WaveMamba: Wavelet-Driven Mamba Fusion for RGB-Infrared Object Detection

Supplementary Material

In this supplementary file, we introduce the details of the datasets and evaluation metrics in Sec. A, and the implementation in Sec. B. We also provide experimental results on the dataset VEDAI and KAIST in Sec. C. Additionally, we provide more visualization results and ablation studies in Sec. D and Sec. E, respectively.

A. Datasets and Evaluation Metrics

Experimental Datasets. We evaluate our methods on four common-used visible-infrared object detection benchmark datasets : M^3FD Dataset [13], DroneVehicle Dataset [20], LLVIP Dataset [9], FLIR Dataset [6] and two additional datasets: VEDAI Dataset [17] and KAIST Dataset [8].

M^3FD dataset is a benchmark dataset for multi-class RGB-IR detection which collects $4,200$ pairs of aligned images from various scenes. These pictures captured under low-light conditions or in adverse weather, pose a significant challenge to the detection performance of the model. This dataset encompasses six categories of objects that frequently appear in autonomous driving or road surveillance scenarios. Since it does not provide an official criterion for dividing the training set and validation set, we adopt the division standard from [11] as $3,360$ pairs for training and 840 pairs for testing.

DroneVehicle dataset is a large-scale RGB-IR vehicle detection dataset, consisting of $28,439$ pairs of images and $953,087$ annotations for five categories: *car*, *truck*, *bus*, *van*, and *freight-car*. The images are collected by drones under varying lighting conditions, angles, and altitudes. The rich perspective variations combined with a large number of dense annotations make it difficult for models to achieve high detection performance on this dataset. Following the official method for dataset partitioning, we use $17,990$ image pairs for training, $1,469$ pairs for validation, and $8,980$ pairs for testing. We report the testing part's results.

LLVIP dataset is an aligned low-light RGB-IR dataset which is specially collected for pedestrian detection with $15,488$ image pairs. According to the official standard, we use $12,025$ image pairs for training and $3,463$ pairs for testing.

FLIR dataset is a relatively difficult RGB-IR detection dataset with five categories: *people*, *car*, *bike*, *dog* and *other cars*. Due to the low-quality annotations and a large number of unaligned image pairs in the original dataset, following by [27], we adopt the FLIR-Aligned dataset which includes $4,129$ pairs of images for training and $1,013$ pairs for testing. We remove the '*dog*' category from the dataset due to its few number of instances.

VEDAI dataset is a remote sensing detection dataset consisting of $1,210$ pairs of RGB-IR images captured from a drone at high altitudes with nine different types of objects such as *car*, *truck* and *pickup*. Since the objects are predominantly small targets, it presents a significant challenge to the detection performance of the model. Since the dataset does not have an official split, we follow the common training and testing set partitioning methodology with $1,089$ pairs for training and 121 pairs for testing.

KAIST dataset is a public low-light multi-spectral pedestrian detection dataset. Due to problems in the original dataset annotations, we utilize the improved training [30] and testing annotations [12] that are widely adopted by researchers. Following the most commonly used data partitioning method provided by [30], we use $8,963$ image pairs for training and $2,252$ pairs for testing.

Evaluation metrics. Since our task is object detection, we choose the most widely used metrics mAP_{50} and mAP to evaluate the performance of models on six datasets. The mAP_{50} metric represents the mean AP under IoU 0.50 and the mAP metric represents the mean AP under IoU ranges from 0.50 to 0.95 with a stride of 0.05 [32]. For the multi-class datasets M^3FD and DroneVehicle, we also provide the AP_{50} results for each category. Due to the high difficulty of FLIR-Aligned, we additionally report the results of precision, recall, and F1 score. All the evaluation metrics indicate better model detection performance when their values are higher. We also present the average inference time of our method, evaluated on an A800 GPU over 15 runs using input image pairs of size 640×640. Additionally, we provide the parameter specifications of our model.

B. Implementation Details

All experiments on our six datasets are conducted on a single A800 GPU, with a batch size of 16 during training and 32 during testing. The input image pairs' size for both testing and training are 640×640 and the training epoch is set to 250 for all four datasets with an initial learning rate of 0.01. We utilize the SGD optimizer with a momentum of 0.937 and a weight decay of 0.0005. The loss function, other hyper-parameters, and data augmentation parameters all adopt the default settings of the original YOLOv8 [23].

Methods	Backbone	mAP_{50}	mAP	Parameters	Inference time (ms)
Ours	ResNet50	87.9	59.4	193.2M	53.2
(TNNLS'23) LRAF-Net [7]	YOLOv5	85.9	59.1	-	-
(PR'24) ICAFusion [19]	YOLOv5	84.8	56.6	164.3M	50.2
(GRSL'24) SSE+FFT [26]	YOLOv5	86.5	56.9	-	-
(ACCV'24) ESM-YOLO [31]	YOLOv5	82.4	-	80.2M	-
(JSTAEORS'25) DCCCNet [22]	YOLOv5	82.0	49.9	-	-
(GRSL'24) CrossYOLO [15]	YOLOv7	79.8	-	-	-
Ours	YOLOv5	88.2	59.6	45.6M	44.1
YOLOv8l-IR	YOLOv8	73.6	52.2	43.7M	22.0
YOLOv8l-RGB	YOLOv8	62.9	46.5	43.7M	22.0
(JSTAEORS'25) MDA [33]	DETR	82.2	-	72.4M	-
(ASC'24) DSM-AVD [29]	YOLOv8	79.3	50.3	-	-
Ours	YOLOv8	88.4	59.8	69.1M	40.0

Table A. Comparison results with SOTA methods on VEDAI dataset. The best, second and third results are highlighted in red, green and blue, respectively.

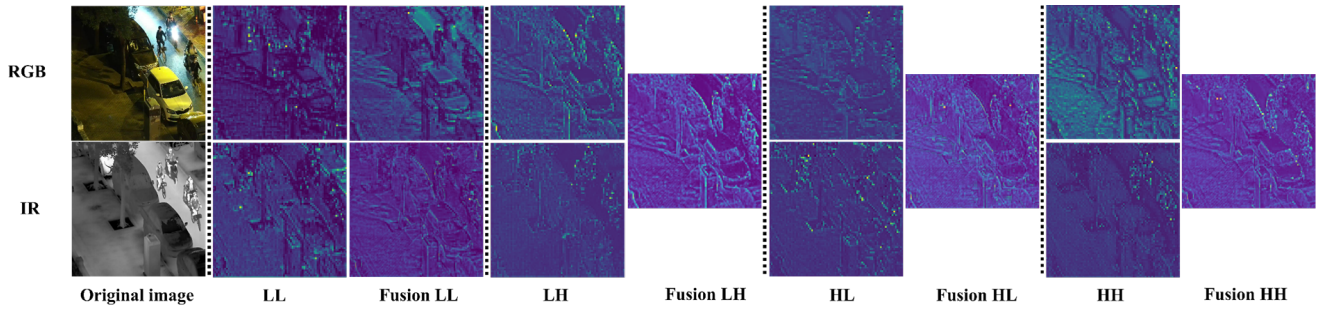

Figure A. The illustration shows how the original features are filtered in two directions (horizontal and vertical) using DWT, resulting in four sub-bands: LL (low-low), LH (low-high), HL (high-low), and HH (high-high). "Fusion" refers to the results obtained after combining these sub-bands through our method.

Methods	Backbone	mAP_{50}	mAP
(CVPRW'19) MMTOD [4]	ResNet50	70.7	31.3
(WACV'21) GAFF [28]	ResNet50	67.1	24.4
(TCSVT'22) CMDet [20]	ResNet50	68.4	28.3
Ours	ResNet50	75.0	34.2
(2021) CFT [16]	YOLOv5	71.2	29.3
(RS'22) RISNet [24]	YOLOv5	72.7	33.1
(PR'24) ICAFusion [19]	YOLOv5	60.3	-
(CVPRW'24) DaFF [1]	YOLOv5	61.9	-
Ours	YOLOv5	75.4	34.4
YOLOv8l-IR [23]	YOLOv8	56.8	22.4
YOLOv8l-RGB [23]	YOLOv8	55.3	21.6
(Sensors'23) Dual-YOLO [2]	YOLOv7	73.2	-
(Sensors'24) IV-YOLO [21]	YOLOv8	75.4	-
Ours	YOLOv8	75.8	34.8

Table B. Comparison results with SOTA methods on KAIST dataset. The best results are highlighted in red. The second and third best results are highlighted in green and blue, respectively.

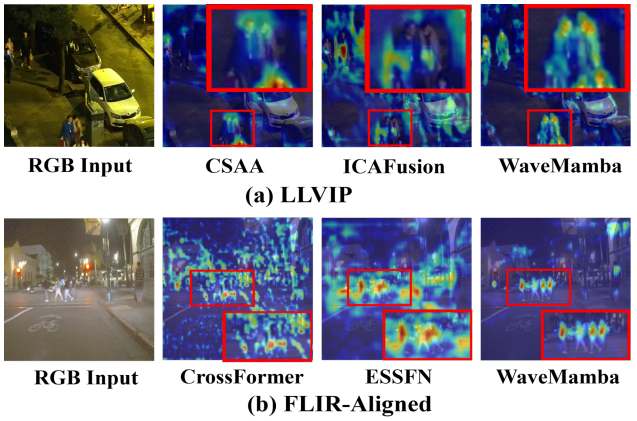

Figure B. Heatmap visualization of several cross-modality object detection methods on LLVIP and FLIR-Aligned.

C. More experiments on VEDAI and KAIST

VEIDA Dataset. The results on VEDAI are summarized on Table A. Our method achieves top-three rankings on both

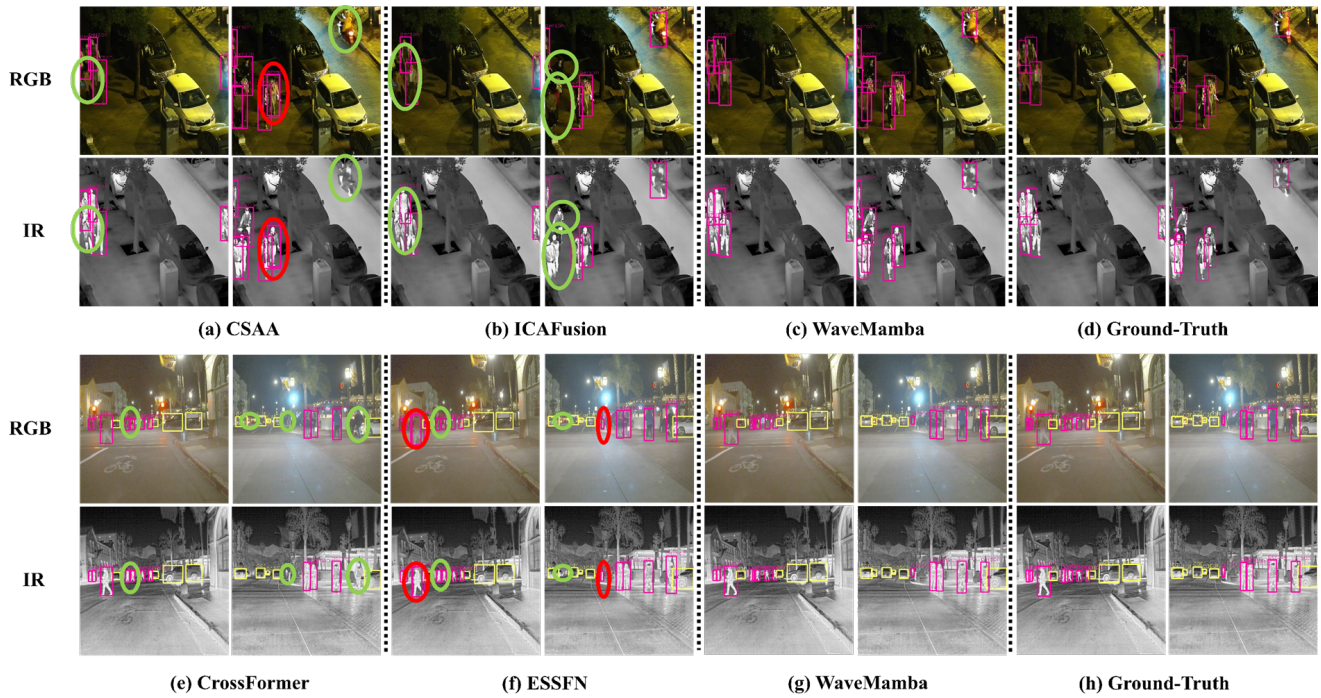

Figure C. Detection results' visualization of several cross-modality object detection methods on LLVIP and FLIR-Aligned. Wherein, (a)-(d) present the results of LLVIP dataset, and (e)-(h) present the results of FLIR-Aligned dataset. The targets encircled by red ellipses are false positives, while those encircled by green ellipses are missed detections. Please zoom in for more details.

Methods	mAP_{50}	mAP	Parameters
$\{P_1,P_2,P_3\}$	91.1	62.8	23.1M
$\{P_2,P_3,P_4\}$	91.8	63.5	58.3M
$\{P_3,P_4,P_5\}$	91.5	62.9	84.6M
$\{P_2,P_4,P_5\}$	91.3	62.5	78.3M
$\{P_2,P_3,P_5\}$ (ours)	92.1	64.4	69.1M

Table C. Effects of positions of WaveMamba Fusion Blocks on M^3FD dataset.

Methods	mAP_{50}	mAP	Parameters
(One Blocks)	91.3	63.8	59.3M
(Two Blocks)	91.6	64.0	61.6M
(Three Blocks) (ours)	92.1	64.4	69.1M

Table D. Effects of different number of WaveMamba Fusion Blocks on M^3FD dataset.

mAP_{50} and mAP, surpassing the fourth-place method by 1.9% and 0.7% and achieving 88.4% and 59.8%, respectively. Moreover, our method, based on the YOLOv5 and YOLOv8 backbones, has the smallest number of parameters, requiring only 44.1 ms and 40.0 ms to process a pair of images, respectively.

KAIST Dataset. Table B shows the results of our method

Methods	mAP_{50}	mAP	Parameters
(MEYR)	91.7	64.0	69.1M
(SYM3)	91.8	64.1	69.1M
(COIF3)	91.7	64.1	69.1M
(DB3)	91.8	64.2	69.1M
(HAAR) (ours)	92.1	64.4	69.1M

Table E. Effects of different wavelet bases in WaveMamba Fusion Blocks on M^3FD dataset.

with other SOTA methods on KAIST. Our method achieves the top three results using three different backbones on both mAP_{50} and mAP, surpassing the fourth method by 0.4% and 1.7%, respectively. Notably, although the feature extraction capability of the YOLOv5 backbone is inferior to that of YOLOv7, our method using YOLOv5 still outperforms Dual-YOLO, which utilize YOLOv7. The results demonstrate the superior performance of our WMFBs.

D. Visualization

D.1. Frequency-domain graph

We visualize the wavelet transform outputs of RGB and IR features derived from the second stage of YOLOv8 backbones trained on each respective modality from a pair of

RGB and IR images, as well as the high- and low-frequency components after fusion. As shown in Fig. A, the low-frequency components of IR more effectively convey shape information compared to RGB, while the high-frequency components of RGB more distinctly emphasize local object contours and details than IR. After fusion, the low-frequency components of both RGB and IR become clearer, and the local object features in the high-frequency components after fusion are significantly enhanced.

D.2. Heatmaps

Based on Grad-CAM [18], we visualize the heatmaps of our model's first inverse wavelet transform layer (for the enhanced high- and low-frequency features obtained at the third fusion block) and compare them with other state-of-the-art methods [3, 10, 19, 25] on LLVIP and FLIR-Aligned. As shown in Fig. B, our model focuses more intently on detecting targets without being excessively distracted by background noise by utilizing wavelet transform and feature fusion to prioritize the targets. In contrast, other methods either exhibit excessive attention dispersion in background areas or cover the detection targets with a wide range of attention, which easily leads to missed detections of targets within the region.

D.3. Detection Results

We visualize the detection results and compare them with several SOTA methods [3, 10, 19, 25]. As presented in Fig. C, under low-light or heavily occluded conditions, our method has reduced the number of missed and false detected targets compared to other methods. It successfully detects more difficult targets and achieves the best detection performance.

E. Ablation Study

E.1. Details of the Improved Head

To demonstrate the effectiveness of our improved YOLOv8 head, we also combine our fusion modules with the original YOLOv8 head for comparison. To seamlessly integrate our method into the original YOLOv8 head, we sum the fused low-frequency components obtained after each fusion layer, concatenate them with the fused high-frequency components for inverse wavelet transformation, and then feed the result into the original YOLOv8 head.

E.2. More Ablation Experiments

Effects of WMFBs' positions. Like previous works [5, 10], we also employ three feature fusion blocks. Table C shows the effect of different position combinations of WMFBs on performance. P_i represents that the WMFB is placed at i^{th} stage. Based on the experimental results from "{ P_1, P_2, P_3}" and [14], the first layer features are not suitable for fusion. Without P_1, there are four remaining possible combinations of layers. When comparing "{ P_2, P_3, P_5}" with others, the importance of using features from the second, third, and fifth layers is verified by the significant accuracy boost.

Effects of the number of WMFBs. Since YOLOv8 head inherently contains three detection modules, we conduct an ablation study on the effects of the number of WMFBS by changing one or two WMFBs to simple AVG method and the results are shown in Table D. Reducing the number of WMFBs from three to one and two results in a decrease of 0.8% and 0.5% in mAP_{50} and a decrease of 0.6% and 0.4% in mAP, respectively.

Effects of different Wavelet Bases. We show the influence of different wavelet bases in Table E. The performance difference in mAP_{50} and mAP is less than 0.4% and the Haar wavelet base achieves the best performance. This result indicates that our method is not sensitive to the selection of wavelet bases.

References

[1] Afnan Althoupety, Li-Yun Wang, Wu-Chi Feng, and Banafsheh Rekabdar. Daff: Dual attentive feature fusion for multispectral pedestrian detection. In *CVPRW*, pages 2997–3006, 2024. 2

[2] Chun Bao, Jie Cao, Qun Hao, Yang Cheng, Yaqian Ning, and Tianhua Zhao. Dual-yolo architecture from infrared and visible images for object detection. *Sensors*, 2023. 2

[3] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *CVPRW*, 2023. 4

[4] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M Sharma, and Vineeth N Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *CVPRW*, pages 0–0, 2019. 2

[5] Qingyun Fang, Dapeng Han, and Zhaokui Wang. Cross-modality fusion transformer for multispectral object detection. *arXiv:2111.00273*, 2021. 4

[6] TELEDYNE FLIR. Free teledyne flir thermal dataset for algorithm training. Online, 2024. 1

[7] Haolong Fu, Shixun Wang, Puhong Duan, Changyan Xiao, Renwei Dian, Shutao Li, and Zhiyong Li. Lraf-net: Long-range attention fusion network for visible–infrared object detection. *TNNLS*, 35(10):13232–13245, 2024. 2

[8] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, pages 1037–1045, 2015. 1

[9] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *ICCVW*, 2021. 1

[10] Seungik Lee, Jaehyeong Park, and Jinsun Park. Crossformer: Cross-guided attention for multi-modal object detection. *Pattern Recognit. Lett.*, 2024. 4

[11] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion

for rgb-thermal perception tasks. *IEEE Robot. Autom. Lett.*, 2023. 1

[12] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016. 1

[13] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, 2022. 1

[14] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion*, 36:191–207, 2017. 4

[15] Jinyan Nie, He Sun, Xu Sun, Li Ni, and Lianru Gao. Cross-modal feature fusion and interaction strategy for cnn-transformer-based object detection in visual and infrared remote sensing imagery. *IEEE Geoscience and Remote. Sens. Letters*, 21:1–5, 2024. 2

[16] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021. 2

[17] Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *JVCIR*, 34:187–203, 2016. 1

[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 4

[19] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *PR*, 2024. 2, 4

[20] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *TCSVT*, 2022. 1, 2

[21] Dan Tian, Xin Yan, Dong Zhou, Chen Wang, and Wenshuai Zhang. Iv-yolo: A lightweight dual-branch object detection network. *Sensors*, 2024. 2

[22] Shu Tian, Li Wang, Lin Cao, Lihong Kang, Xian Sun, Jing Tian, Xiangwei Xing, Bo Shen, Chunzhuo Fan, Kangning Du, Chong Fu, and Ye Zhang. A dynamic cascade cross-modal coassisted network for aav image object detection. *J-STARS*, 18:2749–2765, 2025. 2

[23] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *ADICS*, 2024. 1, 2

[24] Qingwang Wang, Yongke Chi, Tao Shen, Jian Song, Zifeng Zhang, and Yan Zhu. Improving rgb-infrared object detection by reducing cross-modality redundancy. *Remote. Sens.*, 14(9):2020, 2022. 2

[25] Fengxiang Xu, Tingfa Xu, Lang Hong, Peiran Peng, Jiaxin Guo, and Jianan Li. Enhanced spectral–spatial fusion network for multispectral object detection in ground-aerial images. *IEEE Geoscience and Remote. Sens. Letters*, 2024. 4

[26] Fengxiang Xu, Tingfa Xu, Lang Hong, Peiran Peng, Jiaxin Guo, and Jianan Li. Enhanced spectral-spatial fusion network for multispectral object detection in ground-aerial images. *IEEE Geoscience and Remote. Sens. Letters*, 2024. 2

[27] Heng Zhang, Élisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *ICIP*, 2020. 1

[28] Heng Zhang, Elisa Fromont, Sebastien Lefevre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *WACV*, pages 72–80, 2021. 2

[29] Jie Zhang, Tian qing Chang, Li yang Zhao, Jin dun Ma, Bin Han, and Lei Zhang. Efficient cross-modality feature interaction for multispectral armored vehicle detection. *Appl. Soft Comput*, 163:111971, 2024. 2

[30] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *ICCV*, pages 5127–5137, 2019. 1

[31] Qianqian Zhang, Linwei Qiu, Li Zhou, and Junshe An. Esm-yolo: Enhanced small target detection based on visible and infrared multi-modal fusion. In *ACCV*, pages 1454–1469, 2024. 2

[32] Tianyi Zhao, Maoxun Yuan, Feng Jiang, Nan Wang, and Xingxing Wei. Removal then selection: A coarse-to-fine fusion perspective for rgb-infrared object detection. *arXiv:2401.10731*, 2024. 1

[33] Jiahe Zhu, Huan Zhang, Simin Li, Shengjin Wang, and Hongbing Ma. Cross teaching-enhanced multispectral remote sensing object detection with transformer. *J-STARS*, 18:2401–2413, 2025. 2