LightGR-Transformer: Light Grouped Residual Transformer for Multispectral Object Detection

Mingming Li, Fei Wu*, Yinjie Wang
Nanjing University of Posts and Telecommunications
Nanjing, Jiangsu, China
yuemingli319@gmail.com, wufei_8888@126.com, b21041302@njupt.edu.cn

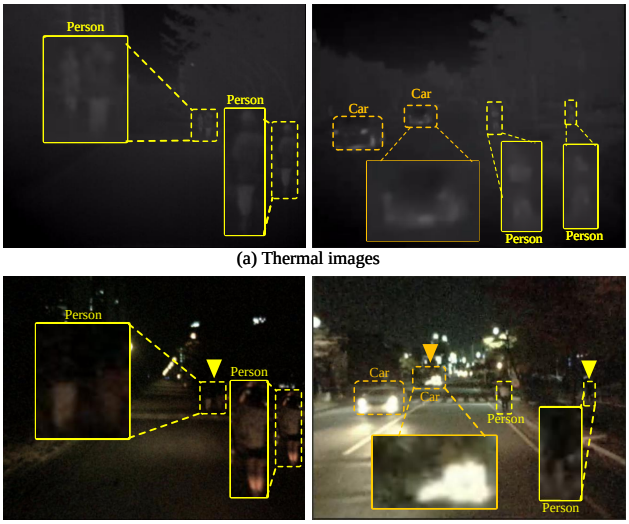
Abstract

In conditions of low visibility, such as nighttime or restricted vision, the fusion of thermal and visible light features becomes particularly important. The Transformer technology has gained popularity in recent research for its ability to capture global information, showing better performance in feature fusion than traditional CNNs. However, the Transformer-based approaches often lead to high computational complexity during forward propagation. To solve this problem, we propose a Light Grouped Residual Transformer (LightGR-Transformer). It uses a single-layer network with channel-wise grouping and learnable residual connections, reducing the complexity of multi-layer Transformers. This design reduces computational cost while preserving important information during feature fusion. It prevents the loss of key features in deeper networks. Additionally, to improve the detection accuracy of small objects in low-light conditions, we introduce deformable convolution layers during the feature extraction stage. These layers dynamically adjust the receptive field of the convolution kernel, enhancing local detail capture. Our experiments on the FLIR dataset show that LightGR-Transformer improves mAP50 by 3.6% compared to existing state-of-the-art methods. On the KAIST dataset for nighttime detection, our network achieves the lowest MR^{-2} score, reaching state-of-the-art performance. Our detector also reduces computational cost by 30%-40% compared to the most advanced models while maintaining top-level performance.

Keywords: ***Multispectral object detection, transformer, feature fusion, cross-modality***

1. Introduction

Single-modality detection has become a fundamental task in modern computer vision, with widespread applications in areas like surveillance [1] [2] and autonomous driving [3] [4]. At nighttime or with restricted visibility due to occlusion, adverse weather, and insufficient illumination, it is difficult for a single thermal or RGB image to provide complete scene information [5]. To address these limitations, multi-sensor modal fusion has emerged as a key technology in detection tasks [6–9].



(a) Thermal images

(b) RGB images

Figure 1. Comparison between Visible Light (RGB) and Thermal Images in a Nighttime Scene.

In recent years, many studies have focused on the fusion of visible and thermal images [10–14]. As shown in Fig.1, both thermal and RGB images have their distinct advantages: thermal images can capture the thermal radiation characteristics of objects in low-light environments, providing clear outline information, while RGB images contain rich details and color information. We use arrows to highlight objects that are unrecognizable under visible light due to poor illumination, yet their contours become discernible in the thermal image, enabling the model to accurately detect them.

By fusing these two modalities, models can leverage the complementary strengths of both modalities. Thermal images provide adaptability to varying environments, while

*Corresponding author

RGB images offer detailed and colorful information. This combination enhances the accuracy and robustness of detection tasks. The fusion approach enables the model to maintain robust target detection under various weather and illumination conditions, which is especially important in application scenarios that require high reliability, such as security surveillance, autonomous driving, and search and rescue.

CNN-based fusion architectures have made significant progress in multimodal fusion and are widely used in state-of-the-art methods. However, due to the limited receptive fields of CNNs, they struggle to effectively model long-range feature relationships. To address this limitation, CFT [15] introduces a Transformer technology for modality feature fusion. Similarly, ICAF [16] proposes a fusion method based on the cross-attention mechanism to enhance the information complementarity between modalities. However, the existing multimodal fusion methods still face two key challenges. First, they involve high computational complexity. Second, their performance in detecting small objects under low-visibility conditions, such as nighttime, is limited, with a notably higher miss rate compared to daytime scenes. Whereas tasks such as search and rescue rely heavily on small target detection and detection accuracy, existing methods do not perform well enough in this regard.

Specifically, the first problem we intend to address is how to reduce computational complexity while maintaining detection accuracy. The traditional Transformer architecture is difficult to adapt to resource-constrained real-world scenarios due to the large number of layers stacked [17] and the huge computational overhead. Therefore, designing a fusion method that reduces computational overhead and efficiently learns inter-modal information is still an urgent challenge. The second problem is how to improve the ability of small target detection and reduce the miss rate at night.

To address the above problems, this paper proposes a Light Grouped Residual Transformer for Multispectral Object Detection.

Our main contributions are concluded as follows:

(1) To address the first problem, we propose a novel multimodal fusion algorithm called Light Grouped Residual Transformer (LightGR-Transformer). The algorithm improves on the traditional transformer-based stacked coding in many ways, adopting channel-wise grouping as well as a simplification strategy, eliminating the multi-layer stacking of the traditional transformer, and using only a single-layer network with a learnable residual structure. This design significantly reduces the computational cost, greatly improving the efficiency of resource utilization while shortening both training and inference time. In addition, LightGR-Transformer can efficiently leverage inter-modal information interactions during the training process, making the fused features more representative and discriminative.

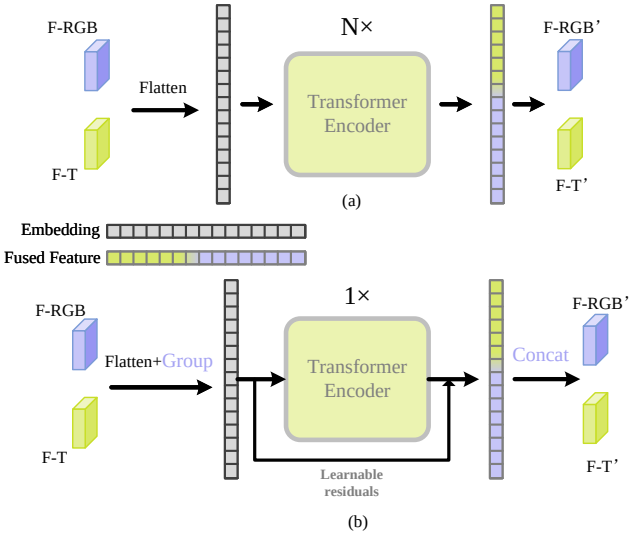


Figure 2. Comparison between existing Transformer architectures and our structure.

(2) To address the second problem, we design a deformable large kernel convolutional layer that dynamically adjusts the receptive field of the convolutional kernel, enhancing the ability to capture local details. This design enables the model to better represent deformed objects and irregular local features, improving small object detection and reducing miss rate under poor visual conditions. Comparison between existing Transformer architectures and our structure is shown in Fig.2.

(3) Through experimental validation on the KAIST dataset [18] and FLIR dataset [19], our method achieves SOTA in both detection accuracy and computational efficiency.

2. Related Work

2.1. Multispectral object detection

Multispectral pedestrian detection has made continuous progress in recent years. Hwang *et al*. [20] established the first multispectral pedestrian detection benchmark, and they proposed a manual feature approach based on Aggregated Channel Features (ACF), which could process color-thermal image pairs simultaneously. To improve the consistency of different modalities, Zhang *et al*. [21] proposed a method for cyclic fusion and refinement of multispectral features. To address the effect of diurnal illumination variations on detection, Guan *et al*. [22] and Li *et al*. [23] proposed illumination-aware modules to predict the light weights from the images and weight the RGB and thermal images through gate function. Considering the inter-modal alignment problem, Zhang *et al*. [24] used a Regional Feature Alignment (RFA) module to predict the feature shift between RGB and thermal images. In order to solve modality

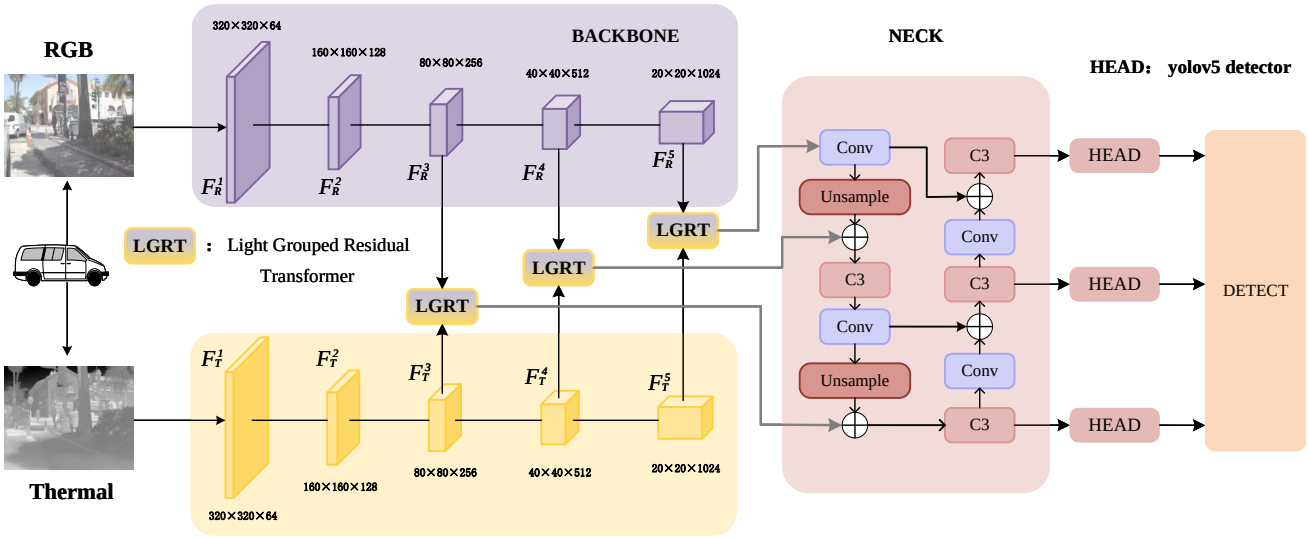


Figure 3. Structure of our object detection framework. The backbone functions as a mono-modal feature extractor, generating five thermal and RGB feature maps of different sizes respectively. Our proposed LGRT modules fuse thermal and RGB features, and the fused features are then fed into the neck network. The neck module performs multiscale feature aggregation, and the head module outputs the final detection results.

imbalance problems, Zhou *et al*. [25] designed two models called DMAF and IAFA, respectively. Kim *et al*. [6] proposed a new loss function to mitigate the modal differences considering that color and thermal cameras have different field-of-views. CFT [15] and LGADet [26] adopt a self-attention mechanism using transformer, which is different from previous CNNs-based approaches, to combine thermal features with RGB features by learning long-range dependencies and integrating global contextual information in the feature extraction stage. Based on CFT, ICAF [16] further proposes cross-attention to enhance inter-modal global feature interaction, aiming at solving the performance degradation problem caused by image misalignment in multispectral image fusion, and also adopting an iterative interaction mechanism with parameters sharing to reduce model complexity and computation cost.

2.2. Attention-based Approach

Attention mechanism is a technique inspired by the human visual system. The key idea is to assign different weights to different parts of the input data. This allows the model to focus on the most important information for the task while ignoring or weakening irrelevant parts. In recent years, there has been a great deal of research work on attention mechanisms. SENet [27] pioneers channel attention and proposes a novel channel-attention network. CBAM [28] proposes a lightweight and general module to enhance the adaptive feature optimization ability of convolutional neural networks by sequentially inferring attention maps in both channel and spatial dimensions and multiplying them with the input feature map. ECANet [29] proposes a local cross-channel interaction strategy without dimensionality reduction and adaptively selects kernel size of 1D convolution with very small parameters for performance improvement.

Inspired by Azad *et al*. [19], we integrate a deformable large kernel attention mechanism into our work. This mechanism efficiently captures the global context through a large convolutional kernel, similar to the self-attention mechanism but with lower computational cost. For image detection tasks, it combines with deformable convolution to flexibly adjust the sampling grid, which can adapt to the shape and structure of different objects, thus improving the capture ability and detection accuracy of target features.

3. Method

3.1. Structure

As shown in Fig.3, our model adopts a dual-branch backbone network for feature extraction for RGB-thermal image pairs. To fully utilize the complementary properties of the two modal information. In the 3rd, 4th, and 5th layers of our network's backbone, we introduce the Light Grouped Residual Transformer module to fuse thermal and RGB image features at multiple levels, producing high-quality hybrid feature maps. The fused feature map is then passed to the neck to enhance multi-scale feature representation, and the detection head accurately localizes and classifies the target.

Given thermal image X_T and RGB image X_{RGB}, a dual-

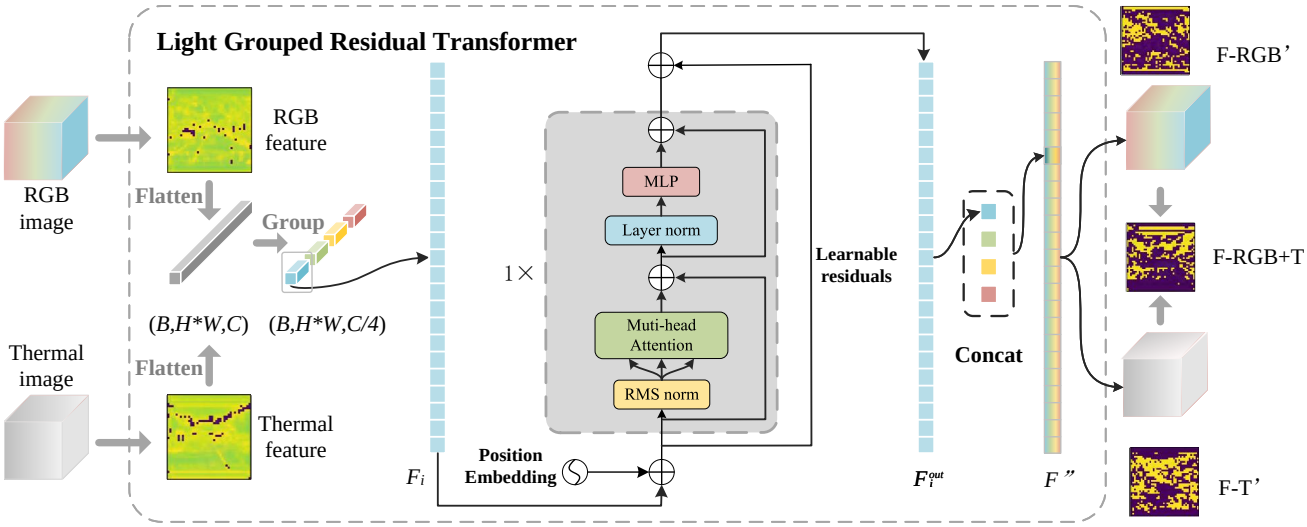


Figure 4. Illustration of the proposed LGRT module. H and W represent the height and width of the feature map, and C is the number of channels. we split C into G groups (with $G = 4$). F_i represents the feature after grouping, and F_i^{out} denotes the output feature of the i-th sub-block after applying the multi-head attention mechanism and learnable residual connection. The final output feature F'' is obtained through concatenation.

branch backbone network is used to extract features from the two images separately. The formula is shown below :

$$F_T = \Phi_{backbone}\ (X_T; \delta_T) \tag{1}$$

$$F_{RGB} = \Phi_{backbone}\ (X_{RGB}; \delta_{RGB}) \tag{2}$$

where $\Phi_{backbone}$ function performs feature extraction on the input image. Commonly used backbone functions are CSPdarkNet [18], ResNet [30], VGG16 [31], *etc*. The original input image is transformed into a multi-layer feature map after backbone extraction. In our model, feature extraction is performed separately on both the RGB image and the thermal image. F_T and F_{RGB} denote the feature maps extracted by the thermal branch and RGB branch, respectively. δ_T and δ_{RGB} denote the parameters of the feature extraction formula.

For feature fusion in the i-th layer, the formula of the LightGR-Transformer module is as follows:

$$F_{fusion}^{i} = \text{LightGR-Transformer}\left(F_T^i; F_{RGB}^i; \theta_f\right) \tag{3}$$

where LightGR-Transformer$(\cdot)$ denotes the formula for the fusion of thermal and RGB image features with parameter θ_f. The work in HalfwayFusion [32] explores four different fusion architectures and the results show that the Halfway fusion approach is more desirable. Based on these findings, we adopt Halfway fusion architecture in our framework. While traditional fusion methods, such as addition operation and NIN [33] have been widely used, they often struggle to capture the complex associations between multimodal features, particularly in complex scenes. In recent years, fusion approaches based on the attention mechanism have gradually become mainstream, exhibiting stronger feature capture capabilities. In this context, we improve this fusion approach for Transformer and propose the LightGR-Transformer module to more effectively mine and fuse the complementary information of thermal and RGB image features to enhance the model's performance in multimodal target detection tasks. F_{fusion}^i denotes the output after the fusion of the features in the i-th layer. F_T^i and F_{RGB}^i denote the features of thermal and RGB images in the i-th layer of the backbone, respectively.

The fused multi-scale feature map F_{fusion}^i is input into the neck structure for further feature processing and enhancement as shown in the following equation:

$$\{\boldsymbol{B}, \boldsymbol{C}\} = \phi_{Head}\left(\phi_{Neck}(F_{fusion}; \theta_n); \theta_h\right) \tag{4}$$

where $\phi_{Neck}(\cdot)$ denotes the multi-scale feature aggregation function of the neck module, which is used for feature enhancement and multi-scale fusion, and in existing research, structures such as FPN and PANet are often used to realize the functions of the neck module [34, 35]. $\phi_{Head}(\cdot)$ denotes the detection head function for target detection, including bounding box classification and regression. θ_n and θ_h denote the parameters of the neck module and the detection head, respectively. $\{\boldsymbol{B}, \boldsymbol{C}\}$ denotes the detection result, where $\boldsymbol{B}$ is the bounding box set and $\boldsymbol{C}$ is the category set. In this paper, we use the detection header of YOLO [18].

3.2. LightGR-Transformer

3.2.1 Design Idea of LightGR-Transformer

Traditional transformer networks usually capture complex feature relationships in the input data by stacking multiple modules (*e.g*. 8 or more) in the process of feature extraction. While this stacked approach improves the model's representation capability, it also increases computational complexity. Additionally, as the number of layers grows, the model tends to introduce more noise and redundant information. In the process of multiple linear transformations and nonlinear activation, important features may be gradually compressed or even lost. This problem is particularly pronounced in deep networks, where the risk of information loss is higher. To address these challenges, we propose a more efficient method: Light Grouped Residual Transformer (LightGR-Transformer) module.

How can we enhance feature preservation while significantly improving computational efficiency through our channel-wise grouping strategy and the use of a single-layer Transformer?

In the design of the single-layer Transformer, each feature group is processed independently through multi-head attention, combined with learnable positional embeddings to form localized feature representations. The multi-head attention mechanism captures both intra-modal and inter-modal feature relationships within each group, while residual connections ensure the preservation of original feature information.

This grouping strategy eliminates the need for multi-layer stacking, achieving efficient compression and retention of information within a single layer. Since each group independently performs learning and attention computation, the model is able to capture different levels of feature relationships in each subspace. This process can be viewed as a finer-grained decomposition of the original high-dimensional space, effectively mapping it into multiple low-dimensional subspaces for learning, which enhances the model's ability to capture semantic information.

3.2.2 The Design of LightGR-Transformer Structure

As shown in Fig.4, the RGB and thermal images are separately spread into sequences, then spliced over the channel to form a feature representation F_{fusion} with comprehensive semantics. (B, $H*W$, C) is the reshaped shape of F_{fusion}. B represents the batch size, C represents the number of channels. H and W represent the image height and width respectively. C_f represents the number of channels of F_{fusion}.

In order to process this channel information more efficiently, this paper proposes to group the channel dimension C. C is divided into G groups with each channel size of $\frac{C_f}{G}$. For example, we set $G = 4$, then the feature F_{fusion} is divided into four groups to get $F_{reshape}$. The formula is shown below :

$$F_{reshape} = [F_1, F_2, F_3, F_4], \quad F_i \in \mathbb{R}^{B*(H*W)*\frac{C_f}{G}} \tag{5}$$

where every learnable position embedding P_i is added separately to preserve the spatial position information in each grouping, which is then processed by the MultiHead attention function MultiHead $(\cdot)$ respectively to realize the full interaction of inter- and intra-modal features. On this basis, the processed features are residually connected with their original grouped values F_i. The function can be represented as follows:

$$F_i^{out}=\text{MultiHead}(F_i+P_i) \ + \rho \cdot F_i \tag{6}$$

where F_i^{out} denotes the output features of the i-th sub-block after the multi-attention mechanism and learnable residual connection. ρ is a learnable weight parameter that adaptively adjusts the weights between the original and fused features.

Finally, splice F_i^{out} together in the channel dimension to form the fused final feature $F^{''}$. The function can be represented as follows:

$$F^{''} = concat(F_1^{out}, F_2^{out}, F_3^{out}, F_4^{out}) \in \mathbb{R}^{B*(H*W)*C_f} \tag{7}$$

This design not only fully utilizes the advantages of the multi-head attention mechanism, but also preserves the information of the original features through residual connection, which makes the final output feature $F^{''}$ have both global information and local details.

In this paper, the number of heads of the multi-head attention mechanism is set to 8, which can help the model to better understand and capture the correlation between thermal and RGB features, and to ensure that the information is adequately expressed in the fusion process.

3.2.3 Comparison of Computation Cost with Other Transformer-based Fusion Modules

The attention mechanism is applied to each $F_i \in \mathbb{R}^{B*N*\frac{C_f}{G}}$ separately.

Given the input matrix I, we map it to the weight matrix Q, K, V, computed as follows:

$$Q = I \cdot W^Q, K = I \cdot W^K, V = I \cdot W^V \tag{8}$$

where the dimension of W is $\mathbb{R}^{C*D_k}$, is used to generate the weight matrix of query, key, and value.

The formula of the attention mechanism is shown below:

$$Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{D_K}}\right) \cdot V \tag{9}$$

Table 1. The comparison of computation cost between existing methods and LightGR-Transformer (where n is the number of stacked transformer encoder blocks, N is the number of tokens, and C is the number of channels, CFE is the module used in ICAF).

Step	CFT(n=8)	CFE(n=8)	Ours(n=1)
QK^T	$O(4N^2 \times C)$	$O(2N^2 \times C)$	$O(\frac{1}{2}N^2 \times C)$
Softmax $\left(\frac{QK^T}{\sqrt{D_K}}\right)$	$O(4N^2)$	$O(2N^2)$	$O(\frac{1}{2}N^2)$
Softmax $\left(\frac{QK^T}{\sqrt{D_K}}\right) \cdot V$	$O(4N^2 \times C)$	$O(2N^2 \times C)$	$O(\frac{1}{2}N^2 \times C)$
FFN	$O(16N \times C^2)$	$O(8N \times C^2)$	$O(N \times C^2)$
TOTAL	$O(4N^2 \times C + 16N \times C^2)$	$O(2N^2 \times C + 16N \times C^2)$	$O(\frac{1}{2}N^2 \times C + N \times C^2)$

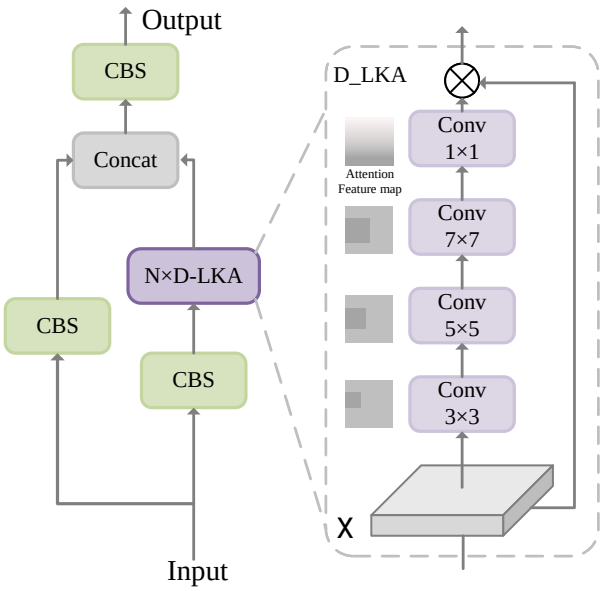


Figure 5. Structure of C3_LK. The CBS module in YOLO is a fundamental unit that combines convolution, batch normalization, and activation functions to efficiently extract and process image features.

where in the input matrix $I \in \mathbb{R}^{L*D_k}$, L is the token length and D_k is the input data dimension. In this study, $D_k = C$, *i.e*. the input and weight matrix are dimensionally identical.

After the attention mechanism, the output of each attention head is passed through a feed-forward network (FFN). The FFN consists of two fully connected layers, with a nonlinear activation function in between. The formula for the FFN can be expressed as:

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \qquad (10)$$

where x is the input from the attention mechanism. $W_1 \in \mathbb{R}^{D_{\text{model}} \times D_{\text{ff}}}$ and $W_2 \in \mathbb{R}^{D_{\text{ff}} \times D_{\text{model}}}$ are weight matrices. b_1 and b_2 are bias terms. D_{ff} is the dimension of the hidden layer (larger than the input/output dimension D_{model}).

Both CFT and ICAF are Transformer-based fusion methods, and they use 8 layers of multi-head attention mechanisms to capture inter- and intra-modal feature relationships. Table 1 shows a detailed comparison of the computational cost for each step.

Our network significantly reduces the number of parameters compared to the CFT and the CFE module in ICAF [16]. Among all Transformer-based fusion methods, our method reduces parameters by over 60% compared to CFT, and by 30%-40% compared to CFE.

3.3. C3 with Deformable Large Kernel Attention

In order to enhance the detection ability of the model for small targets and complex deformed objects, this paper designs a C3_LK module, integrates D_LKA (Deformable Large Kernel Attention) and replaces the C3 module in backbone with it. This module dynamically adjusts the receptive field of the convolutional kernel through a series of deformable convolutional layers to flexibly capture the positional information in the input feature map, thus improving the ability to capture local details. The deformable convolution flexibly adjusts the sampling position of the standard convolution by introducing additional offsets, enabling the convolution kernel to adapt itself to changes in image content.

Compared to the standard C3 module, C3_LK allows the model to better model deformed objects and irregular local features, especially in detecting small targets and deformed objects in complex scenes.

As shown in Fig.5, the first layer of convolution is used to extract the underlying localized features, which has a smaller receptive field to capture more subtle features. The second layer of convolution further expands the receptive field and is used to initially generate the attention map. The third layer of convolution (7x7 inflated deformable convolution) combines the properties of dilation convolution to capture more complex spatial features with a wider range of receptive field while avoiding too much information loss.

The attention maps are generated sequentially through the above convolutional operations, and then the number of channels is finally adjusted by a 1x1 convolutional layer. The function of this step is to multiply the convolutional processed feature information with the original input element by element, to strengthen the feature expression of key regions, and thus promote the model's attention to important local regions.

4. Experiments

4.1. Datasets

KAIST Dataset. KAIST [20] is a widely used multispectral pedestrian detection benchmark dataset containing scenes under different lighting conditions. Because this dataset is taken from the video continuous frame images, and the neighboring images do not differ much, it is cleaned to some extent. The dataset retains only person, cyclist, and merges the two labels into person. Blank labels and their corresponding images are removed. The dataset has 7,595 image pairs for training and 1,400 image pairs for testing. The performance of the KAIST dataset is usually evaluated based on the log-average miss rate metric [36].

FLIR Dataset. FLIR [37] is a multispectral target detection dataset covering both day and night scenes. The dataset contains 5,142 pairs of multispectral images, of which 4,129 pairs are used for training and 1,013 pairs are used for testing. The dataset consists of three types of targets, namely person, cars, and bicycles. Given that there are alignment errors in the images in the original dataset, a corrected FLIR-aligned version [21] is used for comparative analysis to improve the accuracy of the experiments.

4.2. Evaluation Metrics

Average Precision. Average Precision is a fundamental evaluation metric in object detection, used to measure the accuracy of a detection model. It reflects the model's ability to correctly detect objects under different thresholds. Typically, the IoU threshold is set to 0.5, meaning that when the IoU between the model's predicted bounding box and the ground truth is greater than or equal to 0.5, it is considered a True Positive. Mean Average Precision (mAP) is the average of AP across different categories, commonly used in multi-class object detection tasks to indicate overall detection performance. mAP50, a widely used metric, refers to the AP calculated at an IoU threshold of 0.5. In this paper, the evaluation metrics on the FLIR dataset include mAP50, AP for each category, and overall mAP. Higher values for these metrics indicate more accurate detection results.

Log-average Miss Rate. In this paper, the evaluation metric log-average miss rate (MR^{-2}) [36] is used for the KAIST dataset. The relationship between different log (MR) and FPPI can be obtained by adjusting different thresholds of detection confidence. 9 FPPIs within the range of [0.01,1] are uniformly selected, and their corresponding 9 log (MR) values are obtained. These vertical coordinate values are averaged, and then the averaged MR^{-2} is computed. The smaller the metric, the higher the performance of the detector.

4.3. Implementation Details

Our method is implemented on a Windows system with PyTorch version 1.10.0+cu113, utilizing an i5-14600KF CPU and one NVIDIA GeForce RTX 4070 Ti SUPER GPU. We train the models using a stochastic gradient descent (SGD) optimizer with an initial learning rate of 1e-2, momentum of 0.937, and weight decay of 0.0005. The training runs for 500 epochs with workers of 2 and a batch size of 4. To enhance performance, we initialize the models with pretrained weights from the YOLOv5 model, which is trained on the COCO dataset [38].

4.4. Ablation Study

4.4.1 The Effect of the Number of LightGR-Transformer Modules

To evaluate the performance of the LightGR-Transformer module across different backbones, we design three comparison experiments. we design three comparison experiments using CSPDarkNet, VGG16, and ResNet50. For each backbone, we test the following configurations:

(1) Using the LightGR-Transformer module for fusion at the 5th layer;

(2) Using the LightGR-Transformer module for fusion at both the 4th and 5th layers;

(3) Using the LightGR-Transformer module for fusion at the 3rd, 4th, and 5th layers.

These experiments systematically analyze the performance of the LightGR-Transformer module across different backbones and layer combinations. The default number of groups for the LightGR-Transformer is set to 4. The results are shown in Table 2 below.

It is clear that our method significantly reduces the number of parameters compared to the CFE module in ICAF [16]. In terms of performance, our method surpasses the ICAF, primarily due to the efficient feature utilization of the LightGR-Transformer module. By slicing the channels, the interference from redundant information is effectively minimized. This allows the model to focus more on the key features of each sliced block, enhancing the effectiveness of feature representation. Each sub-block created through slicing focuses more on learning local features. This reduces the impact of noise during the overall computation, thereby improving accuracy. Using three LightGR-Transformer modules does not outperform a network with only two modules, even though it achieves the least computational cost.

4.4.2 The Effect of the Number of Groups

To assess the impact of different group numbers on performance, we conduct experiments by dividing the channels

Table 2. Performance comparison of the LightGR-Transformer module across different backbone networks and layer configurations. The best results are highlighted in **bold black**.

Backbone	Fusion Method	Params	(FLIR) mAP50(%)	(KAIST) mAP50(%)	(KAIST) MR^{-2}(%)
CSPdarkNet	3CFE	120.2m	79.20	78.11	**7.17**
	+LightGR-Transformer	88.1m	82.45	79.15	7.45
	+2LightGR-Transformer	79.8m	**82.80**	**81.43**	7.38
	+3LightGR-Transformer	77.6m	81.00	79.11	7.61
VGG	3CFE	62.2m	70.12	75.92	15.46
	+LightGR-Transformer	54.1m	**74.88**	76.14	313.76
	+2LightGR-Transformer	45.8m	74.66	**76.86**	**13.62**
Resnet50	3CFE	313.8m	70.50	74.87	**12.68**
	+LightGR-Transformer	185.6m	72.49	75.92	12.87
	+2LightGR-Transformer	153.2m	**73.39**	**76.03**	13.01

Table 3. Performance comparison of LightGR-Transformer with different group numbers and layer configurations. The best results are highlighted in **bold black**.

	Add LightGR-Transformer	Number of Groups	Params	(FLIR) mAP50(%)	(KAIST) mAP50(%)	(KAIST) MR^{-2}(%)
CSPdarkNet +YOLOv5 detector	I=5	Group=2	89.84m	82.40	79.23	7.76
		Group=4	88.07m	**82.76**	**81.32**	**7.42**
		Group=8	87.63m	81.35	81.21	7.53
	I=4,5	Group=2	82.02m	82.21	81,14	7.44
		Group=4	79.80m	**82.43**	81.04	**7.38**
		Group=8	79.26m	82.35	**81.27**	7.76

Table 4. Experiments of replacing LayerNorm with RMSNorm.(The experiments from left to right are as follows: keeping both LayerNorms, replacing the LayerNorm before Attention with RMSNorm, replacing the LayerNorm after Attention with RMSNorm, and finally replacing both LayerNorms with RMSNorm to compare their effects.)

Metric	both LN	RN Pre-Attention	RN Post-Attention	both RN
FPS(Hz)	42.2	42.5	43.2	43.3
(FLIR)mAP50(%)	82.23	82.69	82.50	81.62

into 2, 4, and 8 groups. The goal is to determine the optimal grouping strategy. We compare the performance of using the LightGR-Transformer module for fusion at the 5th layer alone versus using it for fusion at both the 4th and 5th layers. The results are shown in Table 3. By analyzing the results of these experiments, we evaluate the effect of the number of groups on the overall performance of the model.

In our experiments, we find that using 4 groups has relatively better performance. This is likely because grouping into 4 strikes a good balance between information sharing and computational complexity. Compared to using 2 or 8 groups, 4 groups effectively capture detailed feature information while avoiding excessive isolation or computational overhead, leading to improved overall performance.

Additionally, we observe that replacing LayerNorm with RMSNorm before the multi-attention mechanism modestly improves training performance. Unlike LayerNorm, which re-centers and re-scales the input, RMSNorm simplifies the process by only re-scaling using the root mean square, focusing on scaling invariance rather than translation invariance. The results are shown in Table 4.

Although this replacement accelerates training, the speedup is not particularly significant in our case, as the number of normalization operations is relatively small and the overall computational load is already low. However, RMSNorm still helps preserve useful offset information, enabling the attention mechanism to better capture subtle relationships in the input sequence.

4.5. Comparison Experiment

4.5.1 Comparison with State-of-the-art Methods

On the KAIST dataset, as shown in Table 5, our method performs better than other SOTA methods on the MR^{-2} metric at nighttime, demonstrating superior performance in low-visibility conditions. Additionally, on the MR^{-2} metric, our model performs at a similar level to ICAF, while reducing the computational cost by 30%-40% compared to the ICAF architecture.

On the FLIR dataset, our method is compared with other SOTA methods, and the comparison results are shown in Table 6. Our method surpasses all SOTA methods in both

Table 5. Comparison on the KAIST Dataset. The best results are highlighted in **red** and the second-place results are highlighted in **blue**.

Method	Miss Rate(%)↓			FPS(Hz)	Platform
	All	Day	Night		
FusionRPN+BF [39]	18.29	19.57	16.27	-	-
HalfwayFusion **[32]**	25.77	24.91	26.67	2.33	TITAN X
IAF-RCNN [23]	15.57	14.81	16.70	4.76	TITAN X
IATDNN-IAMSS [22]	14.46	14.18	15.28	4.00	TITAN X
MBNet [25]	8.40	8.62	76.10	14.29	GTX 1080Ti
MLPD [40]	7.58	**7.96**	6.95	-	-
MSDS-RCNN [33]	8.23	8.83	**6.75**	4.55	GTX 1080Ti
ICAF [16]	**7.17**	**6.82**	7.85	38.46	RTX 3090
Ours	**7.38**	8.53	**5.54**	45.68	RTX 4070TiS

Table 6. Comparison on the FLIR Dataset. The best results are highlighted in **red** and the second-place results are highlighted in **blue**.

Method	AP50 (%)			mAP50(%)	mAP(%)
	Bicycle	Car	Person		
MMTOD-CG [41]	50.26	70.63	63.31	61.4	-
MMTOD-UNIT [41]	49.43	70.72	64.47	61.5	-
GAFF [42]	-	-	-	72.9	37.3
CFR [8]	57.77	84.91	74.49	72.4	-
BU-ATT [43]	56.10	87.00	76.10	73.1	-
BU-LTT [43]	57.40	86.50	75.60	73.2	-
CFT [15]	61.40	**89.50**	**84.10**	78.3	40.2
ICAF [16]	**66.90**	89.00	81.60	**79.2**	**41.4**
Ours	**71.80**	**90.10**	**85.60**	**82.8**	**40.7**

AP50 across all categories and mAP50.

4.5.2 Qualitative Analysis

As shown in Fig.6, red markers indicate missed detections, and yellow markers represent false detections. These examples are from five different scenarios. The fusion of thermal and RGB features leverages the complementary strengths of both modalities. The first and second rows show detection results using only RGB and thermal features, respectively, while the third row displays results after fusing RGB and thermal features using our method. The improvements in reducing both missed and false detections are clearly evident.

As shown in Fig.7, this is a comparison of our method with ICAF for small object detection. As shown, the first and second rows are from the same scene. In the first row, due to the long shooting distance and lack of image clarity, the object features are unclear, causing ICAF to miss the detection. Even in the second and third row, where the objects are larger, ICAF still fails to detect them accurately. However, with the application of the deformable large kernel attention mechanism, our method demonstrates excellent performance in small object detection, effectively capturing and recognizing these hard-to-detect objects.

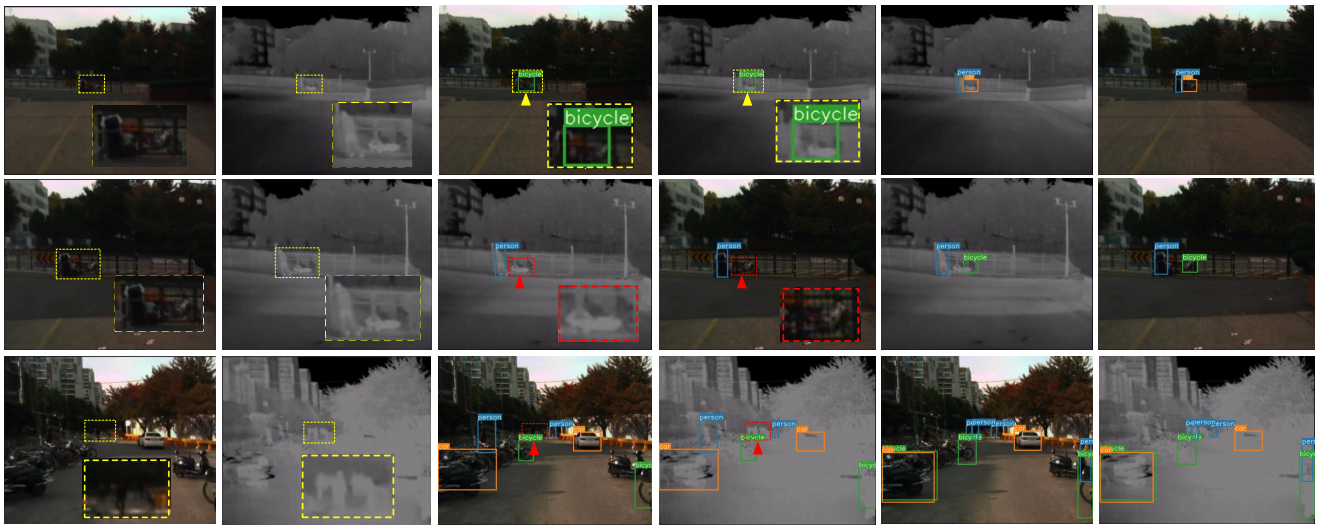
As shown in Fig.8, The first row shows the ground truth, while the second row presents the attention heatmap visualization of the ICAF model. The third row displays the attention heatmap visualization of our method. It is evident that, compared to ICAF, our method more accurately captures the key features of the object. Additionally, our method demonstrates a stronger focus on the object area, effectively suppressing background interference. This improvement significantly enhances the model's detection accuracy.

5. Conclusions

This paper proposes an efficient Light Grouped Residual Transformer (LightGR-Transformer) module for the fusion of thermal and visible features. Compared to traditional multi-layer Transformer structures, LightGR-Transformer significantly reduces computational cost through channel-wise grouping and learnable residual connections while effectively preserving key information. Additionally, a deformable convolutional layer is introduced to improve the detection accuracy of small targets under low-visibility conditions. Experimental results show that LightGR-Transformer reduces computation costs by 30%-40% while maintaining excellent performance on multiple datasets, providing strong support for the low-cost deployment of RGB-T detection.

(1) (2) (3) (4) (5)

Figure 6. The illustration shows detection results based on different feature sets. The first row presents the results using only RGB features. The second row shows the detection based solely on thermal features. The third row displays the results after feature fusion of both RGB and thermal. **Red** arrows highlight the missed detections, while **yellow** arrows mark the false detections.



(a) RGB and thermal images (b) ICAF detection results (c) ours detection results

Figure 7. The results on the KAIST dataset. Columns 1 and 2 represent the RGB and thermal images, respectively. Columns 3 and 4 show the ICAF detection results after feature fusion, while columns 5 and 6 present our module's detection results after feature fusion. **Red** arrows highlight the missed detections, while **yellow** arrows mark the false detections.

6. Acknowledgment

This work was supported by National Natural Science Foundation of China (No. 62076139), and National Key Laboratory of Information System Engineering (No. 05202305).

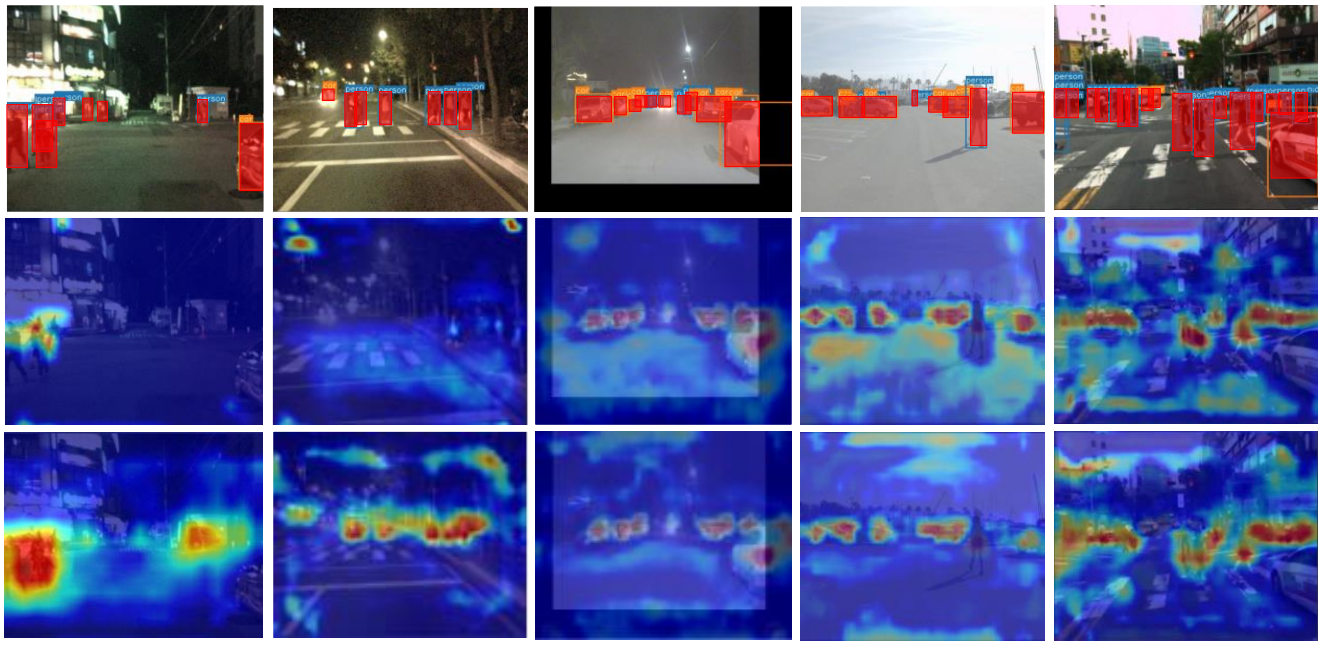


Figure 8. Visualization of Attention Maps on KAIST and FLIR Datasets. The first row shows ground truth, the second row displays the attention heatmaps from the ICAF network, and the third row presents the attention heatmaps from our method.

References

[1] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 193:102907, 2020. 1

[2] Jacinto C Nascimento and Jorge S Marques. Performance evaluation of object detection algorithms for video surveillance. *IEEE Transactions on Multimedia*, 8(4):761–774, 2006. 1

[3] Huanqian Yan, Bo Li, Hong Zhang, and Xingxing Wei. An antijamming and lightweight ship detector designed for spaceborne optical images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4468–4481, 2022. 1

[4] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1

[5] Xingxing Wei and Shiji Zhao. Boosting adversarial transferability with learnable patch-wise masks. *IEEE Transactions on Multimedia*, 26:3778–3787, 2023. 1

[6] Jung Uk Kim, Sungjune Park, and Yong Man Ro. Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1510–1523, 2021. 1, 3

[7] Shaoyue Song, Zhenjiang Miao, Hongkai Yu, Jianwu Fang, Kang Zheng, Cong Ma, and Song Wang. Deep domain adaptation based multi-spectral salient object detection. *IEEE Transactions on Multimedia*, 24:128–140, 2020. 1

[8] Zhengxuan Xie, Feng Shao, Gang Chen, Hangwei Chen, Qiuping Jiang, Xiangchao Meng, and Yo-Sung Ho. Cross-modality double bidirectional interaction and fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):4149–4163, 2023. 1, 9

[9] Kunpeng Wang, Zhengzheng Tu, Chenglong Li, Cheng Zhang, and Bin Luo. Learning adaptive fusion bank for multi-modal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7344–7358, 2024. 1

[10] Maoxun Yuan, Yinyan Wang, and Xingxing Wei. Translation, scale and rotation: cross-modal alignment meets rgb-infrared vehicle detection. In *European Conference on Computer Vision*, volume 13669, pages 509–525, 2022. 1

[11] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. 1

[12] Maoxun Yuan and Xingxing Wei. C^2former: Calibrated and complementary transformer for rgb-infrared object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. 1

[13] Qing Li, Changqing Zhang, Qinghua Hu, Pengfei Zhu, Huazhu Fu, and Lei Chen. Stabilizing multispectral pedes-

trian detection with evidential hybrid fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):3017–3029, 2023. 1

[14] Qing Li, Changqing Zhang, Qinghua Hu, Huazhu Fu, and Pengfei Zhu. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 25:3420–3431, 2022. 1

[15] Qingyun Fang, Dapeng Han, and Zhaokui Wang. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021. 2, 3, 9

[16] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024. 2, 3, 6, 7, 9

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words. *arXiv preprint arXiv:2010.11929*, 7, 2020. 2

[18] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2, 4

[19] Reza Azad, Leon Niggemeier, Michael Hüttemann, Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Yury Velichko, Ulas Bagci, and Dorit Merhof. Beyond self-attention: Deformable large kernel attention for medical image segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1287–1297, January 2024. 2, 3

[20] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1037–1045, 2015. 2, 7

[21] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing*, pages 276–280, 2020. 2, 7

[22] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019. 2, 9

[23] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019. 2, 9

[24] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *the IEEE/CVF International Conference on Computer Vision*, pages 5127–5137, 2019. 2

[25] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *European conference on computer vision*, pages 787–803, 2020. 3, 9

[26] Xin Zuo, Zhi Wang, Yue Liu, Jifeng Shen, and Haoran Wang. Lgadet: Light-weight anchor-free multispectral pedestrian detection with mixed local and global attention. *Neural Processing Letters*, 55(3):2935–2952, 2023. 3

[27] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 3

[28] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *European Conference on Computer Vision*, pages 3–19, 2018. 3

[29] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11534–11542, 2020. 3

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[32] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016. 4, 9

[33] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. *arXiv preprint arXiv:1808.04818*, 2018. 4, 9

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 4

[35] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018. 4

[36] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2011. 7

[37] F. A. Team. FLIR ADAS Dataset. https://www.flir.cn/oem/adas/adas-dataset-form/. [Online; accessed 8-June-2024]. 7

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014. 7

[39] Daniel Konig, Michael Adam, 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, pages 49–56, 2017. 9

[40] Jiwon Kim, Hyeongjun Kim, Taejoo Kim, Namil Kim, and Yukyung Choi. Mlpd: Multi-label pedestrian detector in multispectral domain. *IEEE Robotics and Automation Letters*, 6(4):7846–7853, 2021. 9

[41] Chaitanya Devaguptapu, Ninad Akolekar, Manuj Sharma, and Vineeth N. Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1029–1038, 2019. 9

[42] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *IEEE/CVF winter conference on applications of computer vision*, pages 72–80, 2021. 9

[43] My Kieu, Andrew D Bagdanov, and Marco Bertini. Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(1):1–19, 2021. 9