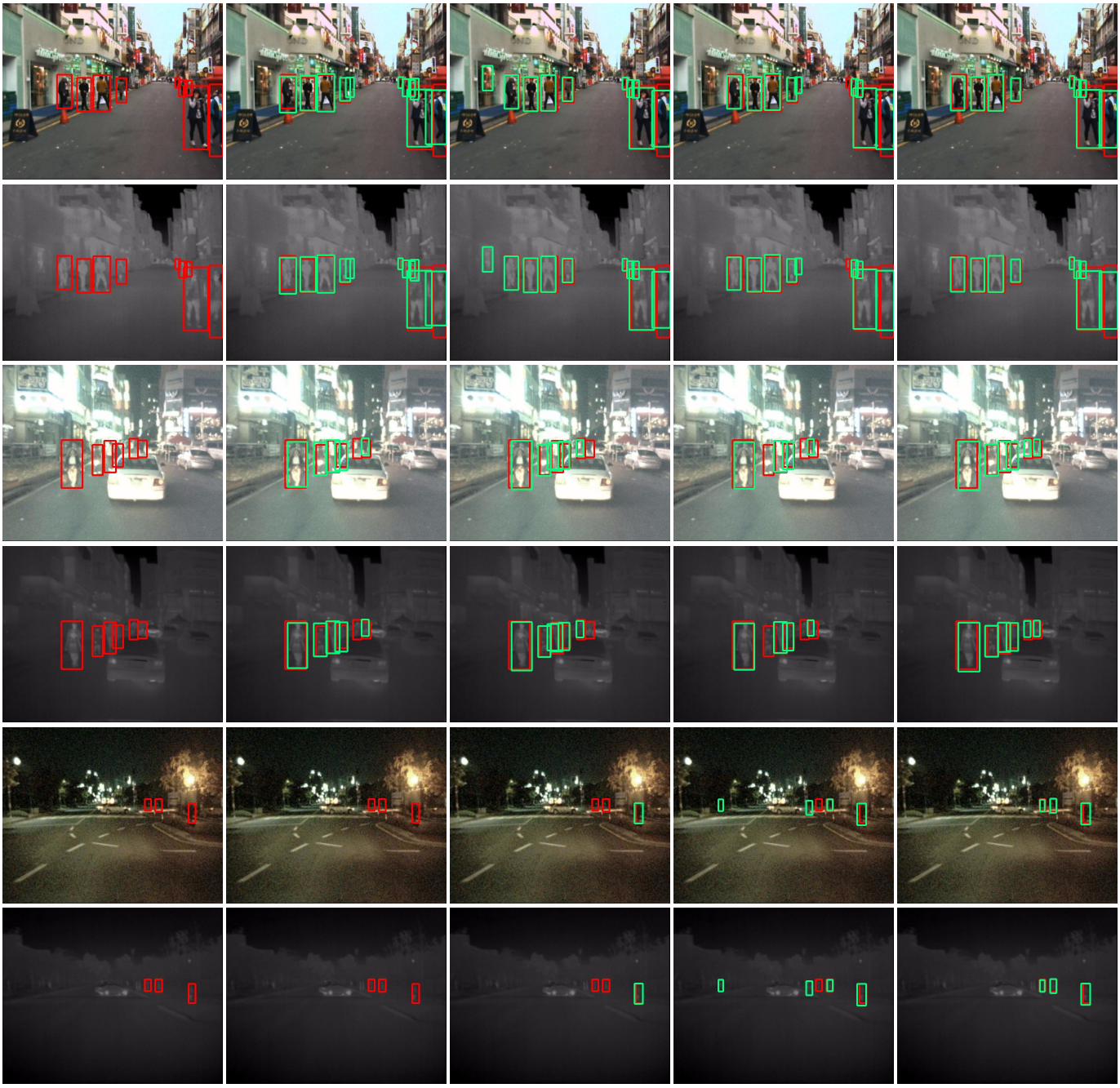


(b) Reasonable Day



(c) Reasonable Night

Fig. 4. The precision-recall curves of RGB-based Faster R-CNN (with FPN), infrared based one, MLF-FRCNN trained only on the day subset and MLF-FRCNN with full set training. The settings of tested set are: (a) Reasonable; (b) Reasonable Day; (c) Reasonable Night.



(a) Ground Truth | (b) RGB Single-Modal FRCNN | (c) Infrared Single-Modal FRCNN | (d) Halfway Fusion FRCNN [3] | (e) Our MLF-FRCNN

Fig. 5. The comparison of detection results on three RGB-infrared image pairs under different light conditions sampled from the KAIST test set. The first column shows (a) Ground Truth. The other four columns show detection results of (b) RGB Single-Modal FRCNN; (c) Infrared Single-Modal FRCNN; (d) Halfway Fusion FRCNN [3]; (e) Our MLF-FRCNN. Green bounding boxes denote detection results while red bounding boxes denote ground truths.

V. Conclusion

In this paper, we propose a novel multi-layer fusion network, namely MLF-FRCNN, for pedestrian detection in low-light environment. In this approach, we deploy feature maps extracted from RGB and infrared channels in different backbone blocks to extract multi-scale features. Additionally, an FPN module is further introduced to facilitate predictions on multi-layer feature maps. The experimental results on the KAIST Multispectral Pedestrian Dataset reveal that the proposed MLF-FRCNN outperforms state-of-the-art fusion approaches and especially effective in detecting pedestrians with greatly varied scales. The results also prove that MLF-FRCNN can fuse RGB and infrared information adaptively according to illumination, which is valuable for autonomous driving in all-day and all-weather situations. For the future research, we plan to add extra supervision like illumination and semantic segmentation in our model and transfer our multi-layer fusion method to the latest high-performing object detection framework to further improve the detection performance.

References

[1] C. Li, D. Song, R. Tong, and M. Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[2] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[3] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[4] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.

[5] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1037–1045.

[6] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?" in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 443–457.

[7] R. Appel, T. Fuchs, P. Dollár, and P. Perona, "Quickly boosting decision trees–pruning underachieving features early," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2013, pp. 594–602.

[8] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3213–3221.

[9] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2017.

[10] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6995–7003.

[11] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 618–634.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 21–37.

[13] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 732–747.

[14] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5187–5196.

[15] J. Mao, T. Xiao, Y. Jiang, and Z. Cao, "What can help pedestrian detection?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3127–3136.

[16] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4950–4959.

[17] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5363–5371.

[18] K. Takumi, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, and T. Harada, "Multispectral object detection for autonomous vehicles," in *Proceedings of the Thematic Workshops of ACM Multimedia*, 2017, pp. 35–43.

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

[20] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[21] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 621–626.

[22] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 49–56.

[23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[24] K. Geng, W. Zou, G. Yin, Y. Li, Z. Zhou, F. Yang, Y. Wu, and C. Shen, "Low-observable targets detection for autonomous vehicles based on dual-modal sensor fusion with deep learning approach," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of automobile engineering*, vol. 233, no. 9, pp. 2270–2283, 2019.

[25] 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*, vol. 50, pp. 148–157, 2019.

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

[27] Y. Chen and H. Shin, "Multispectral image fusion based pedestrian detection using a multilayer fused deconvolutional single-shot detector," *JOSA A*, vol. 37, no. 5, pp. 768–779, 2020.

[28] X. Song, S. Gao, and C. Chen, "A multispectral feature fusion network for robust pedestrian detection," *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 73–85, 2021.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[33] M. Zinkevich, M. Weimer, A. J. Smola, and L. Li, "Parallelized stochastic gradient descent." in *Advances in Neural Information Processing Systems (NIPS)*, vol. 4, no. 1. Citeseer, 2010, p. 4.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.