A Lightweight RGB–Thermal Fusion Method for UAV-Based Wildfire Detection

Keran Feng[1], **KinTak U**[2]

[1]Faculty of Innovation Engineering, Macau University of Science and Technology, Taipa, Macau
[2]Faculty of Innovation Engineering, Macau University of Science and Technology, Taipa, Macau
E-MAIL: 2240024583@student.must.edu.mo, ktu@must.edu.mo

Abstract:

Object detection on Unmanned Aerial Vehicles (UAVs) is often hampered by limited onboard memory and the sensitivity of sensors to environmental noise. While RGB-Thermal (RGBT) fusion improves reliability, existing high-accuracy models are typically too heavyweight for real-time edge deployment. In this study, we proposed a comprehensive RGBT wildfire detection solution specifically optimized for power-constrained UAVs. Based on the YOLOv8 framework, our approach improves both the model structure and the training process. First, we design the HS-GFM module to capture fine details and align features with very few parameters. Second, we use a specialized data augmentation to solve the problems of IR noise and motion blur. Experimental results on the RGBT-3M dataset show that our solution achieves 0.959 mAP$_{50}$. Compared to the original baseline in this dataset, we use less than one-third of the parameters (3.09M vs 11.83M) with only a 2.4% drop in mAP50–95. Consuming just 4.56 W at 480x480 resolution on an NVIDIA Jetson Nano B01 (4GB), this approach provides a robust and efficient pipeline for real-time aerial surveillance.

Keywords:

edge computing, forest fire detection, lightweight neural networks, RGBT fusion, unmanned aerial vehicles (UAVs)

1. Introduction

As a result of global climate change and extreme weather, both the frequency and severity of forest fires have increased[1]. Traditional ground monitoring methods often fail to provide timely alerts in vast and inaccessible forest areas. Therefore, the inherent flexibility and multi-scale sensing capabilities of unmanned aerial vehicles (UAVs) position them as a practical and reliable approach for forest fire monitoring.

However, most platforms rely on visible light (RGB) cameras, which can provide rich texture details but are prone to being affected by unfavorable conditions such as thick smoke, vegetation obstruction, and low light environments. Thermal infrared (TIR) imaging provides thermal information, which enables accurate fire localization in scenes where visibility is limited. Therefore, integrating the RGB and TIR modes - commonly referred to as RGBT detection - has become a reliable method for forest fire monitoring applicable to various weather conditions.

Multimodal fusion strategies in RGBT detection can be divided into three categories: data level, feature level, and decision level fusion [2]. The current research is increasingly inclined to adopt the feature-level fusion method, as previous studies have shown that this approach can effectively enhance the interaction between modalities during the feature extraction stage and uncover complementary information [3, 4]. More and more advanced methods adopt architectures based on attention mechanisms and similar to the Transformer, which demonstrate outstanding capabilities in integrating and fusing different types of RGB-T features [5, 6]. Although these methods can achieve powerful feature interaction, their high computational requirements often limit their real-time execution on edge devices. Therefore, for time-sensitive aerial forest fire monitoring, there is still an urgent need for a lightweight but efficient fusion model to ensure reliable detection while not sacrificing the flexibility of the unmanned aerial vehicle system.

To bridge this gap, our research introduces an efficient wildfire detection framework designed for the resource constraints of unmanned aerial vehicles. This framework is built based on the YOLOv8n architecture[7], with a focus on balancing the synergy effect between thermal imaging and visual information and the execution speed.

1) High-Speed Gated Fusion Module: A lightweight gated fusion unit designed with an efficient feature mixing strategy, combining channel-level gating and spatial convolution to reduce modality-specific noise while integrating complementary features with a small number of additional parameters.

2) Asymmetric Data Augmentation: A targeted training strategy specifically injects the unique interference of the sensors (such as thermal noise and blurring) into the infrared channel, thereby eliminating the instability caused by thermal effects without sacrificing the fine texture of the RGB data.

3) Edge-Deployable RGBT Framework: A lightweight RGBT detection framework with only 3.09 million parameters (approximately 30% of the baseline version), has been verified for its real-time feasibility through a comprehensive assessment of inference speed, power consumption and memory usage on NVIDIA Jetson Nano under different input resolutions.

2. PROPOSED METHOD

Motivated by the need for real-time UAV deployment, we modify the YOLOv8n architecture into a dual-stream framework optimized for RGBT inputs. The subsequent subsections present the pipeline, focusing particularly on the HS-GFM fusion approach and our targeted training strategy.

2.1. Overall Architecture

The framework is built on the YOLOv8 backbone and configured as a dual-stream network, optimized for RGBT fire detection with enhanced multi-modal feature integration. As illustrated in Fig. 1, the model is designed with a symmetrical four-stage pipeline: an input preprocessing stage, a dual-stream backbone, a feature fusion neck, and a multi-scale detection head.

Details of the processing pipeline are described below: Dual-modality data (RGB and IR) is partitioned by **Multiin** and processed through parallel backbones consisting of **Conv** and **C2f** blocks. This produces multi-scale features (P_3–P_5), with an **SPPF** module at P_5 to expand the receptive field. These synchronized features are then adaptively fused via **HSGFM** units. The **Neck** employs an **FPN–PAN** structure, where features are refined by **C2f** blocks through top-down and bottom-up paths to integrate multi-scale localization and semantic cues. Finally, the fused maps are passed to the **Detect** layer for prediction.

2.2. High-Speed Gated Fusion Module (HS-GFM)

The RGB and TIR sensors can capture different features; however, simple stitching is often insufficient for achieving effective integration[8, 9]. Such naive fusion strategies fail to explicitly model cross-modal correlations and are sensitive to modality-specific noise. Therefore, a dedicated fusion module

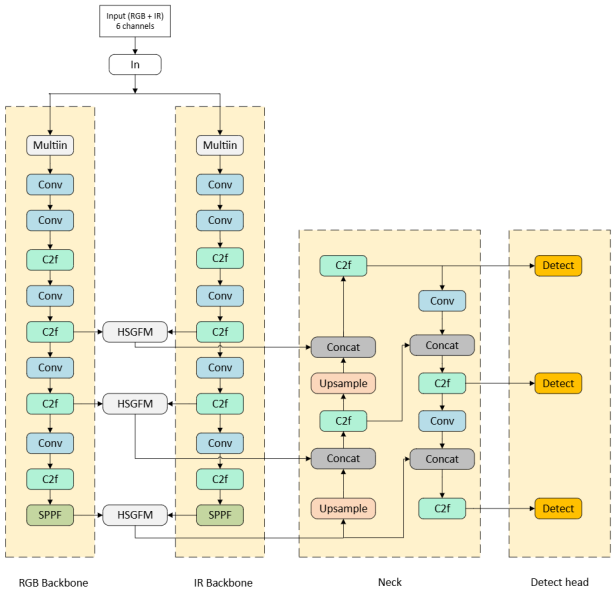

FIGURE 1. The proposed model architecture.

is needed to adaptively select useful features from the two sensors, ensuring that the model focuses on key fire clues while ignoring background noise.

To adaptively regulate the fusion of RGB and TIR features, we introduce a gating mechanism that generates pixel-level or channel-level weights ranging from 0 to 1[4]. These weights, produced via a Sigmoid function, allow the network to express the relative confidence of each modality in a smooth and bounded manner, emphasizing reliable features while suppressing noise. This allows the network to adaptively emphasize more reliable sources at each spatial location. The calculation process can be described as follows:

Following the concept of the Squeeze-and-Excitation (SE) mechanism [10], we first obtain global descriptors from both RGB and TIR feature maps using average pooling. These descriptors are then combined into a single vector, $\mathbf{v} = [\mathrm{GAP}(\mathbf{F}^{rgb}), \mathrm{GAP}(\mathbf{F}^{ir})]$, so that the network can capture the overall contextual information of both modalities. A channel gate is then generated by a lightweight two-layer MLP to maintain efficient inference performance:

$$\mathbf{g} = \sigma(\mathrm{MLP}(\mathbf{v})), \tag{1}$$

and the global fusion is computed as:

$$\mathbf{F}_{global} = \mathbf{g} \odot \mathbf{F}^{rgb} + (1 - \mathbf{g}) \odot \mathbf{F}^{ir}. \tag{2}$$

In parallel, the spatial path captures local features by concatenating the two modalities $\mathbf{U} = [\mathbf{F}^{rgb}, \mathbf{F}^{ir}]$ and applying

a shallow two-layer 3×3 convolutional block to generate the spatial mask $\mathbf{s}$:

$$\mathbf{s} = \sigma(\mathcal{F}conv(\mathbf{U})), \tag{3}$$

where $\mathcal{F}conv$ denotes a sequence of two 3×3 convolutions interspersed with a ReLU activation, designed to extract local spatial saliency while compressing the channel dimension. The resulting map is used for pixel-wise fusion:

$$\mathbf{F}_{spatial} = \mathbf{s} \odot \mathbf{F}^{rgb} + (\mathbf{1} - \mathbf{s}) \odot \mathbf{F}^{ir}. \tag{4}$$

where $\odot$ denotes the element-wise product and $\mathbf{1}$ is an all-ones tensor of the same dimensions as $\mathbf{s}$.

Finally, two learnable scalar parameters α and β balance the contributions of global and spatial paths. The fused output is refined by a 3×3 convolution and batch normalization (BN):

$$\mathbf{F}_{out} = \text{BN}(\text{Conv}_{3\times3}(\alpha\mathbf{F}_{global} + \beta\mathbf{F}_{spatial})). \tag{5}$$

2.3. Specialized Data Augmentation for RGBT

In multi-modal vision tasks, different sensors capture complementary but distinct information, and applying identical augmentations to all modalities can be suboptimal [11, 12]. Inspired by this, we apply separate augmentations for RGB and IR channels: RGB features undergo standard photometric and geometric transformations, while IR features are perturbed with thermal noise and blur to simulate sensor-specific disturbances. These modality-specific augmentations are particularly important for UAV-based wildfire detection, as aerial imagery often faces motion blur, low visibility, and sensor-specific noise.

Based on this design, our data augmentation strategy introduces essentially zero additional computational cost, yet improves the model's mAP_{50-95} and significantly enhances robustness under challenging conditions. Detailed results and ablation studies are provided in the Experimental Section.

3 EXPERIMENTS

3.1 Experimental Settings

The experimental procedure consists of two main stages: training and deployment.

We conducted the model training in an environment featuring Ubuntu 20.04.5 LTS, an NVIDIA GeForce RTX 4090D (24GB) GPU, and Python 3.8.10 (CUDA 11.8). Throughout the training of all networks, we kept the settings consistent by fixing the random seed at 42. The training process lasted 200 epochs in total, which consisted of 150 standard epochs followed by 50 fine-tuning epochs. Specific training parameters, including the AdamW optimizer and batch size, are detailed in Table 1.

TABLE 1. Training environment and hyperparameter settings.

Environment Item	Parameter Settings
CPU	18 vCPU AMD EPYC 9754 128-Core Processor
GPU (Training)	NVIDIA GeForce RTX 4090D (24GB)
Edge Device (Inference)	NVIDIA Jetson Nano B01 (4GB)
Operating System	Ubuntu 20.04.5 LTS / JetPack 4.6.3
Framework	Python 3.8.10, **Ultralytics 8.3.22**, PyTorch 2.0.0
CUDA Version	11.8 (Training) / 10.2 (Inference)
Optimizer	AdamW
Momentum	0.937
Weight Decay	0.0005
Training Epochs	200 (150 + 50 fine-tuning)
Batch Size	16
Input Resolution	640×640

We tested the model on an NVIDIA Jetson Nano B01 to verify its performance in real-world deployment scenarios. The deployment environment, as specified in Table 2, utilizes TensorRT 8.2.1 and cuDNN 8.2.1 to accelerate inference.

TABLE 2. Inference environment and hardware specifications of Jetson Nano.

Environment Item	Parameter Settings
Model	NVIDIA Jetson Nano B01 (4GB)
GPU	128-core NVIDIA Maxwell (Compute 5.3)
CPU	Quad-core ARM Cortex-A57 @ 1.43 GHz
Operating System	Ubuntu 18.04 LTS (JetPack 4.6.3)
Python Version	3.6.9
CUDA Version	10.2.300
cuDNN Version	8.2.1
TensorRT Version	8.2.1
OpenCV Version	4.1.1
Memory	4GB 64-bit LPDDR4

3.2 Evaluation Criteria

Standard evaluation metrics, including Precision (P), Recall (R), mean Average Precision (mAP), and Frames Per Second (FPS), are adopted to evaluate the detection accuracy and inference efficiency.

3.3 Dataset Description

We use RGBT-3M[13] as our benchmark dataset. RGBT-3M is a multi-modal unmanned aerial vehicle dataset used for forest fire detection. This dataset contains 17,862 pairs of synchronized RGB and thermal infrared images, which are divided into the training set and validation set in a 7:3 ratio. It adopts the M-RIFT framework and aligns the cross-modal image pairs

through a two-stage coarse-to-fine affine transformation process. The RGBT-3M dataset also introduces CP-YOLOv11-MF, a benchmark dual-backbone model utilizing CPCA and PPAS modules for cross-modal fusion. We consider this model as the baseline to evaluate our proposed HS-GFM.

For the training set and validation set, we adopted the same label processing method as in RGBT-3M: Since infrared images lack distinctive smoke features, the "smoke" category was removed[13]. The distribution of images and label counts are summarized in Table 3.

TABLE 3. RGBT-3M dataset distribution without smoke labels.

Dataset	Number of Images	Total Labels	Fire	Person
Training Set	7,854	12,062	7,914	4,148
Validation Set	3,366	5,141	3,401	1,740
Total	**11,220**	**17,203**	**11,315**	**5,888**

3.4 Comparative Experiment

In this subsection, we compare our model with single-modality models and the CP-YOLOv11-MF framework proposed in RGBT-3M. Based on the previous subsection, we use only the 'flame' and 'person' labels from the RGBT-3M dataset for model training.

The results are summarized in Table 4. In terms of precision and recall rate, there is a significant performance gap between the bimodal model and the unimodal model. Although CP-YOLOv11-MF performs best in terms of Recall and mAP, it requires nearly four times the parameters of our model. Our model achieves a mAP_{50-95} only 2.4 points lower while substantially reducing model size, offering a more practical balance of accuracy and efficiency for UAV deployment.

TABLE 4. Comparison of detection performance and model complexity on the RGBT-3M dataset.

Model	P (%)	R (%)	mAP50 (%)	mAP50-95 (%)	Params (M)
YOLOv8n-RGB (Single)	91.2	87.3	92.7	53.9	3.01
YOLOv8n-IR (Single))	91.0	87.8	93.1	57.3	3.01
CP-YOLOv11-MF	92.5	**93.5**	**96.3**	**62.9**	11.83
HS-GFM-DS (Ours)	**93.0**	92.3	95.9	60.5	**3.09**

**DS denotes the Data Strengthening strategy.*

3.5 Ablation Study

YOLOv8n is used as the baseline model for unimodal detection on visible and infrared images. Based on this backbone, a dual-stream fusion architecture, YOLOv8-dual, is constructed to integrate RGB and thermal features. By integrating the Gate Fusion Module (GFM) with the High-Speed (HS) into YOLOv8-dual, an effective cross-modal feature interaction mechanism was established. Through the introduction of data augmentation strategies, the HS-GFM-DS model was ultimately obtained.

TABLE 5. Ablation study of the proposed components on the RGBT-3M.

Model Configuration	P (%)	R (%)	mAP50 (%)	mAP50-95 (%)	Params (M)
YOLOv8n-RGB (Single)	91.2	87.3	92.7	53.9	3.01
YOLOv8n-IR (Single)	91.0	87.8	93.1	57.3	3.01
YOLOv8-dual	91.9	89.9	94.6	58.1	2.38
YOLOv8-dual + HS-GFM	91.5	91.1	95.4	59.6	3.09
HS-GFM-DS (Ours)	**93.0**	**92.3**	**95.9**	**60.5**	**3.09**

Different bimodal fusion strategies are compared in Table 5 by gradually enhancing the model architecture. Relative to single-modality baselines, the dual-stream framework shows clear performance improvements, suggesting effective complementarity between RGB and thermal information. However, simple feature concatenation provides only limited performance improvements.

The proposed HS-GFM jointly incorporates gated cross-modal interaction modeling and high-speed feature mixing, enabling more effective RGB–TIR information exchange. With the proposed data augmentation strategy, the HS-GFM-DS model achieves 93.0% precision, 92.3% recall, 95.9% mAP_{50}, and 60.5% mAP_{50-95}. These results indicate that each component contributes positively to the overall performance.

3.6 Robustness Analysis

Considering the instability of thermal infrared imaging during UAV flight, we further investigate the robustness of the proposed method under severe infrared degradation. In this evaluation, two typical distortion types observed in UAV-mounted infrared sensors are considered: infrared blurring and infrared noise.

The original test samples are used to build a more challenging test set for robustness evaluation. Gaussian noise with a standard deviation of $\sigma \in [20, 30]$ is introduced for the noise test, while Gaussian blur with kernel sizes of $k \times k$ ($k \in \{9, 11\}$) is applied for the blur test. Fig. 2 shows representative examples of the resulting degraded images.

Table 6 presents the robustness performance under infrared blur and noise. Without data augmentation (DS), the HS-GFM model suffers a significant performance degradation when infrared images are corrupted. Under severe noise, nearly all evaluation metrics drop to zero.

After applying the DS strategy, the HS-GFM-DS model becomes considerably more stable. On the infrared-blurred dataset, the model performs comparably to the baseline, with only a minor drop in mAP_{50}. More importantly, in the severe

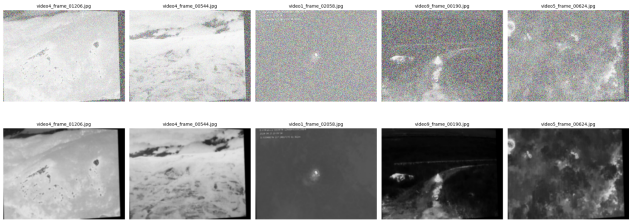

FIGURE 2. Visualization of infrared images with synthetic degradations. The top row shows samples corrupted by Gaussian noise, while the bottom row shows samples affected by Gaussian blur.

IR noise environment, the model still maintains an accuracy and recall rate of over 90%, and mAP_{50} remains at 94.3%. The model with data augmentation remains robust even under severe infrared degradation.

TABLE 6. Robustness analysis under IR Blur and IR Noise conditions.

Condition	Model	P (%)	R (%)	mAP_{50} (%)	mAP_{50-95} (%)
Baseline	HS-GFM	91.5	91.1	95.4	59.6
	HS-GFM-DS	**93.0**	**92.3**	**95.9**	**60.5**
IR Blur	HS-GFM	83.3	70.5	81.4	44.1
	HS-GFM-DS	**92.8**	**91.8**	**95.5**	**59.4**
IR Noise	HS-GFM	4.2	12.8	0.7	0.3
	HS-GFM-DS	**91.8**	**91.2**	**94.3**	**56.6**

**HS-GFM here refers to the YOLOv8-dual framework equipped with HS-GFM modules.*

3.7 Real-time Deployment

The model was deployed on a resource-constrained NVIDIA Jetson Nano platform to evaluate its real-time feasibility. In addition to detection accuracy metrics, inference speed, memory usage, and power consumption are evaluated to assess suitability for edge deployment.

The trained PyTorch model was converted to ONNX format and further optimized using TensorRT with FP16 precision to accelerate inference on the Jetson Nano platform. All deployment results were evaluated directly on the target device.

On resource-constrained edge devices, changing the input resolution leads to clear trade-offs between detection accuracy and inference speed. Therefore, we evaluated the proposed model for various input resolutions to analyze the trade-off between accuracy and efficiency in real-time deployment scenarios. Quantitative results are summarized in Table 7. Reducing the input resolution increases inference speed at the expense of detection accuracy. While the 640×640 setting yields the best fire detection performance, the 320×320 configuration reaches 21.66 FPS. This flexibility enables the HS-GFM-DS framework to be configured for different UAV deployment scenarios by adjusting the trade-off between accuracy and inference speed.

In addition to the inference speed, we also use the `tegrastats` tool to monitor the power consumption and memory usage under different input resolutions. Table 8 summarizes the energy efficiency of the model at different input resolutions. Lower resolutions consistently yield higher FPS/W, with the 320×320 configuration achieving the best efficiency of 5.40 FPS/W.

TABLE 7. Inference performance and detection accuracy across different input resolutions on NVIDIA Jetson Nano.

Resolution	Class	P	R	mAP_{50}	Latency (ms)	FPS
640 × 640	Fire	0.907	0.929	0.892	108.58	9.21
	Person	0.897	0.916	0.894		
	All	**0.904**	**0.925**	**0.893**		
480 × 480	Fire	0.916	0.872	0.807	68.31	14.64
	Person	0.894	0.883	0.812		
	All	**0.909**	**0.876**	**0.809**		
320 × 320	Fire	0.849	0.573	0.529	46.17	21.66
	Person	0.877	0.568	0.517		
	All	**0.858**	**0.571**	**0.523**		

Furthermore, owing to the lightweight model design, the runtime overhead remains stable across different input resolutions. This property is favorable for real-time deployment and long-duration UAV-based fire monitoring tasks.

TABLE 8. Deployment efficiency, memory footprint, and power consumption on NVIDIA Jetson Nano.

Resolution	Average FPS	Avg. Power (W)	Peak RAM (MB)	Efficiency (FPS/W)
640 × 640	9.21	5.11	3106	1.80
480 × 480	14.64	4.56	3103	3.21
320 × 320	21.66	4.01	3083	**5.40**

**Efficiency is defined as the number of frames processed per watt (FPS/W).*
**Power and memory data were collected via the tegrastats utility.*

Overall, the experimental results indicate that the proposed method can meet the real-time requirements of resource-constrained UAV platforms.

4 Conclusion and Future Work

In this study, we proposed a lightweight two-stream fusion framework named HS-GFM-DS for fire detection based on unmanned aerial vehicles on edge devices. This framework improves feature learning by introducing a high-speed (HS) and effectively fuses visible light and infrared information using a gated fusion module (GFM) to alleviate the inconsistency caused by multimodal data augmentation and sensor-specific noise. To mitigate the inconsistency caused by multimodal data augmentation and specific sensor noise, the data reinforcement (DS) strategy was applied, which enhances the model's robustness and generalization performance in complex environments.

Experimental results show that the proposed framework achieves a reasonable balance between detection accuracy and computational efficiency. Although its model size is only 3.09

million parameters, this framework still maintains stable performance on embedded hardware. Moreover, the robustness assessment in the presence of infrared blurring and noise indicates that the model remains reliable under unfavorable imaging conditions.

Deployment experiments conducted on NVIDIA Jetson Nano demonstrated that an input resolution of 480×480 provides a practical balance solution, with a frame rate of 14.64 FPS, $mAP_{50}$of 80.9%, and energy efficiency of 3.21 FPS/W. This configuration meets the real-time and low-power requirements of the UAV-based fire detection task.

Although encouraging results were observed in this study, there are still some limitations before applying the proposed system to large-scale industrial scenarios. These issues also point out potential directions for future research:

(i) Flight Validation: Due to hardware and venue constraints, this assessment focused on static edge devices; future efforts will validate the model under real UAV flight conditions, such as vibrations and rapid maneuvers.

(ii) Cost-Resolution Trade-off: While high-resolution thermal sensors are costly, the strong performance at 480×480 resolution motivates future research into super-resolution and lightweight optimization to enable accurate detection with lower-cost sensors.

(iii) Cross-modal Spatial Alignment: To address the spatial misalignment caused by different camera mounting positions in practical UAV systems, future work will investigate efficient image registration and online calibration to enhance pixel-level fusion in dynamic environments.

References

[1] Calum X. Cunningham, Grant J. Williamson, and David M. J. S. Bowman. Increasing frequency and intensity of the most extreme wildfires on Earth. *Nature Ecology & Evolution*, 8(8):1420–1425, August 2024. Publisher: Nature Publishing Group.

[2] Bahador Khaleghi, Alaa Khamis, Fakhreddine O. Karray, and Saiedeh N. Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44, 2013.

[3] Dongze Jin, Feng Shao, Zhengxuan Xie, Baoyang Mu, Hangwei Chen, and Qiuping Jiang. Cafcnet: Cross-modality asymmetric feature complement network for rgb-t salient object detection. *Expert Systems with Applications*, 247:123222, 2024.

[4] Ying Lv, Zhi Liu, and Gongyang Li. Context-aware interaction network for rgb-t semantic segmentation. *IEEE Transactions on Multimedia*, 26:6348–6360, 2024.

[5] Wujie Zhou, Shaohua Dong, Meixin Fang, and Lu Yu. Cacfnet: Cross-modal attention cascaded fusion network for rgb-t urban scene parsing. *IEEE Transactions on Intelligent Vehicles*, 9(1):1919–1929, 2024.

[6] Yu Pang, Yang Huang, Chenyu Weng, Jialin Lyu, Chuanyue Bai, and Xiaosheng Yu. Enhanced RGB-T saliency detection via thermal-guided multi-stage attention network. *The Visual Computer*, 41(10):8055–8073, August 2025.

[7] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. `https://github.com/ultralytics/ultralytics`, 2023. Version 8.0.0, GitHub repository.

[8] Yuanlin Chen, Zengbao Sun, Cheng Yan, and Ming Zhao. Edge-guided feature fusion network for rgb-t salient object detection. *Frontiers in Neurorobotics*, 18:1489658, 2024.

[9] Weihong Ma, Kun Wang, Jiawei Li, Simon X. Yang, Junfei Li, Lepeng Song, and Qifeng Li. Infrared and visible image fusion technology and application: A review. *Sensors*, 23(2), 2023.

[10] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019.

[11] Zichang Liu, Zhiqiang Tang, Xingjian Shi, Aston Zhang, Mu Li, Anshumali Shrivastava, and Andrew Gordon Wilson. Learning multimodal data augmentation in feature space, 2023.

[12] Chiheng Wei, Lianfa Bai, Xiaoyu Chen, and Jing Han. Cross-modality data augmentation for aerial object detection with representation learning. *Remote Sensing*, 16(24), 2024.

[13] Yalin Zhang, Xue Rui, and Weiguo Song. A uav-based multi-scenario rgb-thermal dataset and fusion model for enhanced forest fire detection. *Remote Sensing*, 17(15), 2025.