Deep Visible and Thermal Image Fusion with Cross-Modality Feature Selection for Pedestrian Detection

Mingyue Li, Zhenzhou Shao, Zhiping Shi, Yong Guan

Deep Visible and Thermal Image Fusion with Cross-Modality Feature Selection for Pedestrian Detection ⋆

Mingyue Li[1,2], Zhenzhou Shao[1,2*[0000−0002−9166−9468]], Zhiping Shi[1,3[0000−0002−3562−8602]], and Yong Guan[1,2,3[0000−0002−2373−2779]] ⋆⋆

[1] College of Information Engineering, Capital Normal University, Beijing, China
[2] Beijing Key Laboratory of Light Industrial Robot and Safety Verification
[3] Beijing Advanced Innovation Center for Imaging Technology
{2181002024, zshao, shizp, guanyong}@cnu.edu.cn

Abstract. This paper proposes a deep RGB and thermal image fusion method for pedestrian detection. A two-branch structure is designed to learn the features of RGB and thermal images respectively, and these features are fused with a cross-modality feature selection module for detection. It includes the following stages. First, we learn features from paired RGB and thermal images through a backbone network with a residual structure, and add a feature squeeze-excitation module to the residual structure; Then we fuse the learned features from two branches, and a cross-modality feature selection module is designed to strengthen the effective information and compress the useless information during the fusion process; Finally, multi-scale features are fused for pedestrian detection. Two sets of experiments on the public KAIST pedestrian dataset are conducted, and experimental results show that our method is better than the state-of-the-art methods. The robustness of fused features is improved, and the miss rate is reduced obviously.

Keywords: Pedestrian detection · Cross-modality features · Feature fusion.

1 Introduction

As a fundamental task in the field of computer vision, object detection has drawn much more attentions in several applications, such as autonomous driving, video surveillance, human-computer interaction, etc. Deep learning-based method [1–3] has made great progress using visible images (*e.g.*, RGB image) in

⋆ Supported by National Key R & D Program of China (2019YFB1309900), National Natural Science Foundation of China (61702348, 61772351), Beijing Nova Program of Science and Technology (Z191100001119075), the National Technology Innovation Special Zone (19-163- 11-ZT-001-005-06) and Academy for Multidisciplinary Studies, Capital Normal University(19530012005).

⋆⋆ *Corresponding author.

recent years. However, considering the adverse environmental conditions, i.e., the visible information is partially or fully missed under the poor lighting condition at night, the accuracy of detection using only RGB image becomes relatively low. Therefore, accurate object detection under adverse environmental conditions is still a challenging problem.

Recently, thermal images have been widely used for facial recognition [4, 5], human tracking [6, 7] and action recognition [8, 9] due to its robustness of biological characteristics. In particular, compared with the visible images, night-time thermal images provide more usable information without the need of enough illumination, so that both modalities are combined accordingly for the multispectral object detection [10–15]. The complementary relationship between both modalities has been proven [12], it paves an alternative way for object detection in the harsh environment, and provides new opportunities for around-the-clock applications.

In this paper, we mainly focus on the pedestrian detection using RGB and thermal images. Motivated by the complementary nature between modalities, extensive research efforts have been made. Hwang *et al.* [10] proposed an extended ACF method that uses aligned RGB and thermal images for all-weather pedestrian detection. With the latest development of deep learning, CNN-based methods [11, 16–18] have significantly improved the performance of object detection based on RGB and thermal image fusion. Liu *et al.* [19] adopted the Faster R-CNN architecture and analyzed the impact of different fusion stages in CNN on the detection results. Kéonig *et al.* [20] employed Region Proposal Network (RPN) and Boosted Forest (BF) frameworks for multispectral data detection. Kihong *et al.* [13] adopted a multi-branch detection model and also introduced a cumulative probability fusion (APF) layer to combine the results from different modes at the regional proposal layer. Zhang *et al.* [15] proposed a regional feature alignment (RFA) module to capture the position offset and adaptively align the regional features of these two modalities to improve the robustness of multi-modal detection. Xu *et al.* [11] first used a deep convolutional network to learn nonlinear mapping, modeled the correlation between RGB and thermal image data, and then transferred the learned feature representation to the second deep network in this way Learned that poor lighting conditions have the characteristics of discrimination and robustness, and it also proves that RGB and thermal image fusion has the possibility of all-weather detection.

However, the aforementioned methods commonly use the channel addition or cascade as the fusion strategy, the confidence of corresponding features from RGB and thermal images is not taken into account, and it cannot guarantee the complementary characteristics between features after the fusion.

In this paper, a pedestrian detection network that can perform cross-modal fusion feature selection is designed for the above-mentioned problems. The main contributions of this work are summarized as follows:

(1) We designed a two-stream fusion network for pedestrian detection. Two branch networks with residual structure with feature squeeze-excitation (SE) [21] modules are used to learn the features of RGB and thermal image data.

(2) The fusion of two modal data will have useless redundant information. Therefore, a cross-modal fusion feature selection mechanism is proposed to extract useful information and compress useless information.

2 Proposed Method

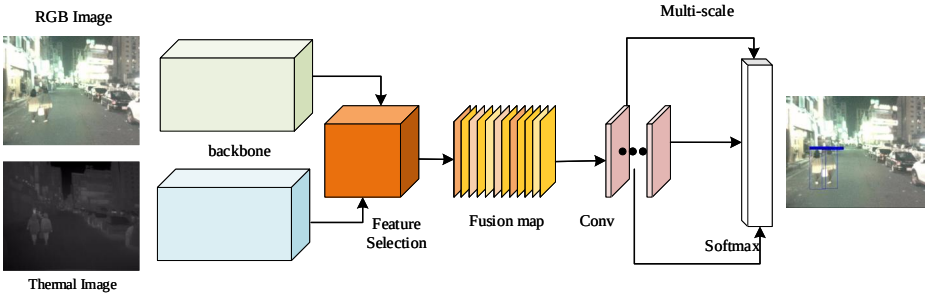

Fig. 1: Overview of the proposed method.

As shown in Fig. 1, the proposed model consists of two parts: two-branch feature extraction backbone network and cross-modal fusion feature selection module. We use the paired RGB and thermal image as the input of the two branches, and the corresponding features are extracted respectively using a two-branch backbone network. Then pass the learned features to the cross-modal feature selection module. Finally, multi-scale convolution operations are applied to fused features for further pedestrian detection.

2.1 Two-Branch Feature Extraction Backbone Network with Squeeze-Excitation Module

In order to extract the representative features of RGB and thermal images, we design a backbone network to extract features for each modality. In the backbone network, in order to improve the efficiency of feature extraction and alleviate the overfitting problem in the deep network [24], similar to yolov3 [25], we add a residual structure during feature extraction. Furthermore, to reduce the useless features for pedestrian detection, squeeze-excitation (SE) module is employed in the residual structure. It is able to improve the efficiency of feature extraction as well.

As shown in Fig. 2, the squeeze-excitation module is combined with ResNet structure. X represents the input feature, and Y represents the output feature. The input layer features first go through two convolutional layer operations of 1×1 and 3×3, and then perform a global pooling operation to obtain a vector representing the importance of a single-modal data channel, which can be

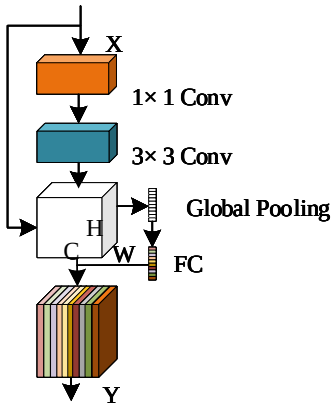

Fig. 2: ResNet structure with squeeze-excitation module.

calculated by the following formula:

$$Z_c = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j), \tag{1}$$

where Z_c represents the calculated channel importance parameter, and W and H represent the height and width of the feature map, respectively. The formula obtains Z_c by adding up all the characteristic points in the mean value. Finally, the importance parameters calculated from all channels are fully connected to obtain the feature channel importance vector.

In order to use the vector obtained by channel squeeze, we multiply the feature output from the ResNet layer with the vector, pass to the ReLu activation function, and finally obtain the output Y through the Softmax activation function, the formula is as follows:

$$Y = s(\delta(Z * X')), \tag{2}$$

where Y represents the final output feature, s denotes Softmax activation function, Z represents the channel importance vector, δ is ReLu activation function, and X'represents the feature obtained after convolution operations with kernel size of 1×1 and 3×3, respectively.

2.2 Cross-Modality Feature Selection Module

After two-modality features are extracted using the backbone network, the effective fusion is carried out. Although multi-modality features have complementary information, they also have mutually redundant information. Therefore, both the complementary and redundant characteristics of RGB and thermal images are taken into account, a cross-modality feature selection module is designed

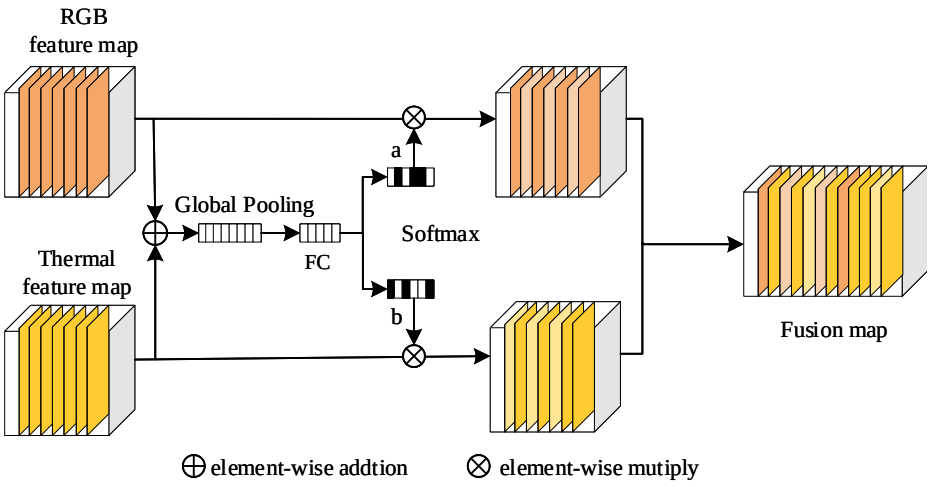

Fig. 3: Cross-modal fusion feature selection module.

to reduce the interference of redundancy or error information on subsequent detection.

As shown in Fig. 3, RGB feature map and thermal feature map respectively represent the feature maps of the two modalities output by the two backbone networks. Global Pooling represents the global pooling operation, FC represents the fully connected layer, Softmax represents the activation function, and finally Fusion map represents the final fusion feature map. The specific operations of the cross-modal fusion feature selection module are as follows: After RGB and thermal image data are extracted through the branch backbone network, the features of the two modalities are first added together to obtain the fusion feature map used to calculate the importance parameters of the two modal data. The preliminary fusion feature map obtained by reuse is subjected to global pooling, full connection, and Softmax operations. The importance parameters of RGB and thermal image features are obtained by the following formula:

$$U = U_{RGB} + U_{Thermal}, \tag{3}$$

where U represents the preliminary fusion feature obtained by adding RGB and Thermal features, and U_{RGB} and $U_{Thermal}$ represent the RGB and Thermal features output by the backbone network, respectively.The global tie pooling operation is written as

$$S_c = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i, j), \tag{4}$$

where H and W represent the height and width of the feature map, and U_c denotes the feature map at cth channel.

$$Z = F_{fc}(S) = \delta(\beta(S)). \tag{5}$$

The parameters obtained by global pooling are used to obtain the fusion feature channel importance vector through full connection, ReLu, and normalization operations. fc represents the fully connected operation, δ represents the ReLu operation, and β represents the normalization operation.

$$a_c = \frac{e^{A_c Z}}{e^{A_c Z} + e^{B_c Z}}, b_c = \frac{e^{B_c Z}}{e^{A_c Z} + e^{B_c Z}}. \tag{6}$$

Eq. (6) indicates that a method similar to the attention mechanism is used to obtain the respective importance vectors of RGB and thermal features according to the obtained feature channel importance vectors. Among them, A_c and B_c represent the characteristics of RGB and Thermal channel count as c respectively.

$$U_{Fuse_c} = a_c * U_{RGB_c} + b_c * U_{Thermal_c}, a_c + b_c = 1, \tag{7}$$

where represents the multiplication operation of RGB and thermal with the obtained importance vector, and then adding them to obtain features frome the final fusion layer. U_{Fuse_c}, U_{RGB_c} and $U_{Thermal_c}$respectively represent the fused features, RGB and thermal features at channel c.

2.3 Optimization

We deploy all parameter calculations to run on GPU devices. In the network model training, the stochastic gradient descent optimization algorithm is adopted. In addition, the use of gradient descent in deep learning model training is itself an approximate solution problem, and asynchronous parallel computing is more efficient than synchronous parallel computing in approximate solution problem, so we use asynchronous parallel computing mode in training model for computation on GPU.

3 Experimental Results

In order to verify the effectiveness of proposed method, we conducted several experiments on the publicly available KAIST pedestrian dataset [25] captured in various traffic scenarios with different lighting conditions. The dataset consists of 95,000 aligned RGB thermal image pairs. One pair of images are sampled every 20 frames from the whole KAIST dataset for training and testing. In our experiment, 7,472 pairs are collected as training samples and 1386 pairs for testing.

3.1 Experimental setup

The proposed method is implemented in the Tensorflow framework, running on two NVIDIA Tesla P100 GPUs with 16G memory. Regardless of changes in day and night or lighting conditions, we put all the selected data together for training and testing. In order to ensure the adequacy of data during the experiment, data augmentation operations (rotation and zoom) are implemented. 30 epochs are set per experiment in the training stage.

The experiment is carried out in a batch format. Each small batch consists of 6 pairs of images, which are randomly selected from the training images. Stochastic gradient descent is used to optimize the model, and its weight attenuation parameter is set to 0.995. The initial learning rate is set to 1e-4, and then reduce the learning rate after the model becomes stable, so as to achieve the goal of avoiding missing the optimal solution.

3.2 Comparison with the State-of-the-Art Methods

We compare our method with advanced methods on the KAIST pedestrian dataset, including: (*i*)ACF-RGB [27], which uses ACF for RGB data; (*ii*) ACF-RGBT + THOG [25], that is, using ACF for RGB thermal data with HOG characteristics; (*iii*) CMT-CNN [11], a detection network that learns cross-modal deep representation; (*iv*) SDS-RCNN [23], which is a convolutional neural network based on simultaneous detection and segmentation. As shown in Table 1, the proposed method outperforms the state-of-the-art methods, and the miss rate is greatly reduced by 39.2%, compared with ACF-RGBT. The performance is comparable with SDS-RCNN.

Table 1: Comparison of different methods. We use miss rate as the evaluation parameter for comparison on the KAIST dataset.

Method	ACF-RGB	ACF-RGBT-HOG	CMT-CNN	SDS-RCNN	**Ours**
Miss Rate	76.16	54.82	49.55	47.44	**46.32**

Fig. 4 illustrates a part of testing results compared with the existing methods. From left to right, the detection results of ACF-RGB, ACF-RGBT-HOG, CMT-CNN, SDS-RCNN and our method are shown in sequence. It can be observed that ACF-RGB method misses some targets, wrong detections occurs. Although ACF-RGBT-HOG has a lower error rate than ACF-RGB, there are more cases of missing detection targets. The detected targets are basically correct, and a small number of targets are not detected using CMT-CNN and SDS-RCNN methods. The detection results using the proposed method is shown in the last column, where the quantities of both wrong and missing targets are reduced obviously.

Fig. 4: Comparison of the detection results of existing methods and our method. From left to right are the detection results of ACF-RGB, ACF-RGBT-HOG, CMT-CNN, SDS-RCNN and our method.

3.3 Ablation Study

In this section, a set of experiments are conducted to prove the effectiveness of the proposed single-modal feature squeeze-excitation (SE) and multi-modal fusion feature selection module. We analysis the proposed methods under 3 different settings: (1) Adjust the number of Res structures at each layer and find out the best backbone network model. (2) Add SE layer to the Res structure, add the SE layer to the Res and fusion layer, and add the SE layer to the fusion layer. (3) Experiments are performed on the fusion layer plus the cross-modal feature selection (FS) layer and the fusion layer without the cross-modal fusion feature selection layer. The detection accuracy rate AP is used as the evaluation matrix.

Table 2: Experimental results of KAIST dataset under different module settings.

	Proposed Method				
Res:1,1,1,1,1	√	–	–	–	–
Res:1,2,2,2,2	–	√	√	√	√
Res+SE	–	–	√	√	√
Fusion+SE	–	–	–	√	–
Fusion+FS	–	–	–	–	√
AP(%)	46.61	51.41	52.85	52.42	53.68

Table 2 shows our comparison results. Use the detection accuracy rate AP as the evaluation parameter. We use the KAIST dataset to conduct ablation research on the model framework of this article. In the process of ablation research,

we comprehensively considered the complexity of the network model and the accuracy of experimental results, and appropriately reduced the residual structure on the basis of the original yolov3 network. Reduce residual structure when considering the RGB and Thermal data feature similarities, in terms of characteristics of reuse use too much residual structural redundancy may be produced, and considering the residual structure can reuse low-level features to a certain extent, alleviate the problem of gradient vanishing. Therefore, the number of residual structures is reduced to (1,1,1,1,1) and (1, 2, 2, 2, 2) for experimental comparison. It can be seen from Table 2 that the effect of residual structure in the model is significantly better than that of (1, 1, 1, 1, 1) when the number of residual structure is (1, 2, 2, 2, 2). In addition, considering that not all features of RGB and thermal data are valid, the characteristic squeeze-excitation module (SE), which can extract effective information and compress unwanted information, is added into the method proposed in this paper. In the experiment, we compared the adding location of feature squeeze-excitation module. The experimental results show that the effect of adding feature squeeze-excitation module after the residual structure is better, because the low-level features reused by the residual structure are similar to the high-level features, the superposition features generate redundant information, and the addition of feature squeeze-excitation module produces better detection results. Finally, we use the feature selection module of cross-mode fusion proposed by us. Because the data of RGB and thermal are similar, the detection accuracy of single mode features extracted by effective feature compression is improved after the fusion selection. In this paper, we propose a two-branch detection network that adds a feature squeeze-excitation module after the residual structure and a cross-modal fusion feature selection module to the fusion layer, which improves the detection accuracy and presents better detection results.

Fig. 5: Comparison of testing results without and with cross-modality feature selection module.

As shown in Fig. 5, the first row is the detection results without cross-modality feature selection module, while the second one illustrates the results of our proposed method. It is obviously observed that pedestrian targets may not be detected or incomplete before the cross-modality feature selection module is applied. Our method can detect almost all pedestrian targets even when the light conditions are not good.

4 Conclusion

This paper proposes a cross-modal feature selection pedestrian detection method based on RGB and thermal images. The residual structure is used when extracting features from RGB and thermal image data, and the feature squeeze-excitation module is embedded after the residual structure. Finally, more practical fusion features are obtained using the cross-modal fusion feature selection module, which effectively improves the detection accuracy. In addition, we also adopted a multi-scale detection method to improve the detection effect of small targets. A set of experiments are carried out on the public KAIST pedestrian dataset. Experimental results demonstrate that the proposed method outperforms the state-of-the-art methods, even when the lighting conditions are not particularly well or there exist many small targets in the scene of detection.

References

1. Bin Yang, Jun jie Yan, Zhen Lei, and Stan Z Li. Convolutional channel features. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 82–90, 2015.
2. Liliang Zhang, Liang Lin, Xiao dan Liang, and Kai ming He.Is faster r-cnn doing well for pedestrian detection? In Proceedings of the European conference on Computer Vision (ECCV), pages 443–457, 2016.
3. Jia nan Li, Xiao dan Liang, Sheng Mei Shen, Ting fa Xu, Jia shi Feng, and Shui cheng Yan. Scale-aware fast r-cnn for pedestrian detection. IEEE Transactions on Multimedia, 20(4):985–996, 2018.
4. Pradeep Buddharaju, Ioannis T Pavlidis, Panagiotis Tsiamyrtzis, and Mike Bazakos. Physiology-based face recognition in the thermal infrared spectrum. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 29(4):613–626, 2007.
5. Seong G Kong, Jingu Heo, Faysal Boughorbel, Yue Zheng, Besma R Abidi, Andreas Koschan, Ming zhong Yi, and Mongi A Abidi. Multiscale fusion of visible and thermal ir imag-es for illumination-invariant face recognition. International Journal of Computer Vision, 71(2):215–233, 2007.
6. Alex Leykin, Yang Ran, and Riad Hammoud. Thermal visible video fusion for moving target tracking and pedestrian classification. In Proceedings of the IEEE Conference on Com-puter Vision and Pattern Recognition (CVPR), pages 1–8, 2007.
7. Atousa Torabi, Guillaume Massé, and Guillaume-Alexandre Bilodeau. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. Computer Vision and Image Understanding, 116(2):210–221, 2012.

8. Yu Zhu and Guo dong Guo. A study on visible to infrared action recognition. IEEE Signal Processing Letters, 20(9):897–900, 2013.
9. Chen qiang Gao, Yin he Du, Jiang Liu, Jing Lv, Lu yu Yang, De yu Meng, and Alexander G Hauptmann. Infar dataset: Infrared action recognition at different times. Neuro computing, 212:36–47, 2016.
10. Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1037–1045, 2015.
11. Dan Xu, Wan li Ouyang, Elisa Ricci, Xiao gang Wang, and Nicu Sebe. Learning cross-modal deep representations for robust pedestrian detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5363–5371, 2017.
12. Alejandro González, Zhijie Fang, Yainuvis Socarras, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M López. Pedestrian detection at day/night time with visible and fir cameras: A comparison. Sensors, 16(6):820, 2016.
13. Kihong Park, Seungryong Kim, and Kwanghoon Sohn. Unfied multi-spectral pedestrian detection based on probabilistic fusion networks. Pattern Recognition, 80:143–155, 2018.
14. Neubauer A, Yochelis S, Paltiel Y. Simple Multi Spectral Detection Using Infrared Nanocrystal Detector[J]. IEEE Sensors Journal, 2019, 19(10): 3668-3672.
15. Zhang, Lu, et al. "Weakly aligned cross-modal learning for multispectral pedestrian detection." Proceedings of the IEEE International Conference on Computer Vision. 2019: 5127-5137.
16. Lu Zhang, Zhi yong Liu, Shi feng Zhang, Xu Yang, Hong Qiao, Kai zhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. Information Fusion, 50:20–29, 2019.
17. Dayan Guan, Yan peng Cao, Jiang xin Yang, Yan long Cao, and ChristelLoic Tisse. Exploiting fusion architectures for multispectral pedestrian detection and segmentation. Applied Optics, 57(18):D108–D116, 2018.
18. Cheng yang Li, Dan Song, Ruo feng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. Pattern Recognition, 85:161–171, 2019.
19. Jing jing Liu, Shao ting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. In British Machine Vision Conference (BMVC),arXiv:1611.02644, 2016.
20. Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 243–250, 2017.
21. Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. In CVPR, 2018: 7132-7141.
22. REDMON, Joseph; FARHADI, Ali. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.
23. Li C, Song D, Tong R, et al. Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation.[J]. arXiv: Computer Vision and Pattern Recognition,arXiv:1808.04818, 2018.
24. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. computer vision and pattern recognition, 2016: 770-778.
25. S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In CVPR, pages 1037–1045, 2015.

26. Rezatofighi H, Tsoi N, Gwak J, et al. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression[C]. computer vision and pattern recognition, 2019: 658-666.
27. P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. TPAMI, 36(8):1532–1545, 2014.