ICAFusion: Iterative Cross-Attention Guided Feature Fusion for Multispectral Object Detection

Jifeng Shen[a,*], Yifei Chen[a], Yue Liu[a], Xin Zuo[b], Heng Fan[c], Wankou Yang[d]

[a]School of Electrical and Information Engineering, Jiangsu University, Zhenjiang, 212013, China
[b]School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, 212003, China
[c]Department of Computer Science and Engineering, University of North Texas, Denton, TX 76207, USA
[d]School of Automation, Southeast University, Nanjing, 210096, China

Abstract

Effective feature fusion of multispectral images plays a crucial role in multispectral object detection. Previous studies have demonstrated the effectiveness of feature fusion using convolutional neural networks, but these methods are sensitive to image misalignment due to the inherent deficiency in local-range feature interaction resulting in the performance degradation. To address this issue, a novel feature fusion framework of dual cross-attention transformers is proposed to model global feature interaction and capture complementary information across modalities simultaneously. This framework enhances the discriminability of object features through the query-guided cross-attention mechanism, leading to improved performance. However, stacking multiple transformer blocks for feature enhancement incurs a large number of parameters and high spatial complexity. To handle this, inspired by the human process of reviewing knowledge, an iterative interaction mechanism is proposed to share parameters among block-wise multimodal transformers, reducing model complexity and computation cost. The proposed method is general and effective to be integrated into different detection frameworks and used with different backbones. Experimental results on KAIST, FLIR, and VEDAI datasets show

*Corresponding author
  Email address: shenjifeng@ujs.edu.cn (Jifeng Shen)

that the proposed method achieves superior performance and faster inference, making it suitable for various practical scenarios. Code will be available at https://github.com/chanchanchan97/ICAFusion.

Keywords: Multispectral Object Detection, Cross-Attention, Transformer, Iterative Feature Fusion

1. INTRODUCTION

Thermal spectrum range image provides a special way to perceive the natural scenes, which is believed to complement visible spectrum images in computer vision. Multispectral image feature representation and fusion is a challenging problem, serving a variety of downstream vision tasks, such as object detection, semantic segmentation and object tracking. As a fundamental vision task, object detection remains a hot topic in both academia and industry, and has made considerable progress in small object detection [1], face and pedestrian detection [2] and oriented object detection [3, 4], thanks to the rapid development of convolutional neural networks. However, these methods are still vulnerable to environmental factors, such as severe weather conditions and changing illumination. In order to improve the robustness and accuracy of object detectors in all-weather conditions, multispectral object detection based on both RGB and thermal images has become a viable solution and is gaining popularity in recent academic studies.

Compared with mono-modal object detection, the use of multiple modalities in object detection can provide a richer visual representation of objects, enabling them to effectively compensate for each other. As illustrated in Fig. 1(a), RGB images can provide detailed colors, textures, and contours of objects under good illumination conditions, but these features may not be visible in thermal images. In contrast, Fig. 1(b) shows that in poor illumination conditions, such as night or dark places. It is difficult to distinguish object details and edges from the background in RGB images, but thermal images can still provide perceptible contour features due to their unique energy radiation imaging mechanism. These

<div align="center">(a)   (b)</div>

Figure 1: Visualization of objects in paired RGB-thermal images. The first and second rows are RGB and thermal images respectively. The objects are more perceptible in RGB images (a), while they are much easier to observe in thermal images (b). The dashed windows are cropped and enlarged for better visualization.

characteristics indicate the complementary nature of RGB and thermal modalities. As a result, effective feature fusion from different modalities is critical for multispectral object detection.

In the previous studies, convolution-based feature fusion has been widely used in current state-of-the-art methods. The pioneering work [5] has explored different convolutional neural network (CNN) based fusion architectures and has shown that halfway fusion can lead to desirable performance. Zhang et al. [6] encode the interaction between different modalities and fuse features adaptively. Zhang et al. [7] introduce auxiliary pedestrian masks to guide the feature representation of inter-modal and intra-modal. Fu et al. [8] introduce the attention modules to perform the pixel-level feature fusion from the spatial and channel dimension. However, convolution-based feature fusion only pays attention to local feature information, and it lacks the ability to model long-range feature relationships due to the limited receptive field of CNN. To this end, we propose a novel dual cross-attention transformer fusion method that aggregates the feature information of RGB and thermal modalities from both local and global perspectives. Although transformers have the potential to address this limitation of CNNs, we observe that naive usage in feature fusion can lead to massive feature redundancy, resulting in excessive computational load and memory re-

quirements. To address these issues, we have carefully studied the structure of fusion transformers and have attempted to answer two key questions.

How to effectively borrow complementary information from other modalities ? The performance of multispectral object detection is strongly correlated with the quality of feature fusion. The traditional fusion method relied on feature concatenation or addition, which is susceptible to image misalignment due to the inherent constraint of limited local-range feature interaction. Inspired by pretraining cross-modal feature representations in vision-language tasks [9], we have proposed a cross-attention fusion transformer for multispectral image feature fusion to tackle this issue. This method is designed to capture the complementary features from other modalities, and specifically tailored to enhance both feature branches simultaneously. Additionally, our proposed fusion transformer naturally benefits from its long-range modeling feature interactions, which aid in the discovery of discriminative complementing information from other modalities. Different from the single transformer fusion methods [10] that concatenate the tokens of each modality and compute the queries, keys, values from all modality information, our proposed method computes the correlation across modalities only with queries from the auxiliary modality.

How to efficiently integrate and refine multispectral image features ? Transformer-based models are well known in the vision world for their enormous computational complexity. Besides, most of the existing methods [11] stack numerous blocks to boost performance, resulting in a surge of computational cost. However, humans generally repeatedly review the knowledge after learning new ones, which aids in the retention of what they have learned. Inspired by this, we have proposed an iterative learning strategy. This method not only learns global complementary information based on the bidirectional feature flow interaction between the RGB and thermal branches, but also iteratively refines the feature representation of inter-modality and intra-modality simultaneously, thereby strengthening the discriminative feature information. In contrast to standard methods that stack multiple blocks, our proposed iterative learning strategy shares parameters in each block and improves the balance

4

between model performance and complexity.

To summarize, our main contributions are as follows:

- A novel dual cross-attention feature fusion method is proposed for multispectral object detection, which simultaneously aggregates complementary information from RGB and thermal images.

- An iterative learning strategy is tailored for efficient multispectral feature fusion, which further improves the model performance without additional increase of learnable parameters.

- The proposed feature fusion method is both generalizable and effective, which can be plugged into different backbones and equipped with different detection frameworks.

- The proposed CFE/ICFE module can function with different input image modalities, which provide a feasible solution when one of the modality is missing or has pool quality.

- The proposed method can achieve the state-of-the-arts results on KAIST [12], FLIR [13] and VEDAI [14] datasets, while also obtains very fast inference speed.

The rest of this paper is organized as follows. Section II introduces the related work published in recent years. Section III describes our proposed method in detail. The experimental results are given in Section IV and we conclude the paper in Section V.

2. Related work

2.1. Multispectral Object Detection

Recent multispectral object detection research has made consistent progress, particularly in multispectral pedestrian detection. Hwang et al.[12] have built

the first multispectral pedestrian benchmark and provided a hand-crafted approach based on the Aggregated Channel Feature (ACF) by extending the infrared channel features. To make full use of complementary information between different modalities, Zhou et al. [15] use the ethos of differential amplifiers to understand the consistency and difference between distinct modalities by leveraging common-mode and differential-mode information. Zhang et al. [16] employ a method for cyclically fusing and refining multispectral features, which aims to improve the consistency of both modalities. To capture the discriminative object features in the multispectral images, MSDS-RCNN [17] is the first work that uses semantic segmentation to guide multispectral object detection via multi-task learning. Shen et al. [18] propose a mask-guided mutual attention module and score fusion module based on the anchor-free detector, which achieves the trade-off between accuracy and speed. Taking into account the variation in lighting between day and night, Li et al. [19] have developed an Illumination-aware Network (IAN) that predicts an illumination weight from RGB images via a Gate Unit and weights the results from the RGB and thermal branches. With the goal of addressing the misalignment problem between two modalities, Zhang et al. [20] utilize a Region Feature Alignment (RFA) module to anticipate feature offset between RGB and thermal pictures. Kim et al. [21] take into account the model's many uncertainties and provided a loss function to direct the visual representation of two modalities in the feature-level to be similar. To tackle the issue that the existing methods lack the ability to model long-range dependencies across modalities, CFT [10] and LGADet [22] are proposed to improve the quality of feature fusion with local and global attention mechanisms. However, these methods only simply integrate Transformer or non-local network into detection framework without taking full advantage of complementary information between global features. In this paper, we propose a novel iterative cross-attention interaction method that fully utilizes both local and global feature information between different modalities.

2.2. Attention-based methods

Attention mechanism originates from research on human vision, which is widely used in the computer vision field. SENet [23] proposes a simple yet effective structure to learn the weights between different channels with fully connected network. Inspired by this, SKNet [24] proposes a dynamic selection mechanism that allows each neuron to adaptively adjust its receptive field size based on multiple scales of input information. CBAM [25] proposes a lightweight and general module to adaptively refine the features in both channel and spatial dimensions. ECANet [26] proposes a local cross-channel interaction strategy with an adaptive one-dimensional convolution, which only involves a handful of parameters while bringing clear performance gain. More recently, CANet [27] is proposed with an effective class-specific attention encoding module that learns a class-specific dictionary to encode class attention maps. In this paper, we propose a cross-modal attention module which leverages the complementary information from auxiliary modality to enhance the mono-modality feature representation.

2.3. Transformer for Multimodal Learning

Transformer has been applied to multimodal tasks as a result of its significant performance improvement in NLP and CV. Multi-Modality Cross Attention (MMCA) [28] is proposed for image and text matching, which jointly models the intra-modal and inter-modal relationships between image and sentence in a unified depth model. TransFusion [29] provides a robust solution for LiDAR-camera fusion with a soft-association mechanism to handle inferior image conditions. Botach et al. [30] propose an architecture of multimodal tracking transformer, which models referring video object segmentation task as sequence prediction problems. A token-based multi-task decoder [31] for RGB-D salient object detection method is developed by introducing task-related tokens and a novel patch-task-attention mechanism. Li et al. [32] propose a Transformer-based RGB-D egocentric action recognition framework and model the temporal structure of the data from different modalities with self-attention. Xiao et al. [33]

7

Figure 2: Overview of our multispectral object detection framework. (The upper and bottom branch are the RGB and thermal feature extraction module, C1∼C5 represent different scales of feature maps, DMFF module is our proposed feature fusion method, Neck module is the multiscale feature aggregation network, and Head module outputs the final detection results.)

have designed five attribute-specifc fusion branches to integrate RGB and thermal features under various challenges of RGB-T tracking, and strengthened the aggregated feature and modality-specifc features with an enhancement fusion transformer. These studies have proved that Transformer is effective in various multimodal tasks. In this paper, we introduce Transformer into multispectral object detection that aims to better harvest the complementary information between RGB and thermal modality from a global perspective.

3. The proposed method

3.1. Architecture

As illustrated in Fig. 2, the proposed method is a dual-branch backbone network, which is tailored for feature extraction from RGB-thermal image pairs. Our method mainly comprises of three stages: the mono-modal feature extraction, the dual-modal feature fusion and the detection neck and head. The mono-modal feature extraction is firstly used for RGB and thermal images independently, which can be formulated in Eq. 1.

$$\boldsymbol{F}_R^i = \Psi_{backbone}(\boldsymbol{I}_R; \boldsymbol{\theta}_R), \boldsymbol{F}_T^i = \Psi_{backbone}(\boldsymbol{I}_T; \boldsymbol{\theta}_T) \tag{1}$$

where $\boldsymbol{F}_R^i$, $\boldsymbol{F}_T^i \in \mathbb{R}^{W \times H \times C}$ denote the feature maps from the i-th layer (i=3, 4, 5) of the RGB and thermal branch respectively. H, W and C denote height,

8

width and channel number of feature maps. $\boldsymbol{I}_R$, $\boldsymbol{I}_T \in \mathbb{R}^{W \times H \times C}$ represent the input RGB and thermal images, $\Psi_{backbone}$ denotes the feature extraction function with parameters $\boldsymbol{\theta_R}$ and $\boldsymbol{\theta_T}$ for RGB and thermal branch respectively. In generic object detection, VGG16 [34], ResNet [35] and CSPDarkNet [36] are commonly used as function $\Psi_{backbone}$. In the feature extraction stage, the multi-scale features are commonly utilized to capture objects with different size.

Secondly, given feature maps of $\boldsymbol{F}_R^i$ and $\boldsymbol{F}_T^i$, cross-modal feature fusion is required to aggregate features from different branches in multispectral object detection, which can be defined in Eq. 2.

$$\boldsymbol{F}_{R+T}^i = \Phi_{fusion}(\boldsymbol{F}_R^i; \boldsymbol{F}_T^i; \boldsymbol{\theta_f}) \tag{2}$$

where $\boldsymbol{F}_{R+T}^i \in \mathbb{R}^{W \times H \times C}$ denotes the fused features in the i-th layer. $\Phi_{fusion}(\cdot)$ denotes the feature fusion function with parameter $\boldsymbol{\theta_f}$. Given that previous researches [5, 19] have explored different fusion architectures and validated that halfway fusion outperforms the other fusion methods, we use halfway fusion as the default setting and fuse the multimodal features from convolution layer C3∼C5 as shown in Fig. 2. In general, addition operation or NIN fusion [17] [1] is commonly used as feature fusion function $\Phi_{fusion}(\cdot)$. In this paper, a dual cross-attention feature fusion transformer is proposed to model $\Phi_{fusion}(\cdot)$, which will be described in section 3.2.

Finally, the feature maps from $\{\boldsymbol{F}_{R+T}^i\}_{i=1}^L$ are fed to the detector neck for multi-scale feature fusion, and then delivered to the detector head for subsequent classification and regression which is formulated in Eq. 3.

$$[\boldsymbol{D}_{cls}, \boldsymbol{D}_{bbox}] = \phi_{head}(\phi_{neck}(\{\boldsymbol{F}_{R+T}^i\}_{i=1}^L); \boldsymbol{\theta_h}) \tag{3}$$

where ϕ_{neck} and ϕ_{head} represent the multi-scale feature aggregation and detection head function. FPN [37] and PANet [38] is commonly utilized as function ϕ_{neck} to enhance the semantic expression and localization ability of features,

---

[1] $\Phi_{fusion} = conv_{1 \times 1}([\boldsymbol{F}_R, \boldsymbol{F}_T])$, where $conv_{1 \times 1}$ is a 1×1 convolution, $[\cdot]$ denotes the concatenate operation.

Figure 3: Illustration of the proposed DMFF module.(In the upper row, the proposed DMFF module comprised of Spatial Feature Shrinking (SFS) module, Iterative Cross-modal Feature Enhancement (ICFE) module and the bimodal feature fusion module with NIN fusion. The SFS module compresses the size of the feature map for subsequent CFE module, the ICFE module refines the cross-modal features by dual CFE modules, and the bimodal feature fusion module conducts the local feature fusion from the output of ICFE module. The bottom row illustrates the details of the CFE module for the enhancement of thermal modality.)

while ϕ_{head} acts as a role of classification and bounding box regression with parameter $\boldsymbol{\theta_h}$, such as detection head of YOLO [36] and FCOS [39]. For a fair comparison, we adopt these default setting of detection necks and heads in the original paper.

3.2. Dual-modal Feature Fusion (DMFF)

Fig. 3 illustrates the structure of our Dual-modal Feature Fusion (DMFF) module, which mainly contains three components: the Spatial Feature Shrinking (SFS) module, the Iterative Cross-modal Feature Enhancement (ICFE) module, and the bimodal feature fusion module with NIN fusion. The modules will be detailed in the following sections.

3.2.1. Cross-modal Feature Enhancement (CFE)

Different from the previous studies which capture the local features of different modalities, the proposed CFE module enables mono-modal to learn more complementary information from auxiliary modality in a global perspective.

10

The proposed CFE module not only retrieves the complementary relationship between the RGB and thermal modality, but also overcome the deficiency in modeling the long-range dependencies of cross-modal features.

Given input feature maps $\boldsymbol{F}_R$ and $\boldsymbol{F}_T \in \mathbb{R}^{H \times W \times C}$[2], we first flatten each feature map into a set of tokens and add a learnable position embedding, which is a trainable parameter of dimension $HW \times C$ that encodes the spatial information between different tokens. After that, we can obtain a set of tokens $\boldsymbol{T}_R, \boldsymbol{T}_T \in \mathbb{R}^{HW \times C}$ with position embeddings as the input of CFE module. Since the RGB-thermal image pairs are usually not perfectly aligned, we employ dual CFE modules to harvest the complementary information for the enhancement of the RGB and thermal features respectively. The parameters are not shared between the two CFE modules. In Fig. 3(bottom), we only illustrate the CFE module of thermal branch for clarity, which is formulated in Eq. 4.

$$\hat{\boldsymbol{T}}_T = \mathcal{F}_{CFE-T}(\{\boldsymbol{T}_R, \boldsymbol{T}_T\}) \tag{4}$$

where $\boldsymbol{T}_R$ and $\boldsymbol{T}_T$ denotes the RGB and thermal feature tokens input to the CFE module. $\hat{\boldsymbol{T}}_T$ indicates the enhanced thermal features with the CFE module. $\mathcal{F}_{CFE-T}(\cdot)$ denotes our proposed CFE module for thermal branches.

The details of the CFE module are as follows. Firstly, the tokens of thermal modality $\boldsymbol{T}_T$ are projected to two separate matrices $\boldsymbol{V}_T, \boldsymbol{K}_T \in \mathbb{R}^{HW \times C}$ to compute a set of values and keys (Eq. 5). And then, the tokens of RGB modality $\boldsymbol{T}_R$ are projected to another separate matrix $\boldsymbol{Q}_R \in \mathbb{R}^{HW \times C}$ to compute a set of queries (Eq. 5).

$$\boldsymbol{V}_T = \boldsymbol{T}_T \boldsymbol{W}^V, \boldsymbol{K}_T = \boldsymbol{T}_T \boldsymbol{W}^K, \boldsymbol{Q}_R = \boldsymbol{T}_R \boldsymbol{W}^Q \tag{5}$$

where $\boldsymbol{W}^V$, $\boldsymbol{W}^K$ and $\boldsymbol{W}^Q \in \mathbb{R}^{C \times C}$ denote the weight matrices.

Secondly, the correlation matrix is built via dot-product operation, followed by a softmax function normalizes the correlation scores, which represents the

---

[2]The vector $\boldsymbol{F}_R$ and $\boldsymbol{F}_T$ represent the feature maps from the i-th layer of different branches, similar to that in Eq. 1. For simplicity, we remove the superscript i.

similarity between different features of RGB and thermal modality. After that, the vector $\boldsymbol{Z}_T$ is obtained by multiplying the correlation matrix with vector $\boldsymbol{V}_T$ (Eq. 6), which refines the RGB features by leveraging the similarity across modalities.

$$\boldsymbol{Z}_T = \text{softmax}(\frac{\boldsymbol{Q}_R \boldsymbol{K}_T{}^T}{\sqrt{D_K}}) \cdot \boldsymbol{V}_T \tag{6}$$

$$\boldsymbol{T}'_T = \alpha \cdot \boldsymbol{Z}_T \boldsymbol{W}^O + \beta \cdot \boldsymbol{T}_T \tag{7}$$

$$\hat{\boldsymbol{T}}_T = \gamma \cdot \boldsymbol{T}'_T + \delta \cdot \text{FFN}(\boldsymbol{T}'_T) \tag{8}$$

Besides, we also employ multi-head cross-attention mechanism with 8 parallel heads in this paper, which enables the model to jointly understand the correlation between RGB and thermal features from different perspectives.

Thirdly, the vector $\boldsymbol{Z}_T$ is reprojected back to the original space through nonlinear transformation, and added to the input sequence through a residual connection [35] (Eq. 7), where $\boldsymbol{W}^O \in \mathbb{R}^{C \times C}$ denotes a output weight matrix before FFN layer.

Finally, the feed-forward network (FFN) with two fully-connected layers as that in the standard Transformer [11] is applied to further refine the global information to improve the robustness and accuracy of the model and output the enhanced features $\hat{\boldsymbol{T}}_T$ (Eq. 8).

Inspired by [40], we apply learnable coefficients on each branch of residual connection in Eq. 7 and Eq. 8, adaptively learning the data from different branches to achieve performance gain, where $\alpha, \beta, \gamma, \delta$ are the learnable parameters initialized as 1 during training.

Similar to the thermal branch, the other CFE module is also utilized to enhance the features of RGB branch, which can be formulated in Eq. 9:

$$\hat{\boldsymbol{T}}_R = \mathcal{F}_{CFE-R}(\{\boldsymbol{T}_R, \boldsymbol{T}_T\}) \tag{9}$$

It is worth to mention that CFT [10] is also a transformer-based method, which directly concatenates the tokens of each modality and computes the correlation across modalities with a single transformer encoder. Differently, we employ two improved cross-attention transformers to compute the correlation

across modalities only with queries from auxiliary modality, which have lower computational complexity and fewer parameters. The detailed computational complexity comparison between CFT and our method is given in Tab. 1.

Table 1: The computational complexity comparison between CFT and our proposed method.(T is the length of tokens, while C is the channels of token.)

Step	CFT	Ours
QK^T	$O(4T^2 \times C)$	$O(2T^2 \times C)$
$\text{softmax}(\frac{QK^T}{\sqrt{D_K}})$	$O(4T^2)$	$O(2T^2)$
$\text{softmax}(\frac{QK^T}{\sqrt{D_K}}) \cdot V$	$O(4T^2 \times C)$	$O(2T^2 \times C)$
FFN	$O(16T \times C^2)$	$O(8T \times C^2)$
Total	$O(4T^2 \times C + 16T \times C^2)$	$O(2T^2 \times C + 16T \times C^2)$

3.2.2. Spatial Feature Shrinking (SFS)

Although the initial feature maps used in fusion are downsampled using backbone, the model's parameters and memory cost can still much surpass the operating requirements of standard processors. To lower down the subsequent computational cost of our module with less information loss in the feature maps, we apply a SFS module before the CFE module that compresses the feature maps. In this module, we attempt two different methods with convolution and pooling operations, and the details are as follows.

Convolution operation. We first design a method for dimension reduction based on the convolution operation, as shown in Eq 10. Specifically, we transform the spatial information of features to the channel dimension by reshaping the the dimensions of feature maps, and then compress the channel dimension with 1×1 convolution operation.

$$\boldsymbol{F}_{conv} = conv_{1\times1}(Reshape(\boldsymbol{F})) \tag{10}$$

where $\boldsymbol{F}$ denotes the input feature maps. $\boldsymbol{F}_{conv}$ denotes the compressed feature maps by 1×1 convolution.

13

Pooling operation. Average pooling and max pooling are two conventional pooling methods, which commonly used to reduce the spatial dimension of feature maps without additional parameters. Average pooling computes the mean of all the elements in the pooling region and retains the background information in the images, while max pooling considers the maximum element in the pooling region and mainly retains the texture features of objects. Thus, we employ a method aggregating average pooling and max pooling adaptively inspired from mixed pooling [41], as shown in Eq. 12.

$$\boldsymbol{F_a} = \text{AvgPooling}(\boldsymbol{F}, S), \ \boldsymbol{F_m} = \text{MaxPooling}(\boldsymbol{F}, S) \tag{11}$$

$$\boldsymbol{F_o} = \lambda \cdot \boldsymbol{F_a} + (1 - \lambda) \cdot \boldsymbol{F_m} \tag{12}$$

where $\boldsymbol{F}$ denotes the input feature maps. S denotes scaling factors of feature maps. $\boldsymbol{F_a}$ and $\boldsymbol{F_m}$ denote the compressed feature maps via AvgPooling$(\cdot)$ and MaxPooling$(\cdot)$ respectively. λ denotes the weight between 0 and 1, which is a learnable parameter in this paper.

Compared with the original feature maps of dimension $H \times W \times C$, the compressed feature maps have dimensions $(H \times W)/S \times C$, resulting in the dimension of tokens reduced from $HW \times C$ to $HW/S \times C$. Thus, the dimensions of keys, queries and values in the CFE module become $K, Q, V \in \mathbb{R}^{HW/S \times C}$. Finally, the total computational complexity is reduced from $O(W^2 H^2 \times C + 8WH \times C^2)$ to $O(W^2 H^2/S^2 \times C + 8WH/S \times C^2)$.

3.2.3. Iterative Cross-modal Feature Enhancement (ICFE)

For the sake of strengthening the memory of complementary information from inter-modal and intra-modal features to further improve the model performance, we introduce an iterative learning strategy based on the CFE module and dubbed as ICFE module. As illustrated in Fig. 4 (a), the traditional methods generally improve the performance by stacking serveral modules, however this strategy of dramatically expanding the depth of the model may not only increase the parameters significantly, but also lead to overfitting. Instead, our proposed

14

Figure 4: Visualization of difference between the traditional method and ours. (a) Traditional method stacks multiple blocks in series, and parameters are not shared in each block. (b) Our proposed ICFE module iteratively refines the features across modalities, and parameters are shared in each block. Block in the image denotes our proposed dual CFE modules.

iterative learning strategy deepens the depth of the network over multiple iterations with shared parameters, and progressively refines the complementary information across modalities without increasing the number of parameters, as shown in Fig. 4 (b). Taking n iterations as an example, it can be simplified as follows (Eq. 13):

$$
\begin{aligned}
\{\hat{\boldsymbol{T}}_R^n, \hat{\boldsymbol{T}}_T^n\} &= \mathcal{F}_{ICFE}(\{\boldsymbol{T}_R, \boldsymbol{T}_T\}, n) \\
&= \underbrace{\mathcal{F}_{CFE}(\cdots \mathcal{F}_{CFE}}_{n}(\{\boldsymbol{T}_R, \boldsymbol{T}_T\}))
\end{aligned}
\tag{13}
$$

where $\{\hat{\boldsymbol{T}}_R^n, \hat{\boldsymbol{T}}_T^n\}$ denotes the output sequence obtained after n iterative operations, $\{\boldsymbol{T}_R, \boldsymbol{T}_T\}$ denotes the input sequence of the ICFE module. $\mathcal{F}_{ICFE}(\cdot)$ denotes our proposed ICFE module, which integrates two CFE modules for RGB and thermal branch respectively. The output of each iterative operation is used as the input of the next iterative operation, and parameters are shared between each iterative operation. Besides, The output sequences $\hat{\boldsymbol{T}}_R^n$ and $\hat{\boldsymbol{T}}_T^n$ from ICFE module are first converted into the feature maps, and then re-calibrated to the original size of feature map with bilinear interpolation.

Figure 5: Different fusion modes of CFE module. (a) Single CFE module for RGB modality. (b) Single CFE module for thermal modality. (c) Dual CFE module (shared) for both RGB and thermal modality. (d) Dual CFE modules for both RGB and thermal modality. (e) Baseline feature fusion for both RGB and thermal modality. (f) Detection head from all these output features.

3.2.4. Fusion Modes for Detection Heads

Figure 3 shows how our proposed CFE module can work with different input modalities. We have investigated four alternative fusion modes to validate the effectiveness of the CFE module, as shown in Fig. 5. Only a single modality feature is outputted in Fig. 5 (a) and (b), forcing the CFE-R and CFE-T module to collect complimentary features from thermal and RGB image features respectively. In addition, we have also explored two different work modes with dual CFE modules, which used shared and unshared parameters, as shown in Fig. 5 (c) and (d). Fig. 5 (e) illustrates the baseline feature fusion method with NIN fusion method. Finally, all these fused feature maps $(F_i, F_i^{'}, i = \{R, T, fused\})$ will be fed into the detection head as shown in Fig. 5 (f). It's noteworthy to see that our method naturally favors both dual and single image modalities. Thanks to the cross-attention mechanism, even if one of the input modalities is missing or the image quality is poor, our method can still produce satisfactory results. The detailed experimental study which can support our assertion will be postponed in Section 4.3.

16

4. Experiments

4.1. Datasets and Evaluation metrics

KAIST Dataset. KAIST [12] is a popular multispectral pedestrian detection benchmark, which involves scenes with different illuminations. There are 8,963 and 2,252 weakly-aligned image pairs with the resolution of 640×512 for training [15] and testing respectively. The performance evaluation on KAIST dataset is typically in accordance with the metric log-average miss rate [42]. For more accurate annotations, we use the sanitized annotations for training [20] and testing [5].

FLIR Dataset. FLIR [13] is a challenging multispectral object detection dataset including daytime and night scenes. There are 5,142 aligned multispectral image pairs, of which 4,129 are used for training and 1,013 for testing. It contains three classes of objects, namely "person", "car" and "bicycle". Since the images are misaligned in the original dataset, we select the FLIR-aligned version [16] for comparisons in our experiment.

VEDAI Dataset. VEDAI [14] is a public dataset for small target detection in aerial imagery, which contains more than 3,700 annotated targets in 1,268 RGB-infrared image pairs. There are 9 vehicle categories in this dataset. We use the images with size of 1024×1024 for training and testing, and convert the annotations to the horizontal-box format with [43] as reference since the original version is annotated as a rotating box with four-corner coordinates.

Log-average Miss Rate. The log-average miss rate (MR^{-2}) [42] is used for the evaluation on KAIST dataset. It represents the average miss rate under 9 FPPI values, which are sampled uniformly in the logarithmic interval $[10^{-2},1]$. The lower values of MR^{-2}, the better performance.

Average Precision. Average Precision (AP) is a common evaluation metric for object detection. The positive and negative samples should be divided according to the the correctness of classification and Intersection over Union (IoU) threshold. Usually, 0.5 is used as the IoU threshold. In general, mean

17

Average Precision (mAP) represents the average of AP under all categories. Different from MR^{-2}, the higher values of AP and mAP, the better performance.

4.2. Implementation Details

Our method is implemented using PyTorch 1.7.1 framework on a Ubuntu 18.04 server with CPU i7-9700, 64G Memory and Nvidia RTX 3090 24G GPU. The training phase takes 60 epochs with the batch size of 8. The SGD optimizer is used with the initial learning rate of 1.0×10^{-2} and the momentum of 0.937. In addition, the weight decay factor is 0.0005 and the learning rate decay method is cosine annealing. The input size of images are 640×640 for training and 640×512 for testing. Besides, mosaic and random flipping is used for data augmentation. The loss function is utilized following the detectors of YOLOv5 and FCOS in the orignal paper. In the ablation studies, we use YOLOv5 with NIN fusion [17] as default baseline for comparison.

4.3. Ablation Study

4.3.1. Effects of learnable parameters applied on residual connection

Given that Shen et al. [40] has proved learning parameters applied on both branches is slightly better than that on the single branch, we evaluate the effectiveness of learnable parameters applied on both branches of residual connection in our proposed CFE module. The experimental results are given in Table. 2. Compared with the CFE module without learnable parameters, adding learnable parameters on both branches reduces MR from 7.86% to 7.63% on the KAIST dataset, and improves mAP50 from 77.1% to 77.5% on the FLIR dataset. As a result, the learnable parameters applied on both branches of residual connection are effective to achieve performance gain without significantly increasing the computational cost in our CFE module.

4.3.2. Effects of CFE module for mono-modality and dual-modality

The experimental results of CFE module for mono-modality on both KAIST and FLIR datasets are presented in Tab. 3, which is separated into three groups

18

Table 2: Effects of learnable parameters in CFE module on both KAIST and FLIR datasets. (LP denotes learning parameters applied on both branches.)

Method	LP	KAIST	FLIR	Params(M)
		MR(%)↓	mAP50(%)↑	
Baseline+CFE		7.86	77.1	120.2
	✓	7.63(-0.23)	77.5(+0.4)	120.2

acoording to the output modality. In the first group (first row), we apply the CFE module to enhance RGB features leveraging the compensatory information from the thermal images and only output the enhanced RGB features for the subsequent detection, as shown in Fig. 5(a). Our dual branch method with CFE module outperforms the RGB-only single branch method by 0.65% and 0.90% on the KAIST and FLIR dataset respectively. Similarly, in the second group (second row), the enhanced thermal feature with CFE module in Fig. 5(b) achieves a gain of 0.59% and 1.20% over the thermal-only detector on the KAIST and FLIR dataset respectively. The top two rows in Tab. 3 indicate that the quality of RGB features on the KAIST dataset is superior to that of thermal features for detection, whereas thermal features on the FLIR dataset are superior to RGB features in quality. This might be caused by the properties of dataset, the camera model and other elements. Thus, the dual CFE modules are applied to RGB and thermal branch to collect the complementary information from each other, and fused features from enhanced RGB and thermal modality are utilized for the subsequent detection in the last group (last row), which outperforms the baseline method by 0.70% and 1.00% on the KAIST and FLIR dataset respectively. As a result, the experimental findings presented above demonstrate the effectiveness of our proposed CFE module, which favors both RGB and thermal-based global feature fusion.

4.3.3. Effects of the number of stacked modules

In this section, we provide the mAP values for different number of stacked CFE modules on FLIR dataset. Tab. 4 shows that as the number of stacked

Table 3: Effects of CFE module for each modality on both KAIST and FLIR datasets.(The lower the MR, the better. The higher mAP, the better performance. In the third column, the letter (a)∼(f) denote the fusion mode in Fig. 5)

Input Modality	Output Feature	Method	KAIST	FLIR
			MR(%)↓	mAP50(%)↑
RGB	F_R	Baseline-RGB (f)	18.39	67.8
RGB+Thermal	F_R	Baseline+CFE (a)	17.74(-0.65)	68.7(+0.9)
Thermal	F_T	Baseline-Thermal (f)	18.94	73.9
RGB+Thermal	F_T	Baseline+CFE (b)	18.35(-0.59)	75.1(+1.2)
RGB+Thermal	F_{fused}	Baseline (e)	8.33	76.5
RGB+Thermal	F_{fused}	Baseline+CFE (c)	10.78(+2.45)	76.0(-0.5)
RGB+Thermal	F_{fused}	Baseline+CFE (d)	7.63(-0.70)	77.5(+1.0)

Table 4: Comparison with different number of stacked modules on the FLIR dataset.

Number	mAP50(%)↑	Mem(M)	Params(M)	FPS(Hz)
1	77.5	528.5	120.2	40.5
2	77.6	686.0	164.3	31.9
4	77.4	1037.5	252.5	26.9
6	77.8	1795.5	340.7	22.9
8	77.9	1747.0	428.9	19.6
10	78.2	2100.5	517.1	17.3

modules increases to 10, the parameter numbers and GPU memory increases by more than 4× times, while the running speed decreases dramatically from 40.5 Hz to 17.3 Hz with a marginal benefit of 0.70% in terms of mAP. The previous studies [44, 45] find that the attention maps across adjacent layers of vision transformer exhibit very high similarity. As shown in Fig. 6(right), after visualizing the feature maps of different stacked numbers, we also find this phenomenon in our experiment. So, we think that the high similarity of feature maps can result in marginal performance improvement. As a result, stacking blocks in series is not an efficient solution for feature fusion.

4.3.4. Effects of different number of iterations

The experimental results of varying number of iterations on the KAIST and FLIR datasets are shown in Tab. 5. With only one iteration, the iterative

Figure 6: Visualization results of CFE module with different stacking number and ICFE module with different iteration number on the FLIR dataset. The top row is an image pair at daytime, and the bottom row is an example at night. The 1st column is the input image; the $2^{nd} \sim 5^{th}$ columns are feature maps from iterative learning; the $6^{th} \sim 9^{th}$ columns are feature maps from different numbers of stacking.

learning method reduces MR from 7.63% to 7.17% on the KAIST dataset and improves mAP50 from 77.50% to 79.20% on the FLIR dataset. Interestingly, we find that extra iterations do not boost performance, and one iteration achieves the best results in our experiment. As shown in Fig. 6(middle), we also visualize the feature maps of ICFE modules. As the number of iterations rises, we find that the interaction between different modality features results in negative effects and the background information are gradually enhanced. We think that the enhanced interference of background noises may lead to performance degradation. Furthermore, since the iterative learning technique uses shared parameters, more iterations will not incur extra parameters or memory costs.

Comparing the experimental results in Tab. 4 and Tab. 5, we can infer that interative learning is more effective for cross-modal feature interaction and also outperforms the stacking method. Besides, our method has a much faster inference speed of 36.7 FPS than the stacking method.

4.3.5. Effects of different spatial feature shrinking methods

We evaluate multiple existing approaches in order to find a reliable down-sampling method with less loss of feature information, and the experiment re-

21

Table 5: Comparison with different number of iterations.

Number	KAIST MR(%)	FLIR mAP50(%)	Params(M)	Mem(M)	FPS(Hz)
0	7.63	77.5	120.2	528.5	40.5
1	**7.17**	**79.2**	120.2	528.5	36.7
2	7.87	76.9	120.2	528.5	32.8
3	7.87	77.5	120.2	528.5	29.6

sults are shown in Tab. 6. In comparison to the other down-sampling approaches, mixed pooling produces best results, with MR of 7.17% and mAP50 of 79.20% on the KAIST and FLIR datasets, respectively. As a result, we use mixed pooling to compress the feature maps and reduce computational complexity in this paper.

Table 6: Comparison with different spatial feature shrinking methods.

Methods	KAIST MR(%)	FLIR mAP50(%)
Average Pooling	7.58	77.0
Max Pooling	7.94	78.4
Ours-Conv	7.42	78.6
Ours-Pool	7.17	79.2

4.3.6. Discussion on Different Input Modalities

In this section, we have also conducted four groups of experiments in order to validate the effectiveness of using different input modalities for CFE, and the experimental results are shown in Tab. 7. The first group (row $1 \sim 2$) demonstrates the experimental result in the KAIST and FLIR datasets for the YOLOv5 detector with a single input image modality (RGB or thermal). In the second group (row $3 \sim 5$), we provide the results of the YOLOv5+NIN method with one or two input image modalities (RGB or RGB+Thermal). It is clear to observe that the performance of the YOLOv5+NIN method with two different input modalities (row 3) is superior to the YOLOv5 method (row 1 and 2) by a large margin. However, the YOLOv5+NIN method will bring a large

performance degradation when using the same two modalities (row 4 ~ 5) as input. Furthermore, we also conducts experiments with our proposed method (YOLOv5+ICFE) in the third group (row 6 ~ 9). It is surprisingly found that using the same two modality images can still achieve competitive results comparing to the those with both RGB and thermal images. It indicates that our method can provide discriminative mono-modal features for the subsequent detection stage with a slight performance degradation as shown in row 7 and 9 on both KAIST and FLIR datasets. This is beneficial to the scenarios where one of the input modalities is missing or the image quality is poor. In the last group (row 10 ~ 12), we also have observed significant drop in the detection performance of our proposed method (YOLOv5+ICFE+NIN) due to the append of the NIN module to the output of the dual CFE modules. The observation from both row 4 ~ 5 and 11 ~ 12 indicates that NIN is harmful to both the YOLOv5+NIN and our proposed method when only one input modality is accessible.

Table 7: Comparison with different input modalities. (R denotes RGB, T denotes Thermal. R+T represents the input with dual modalities, while R+R or T+T denotes input with single modality and ignores the other modality. In the third column, the letter (a)~(f) denote the fusion mode in Fig. 5)

Number	Methods	Input	Output	KAIST MR(%)↓	FLIR mAP50(%)↑
1	YOLOv5	R (f)	F_R	18.39	67.8
2		T (f)	F_T	18.94	73.9
3	YOLOv5+NIN	R+T (e)	F_{fused}	8.33	76.5
4		R+R (e)	F_{fused}	43.79(+35.46)	57.1(-19.4)
5		T+T (e)	F_{fused}	43.79(+35.46)	65.3(-11.2)
6	YOLOv5+ICFE	R+T (a)	F_R	17.74	68.7
7		R+R (a)	F_R	20.50(+2.76)	66.3(-2.4)
8		R+T (b)	F_T	18.35	75.1
9		T+T (b)	F_T	20.07(+1.72)	74.2(-0.9)
10	YOLOv5+ICFE+NIN	R+T (d)	F_{fused}	7.17	79.2
11		R+R (d)	F_{fused}	31.23(+23.06)	57.8(-21.4)
12		T+T (d)	F_{fused}	37.42(+29.79)	66.0(-13.2)

4.3.7. Comparisons with different backbones and heads

In order to evaluate the effectiveness and generality of our proposed DMFF module, we first conduct the experiments on YOLOv5 detector with three different backbones: VGG16, ResNet50 and CSPDarkNet53. As shown in Tab. 8, the results on the KAIST dataset show that our approach outperforms the baseline method by 0.66%, 0.97% and 1.16% on VGG16, ResNet50 and CSP-Darknet53 respectively. The results on the FLIR dataset demonstrate that our approach also achieves a gain of 0.50%, 1.50% and 2.70% over the baseline method on VGG16, ResNet50 and CSPDarknet53 respectively. As a result, we conclude that our proposed DMFF module is applicable to various backbones and is effective under different evaluation metrics.

We also evaluate on the FCOS detector to further examine the effectiveness and generality of our proposed DMFF module. The experimental results are given in Tab. 8. Compared with the baseline method, FCOS with DMFF module reduces MR from 14.03% to 12.96% with a gain of 1.07% on the KAIST dataset, and improves mAP50 from 69.80% to 71.70% with a gain of 1.90% on the FLIR dataset. The above results indicate that our proposed DMFF module works wells for both anchor-base and anchor-free detectors. Finally, it is clear to find that YOLOv5 detector with CSPDarknet53 backbone achieves the best performance with a fair number of parameters when compared to the other backbones and detectors.

Table 8: Comparison with different detectors and backbones.

Detector	Type	Backbone	Method	KAIST MR(%)↓	FLIR mAP50(%)↑	Params(M)	FPS(Hz)
YOLOv5	Anchor-based	VGG16	Baseline	16.12	69.3	42.67	72.67
			Ours	15.46(-0.66)	69.8(+0.5)	62.17	54.48
		ResNet50	Baseline	13.65	70.5	136.13	31.38
			Ours	12.68(-0.97)	72.0(+1.5)	313.76	31.44
		CSPDarkNet53	Baseline	8.33	76.5	75.44	50.00
			Ours	7.17(-1.16)	79.2(+2.7)	120.21	38.46
FCOS	Anchor-free	ResNet50	Baseline	14.03	69.8	58.86	19.21
			Ours	12.96(-1.07)	71.7(+1.9)	65.24	17.32

Table 9: Comparsion on the KAIST dataset. (Bold numbers represent the best result in each column. Methods with suffix † and suffix ‡ use ResNet50 and CSPDarkNet53 backbone respectively, while the others use VGG16 as defaults.)

Method	Miss Rate(%)↓			FPS(Hz)	Platform
	All	Day	Night		
ACF+T+THOG [12]	47.32	42.65	56.18	-	-
AR-CNN [20]	10.22	10.80	9.02	8.33	TITAN X
CIANet [6]	14.13	14.78	11.14	16.67	GTX 1080Ti
FusionRPN+BF [46]	18.29	19.57	16.27	-	-
HalfwayFusion [5]	25.77	24.91	26.67	2.33	TITAN X
IAF-RCNN [19]	15.57	14.81	16.70	4.76	TITAN X
IATDNN-IAMSS [47]	14.46	14.18	15.28	4.00	TITAN X
MBNet [15]†	8.40	8.62	8.27	14.29	GTX 1080Ti
MLPD [48]	7.58	7.96	6.95	-	-
MSDS-RCNN [17]	8.23	8.83	**6.75**	4.55	GTX 1080Ti
Ours‡	**7.17**	**6.82**	7.85	**38.46**	RTX 3090

4.4. Comparison with State-of-the-art Methods

KAIST Dataset. Tab. 9 shows the comparison of our method with existing methods on the KAIST dataset. It can be observed that our approach surpasses most of the state-of-the-art methods in the settings of reasonable, and obtains the lowest miss rate under daytime subset. Furthermore, Tab. 9 also illustrates that our approach runs at 38.46 Hz on the RTX 3090 platform. As a result, our method is beneficial for scenarios of object detection where high detection speed is required.

FLIR Dataset. Tab. 10 shows the comparison of our method with existing methods on the FLIR dataset. It is clear to observe that our approach outperforms all the existing methods and achieves state-of-the-art performance. Specifically, our method achieves 79.20%, 36.9% and 41.4% in terms of mAP50, mAP75 and mAP metrics. Furthermore, our method achieves 66.90%, 89.00% and 81.60% for the categories of Bicycle, Car and Person respectively. Besides, we also employ the CFT baseline with our proposed modules (Ours*) for a fair comparison. It is clear to see that our method outperforms Ours* in terms of all mAP50, mAP75 and mAP metrics.

Table 10: Comparison on the FLIR Dataset.

Method	AP50(%)			mAP50(%)	mAP75(%)	mAP(%)
	Bicycle	Car	Person			
MMTOD-CG [49]	50.26	70.63	63.31	61.4	-	-
MMTOD-UNIT [49]	49.43	70.72	64.47	61.5	-	-
GAFF [7]	-	-	-	72.9	30.9	37.3
CFR [16]	57.77	84.91	74.49	72.4	-	-
BU-ATT [50]†	56.10	87.00	76.10	73.1	-	-
BU-LTT [50]†	57.40	86.50	75.60	73.2	-	-
CFT [10]‡	61.40	89.50	**84.10**	78.3	35.5	40.2
Ours*‡	65.20	**90.20**	80.40	78.6	35.8	40.8
Ours‡	**66.90**	89.00	81.60	**79.2**	**36.9**	**41.4**

Table 11: Comparison on the VEDAI Dataset.

Method	mAP50(%)	mAP(%)
Input Fusion [43]‡	74.40	45.65
Mid Fusion [43]‡	74.80	**46.30**
SuperYOLO [51]‡	73.61	-
Ours(baseline)‡	74.66	44.09
Ours‡	**76.62**	44.93

VEDAI Dataset. The experimental comparisons on the VEDAI dataset are given in Tab. 11. Although we do not apply any tricks for small object detection, our method still outperforms the baseline method by 1.96% over mAP, and achieves competitive results with a mAP of 76.62% among the existing methods. However, under the more strict evaluation metrics mAP, our method is 0.28%, 1.37% lower comparing to the Input Fusion and Mid Fusion method respectively.

4.5. Qualitative Analysis

Fig. 7 illustrate sample visualization results of attention maps during daytime and nighttime on the KAIST and FLIR datasets. As shown in Fig. 7(a), it's difficult to detect the pedestrians under poor illumination conditions in the RGB images with naked eyes, however our method can still identify and locate the objects by aggregating RGB and thermal images. Besides, the complex ur-

ban traffic scenes mixed with pedestrians and vehicles bring great challenges, whereas our method is able to distinguish between different categories of objects with the assistance of auxiliary modality. Fig. 7(b) illustrate that the baseline method shows interest in different areas of the input images, resulting in more false positives. However, our method can exploit the global spatial location information and the correlation between different objects to capture highly-discriminative features, as shown in Fig. 7(c).

<div align="center">(a) Ground Truth   (b) Baseline   (c) Ours</div>

Figure 7: Visualization results of attention maps on KAIST and FLIR dataset. From left to right column: ground truth in RGB and thermal images, heatmaps of NIN fusion [17] method (Baseline), and our proposed method.

4.6. Limitations

In this section, we have provided some the failure cases and analyzed the limitations of our proposed method. Fig. 8 (a) illustrates that our model misidentifies the traffic signs or trees as person in some scenarios. In our opinion, the main reasons for the false positives are the visual appearance similarity to the traffic signs or trees, and low image quality of the KAIST dataset. In

Fig. 8 (b), the occlusion between two overlapping pedestrians can also leads to false negatives on the FLIR dataset. Furthermore, Fig. 8 (c) shows that our model may misidentify some devices mounted on rooftops as cars on the VEDAI dataset because they have similar shapes and colors when viewed from a bird's-eye perspective.

Figure 8: Failure cases on the KAIST, FLIR and VEDAI datasets. From left to right columns are failure cases on the KAIST dataset (a), FLIR dataset (b) and VEDAI dataset (c). The red triangles indicate the false positives or false negatives in the images. Zoom in for more details.

5. Conclusions

In this paper, we proposed a novel cross-modal feature fusion framework for multispectral object detection, which addresses the issue that exsiting methods mainly focus on local feature correlations between different modalities. More particularly, the cross-modal feature enhancement module is proposed to enhance the feature representation of mono-modality by leveraging the global information from complementary modality. Furthermore, we introduce iterative learning strategy to refine the complementary information, which improves the model performance without adding extra parameters. In terms of detection accuracy and running speed, our proposed method outperforms the other state-of-the-art methods on the KAIST, FLIR and VEDAI datasets. In the future, we intent to further explore the efficient and lightweight cross-modal feature fusion framework, and extent our approach to other multimodal tasks.

28

Acknowledgement

References

References

[1] B. Bosquet, D. Cores, L. Seidenari, V. M. Brea, M. Mucientes, A. D. Bimbo, A full data augmentation pipeline for small object detection based on generative adversarial networks, Pattern Recognition 133 (2023) 108998.

[2] W. Liu, I. Hasan, S. Liao, Center and scale prediction: Anchor-free approach for pedestrian and face detection, Pattern Recognition 135 (2023) 109071.

[3] G. Cheng, J. Han, P. Zhou, D. Xu, Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection, IEEE Transactions on Image Processing 28 (1) (2018) 265–278.

[4] X. Xie, G. Cheng, J. Wang, X. Yao, J. Han, Oriented r-cnn for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3520–3529.

[5] J. Liu, S. Zhang, S. Wang, D. N. Metaxas, Multispectral deep neural networks for pedestrian detection, in: 27th British Machine Vision Conference, BMVC 2016, 2016.

[6] L. Zhang, Z. Liu, S. Zhang, X. Yang, H. Qiao, K. Huang, A. Hussain, Cross-modality interactive attention network for multispectral pedestrian detection, Information Fusion 50 (2019) 20–29.

[7] H. Zhang, E. Fromont, S. Lefèvre, B. Avignon, Guided attentive feature fusion for multispectral pedestrian detection, in: Proceedings of the

IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 72–80.

[8] L. Fu, W.-b. Gu, Y.-b. Ai, W. Li, D. Wang, Adaptive spatial pixel-level feature fusion network for multispectral pedestrian detection, Infrared Physics & Technology 116 (2021) 103770.

[9] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Advances in neural information processing systems 32.

[10] F. Qingyun, H. Dapeng, W. Zhaokui, Cross-modality fusion transformer for multispectral object detection, arXiv preprint arXiv:2111.00273.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.

[12] S. Hwang, J. Park, N. Kim, Y. Choi, I. So Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1037–1045.

[13] F. A. Team, [EB/OL], https://www.flir.cn/oem/adas/adas-dataset-form/ Accessed July 6, 2021.

[14] S. Razakarivony, F. Jurie, Vehicle detection in aerial imagery: A small target detection benchmark, Journal of Visual Communication and Image Representation 34 (2016) 187–203.

[15] K. Zhou, L. Chen, X. Cao, Improving multispectral pedestrian detection by addressing modality imbalance problems, in: European Conference on Computer Vision, Springer, 2020, pp. 787–803.

[16] H. Zhang, E. Fromont, S. Lefevre, B. Avignon, Multispectral fusion for object detection with cyclic fuse-and-refine blocks, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020, pp. 276–280.

[17] C. Li, D. Song, R. Tong, M. Tang, Multispectral pedestrian detection via simultaneous detection and segmentation, in: British Machine Vision Conference (BMVC), 2018.

[18] J. Shen, Y. Liu, Y. Chen, X. Zuo, J. Li, W. Yang, Mask-guided explicit feature modulation for multispectral pedestrian detection, Computers and Electrical Engineering 103 (2022) 108385.

[19] C. Li, D. Song, R. Tong, M. Tang, Illumination-aware faster r-cnn for robust multispectral pedestrian detection, Pattern Recognition 85 (2019) 161–171.

[20] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, Z. Liu, Weakly aligned cross-modal learning for multispectral pedestrian detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5127–5137.

[21] J. U. Kim, S. Park, Y. M. Ro, Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection, IEEE Transactions on Circuits and Systems for Video Technology 32 (3) (2021) 1510–1523.

[22] X. Zuo, Z. Wang, Y. Liu, J. Shen, H. Wang, Lgadet: Light-weight anchor-free multispectral pedestrian detection with mixed local and global attention, Neural Processing Letters (2022) 1–18.

[23] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[24] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 510–519.

[25] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[26] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11534–11542.

[27] G. Cheng, P. Lai, D. Gao, J. Han, Class attention network for image recognition, Science China Information Sciences 66 (3) (2023) 132105.

[28] X. Wei, T. Zhang, Y. Li, Y. Zhang, F. Wu, Multi-modality cross attention network for image and sentence matching, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10941–10950.

[29] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, C.-L. Tai, Transfusion: Robust lidar-camera fusion for 3d object detection with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1090–1099.

[30] A. Botach, E. Zheltonozhskii, C. Baskin, End-to-end referring video object segmentation with multimodal transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4985–4995.

[31] N. Liu, N. Zhang, K. Wan, L. Shao, J. Han, Visual saliency transformer, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 4722–4732.

[32] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, W. Li, Trear: Transformer-based rgb-d egocentric action recognition, IEEE Transactions on Cognitive and Developmental Systems 14 (1) (2021) 246–252.

[33] Y. Xiao, M. Yang, C. Li, L. Liu, J. Tang, Attribute-based progressive fusion network for rgbt tracking, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2831–2838.

[34] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[36] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, Yolov4: Optimal speed and accuracy of object detection, arXiv preprint arXiv:2004.10934.

[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[38] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8759–8768.

[39] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9627–9636.

[40] Z. Shen, Z. Liu, E. Xing, Sliced recursive transformer, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV, Springer, 2022, pp. 727–744.

[41] D. Yu, H. Wang, P. Chen, Z. Wei, Mixed pooling for convolutional neural networks, in: Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24-26, 2014, Proceedings 9, Springer, 2014, pp. 364–375.

[42] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, IEEE transactions on pattern analysis and machine intelligence 34 (4) (2011) 743–761.

[43] F. Qingyun, W. Zhaokui, Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery, Pattern Recognition (2022) 108786.

[44] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks?, Advances in Neural Information Processing Systems 34 (2021) 12116–12128.

[45] S. Venkataramanan, A. Ghodrati, Y. M. Asano, F. Porikli, A. Habibian, Skip-attention: Improving vision transformers by paying less attention, arXiv preprint arXiv:2301.02240.

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

[47] D. Guan, Y. Cao, J. Yang, Y. Cao, M. Y. Yang, Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection, Information Fusion 50 (2019) 148–157.

[48] J. Kim, H. Kim, T. Kim, N. Kim, Y. Choi, Mlpd: Multi-label pedestrian detector in multispectral domain, IEEE Robotics and Automation Letters 6 (4) (2021) 7846–7853.

[49] C. Devaguptapu, N. Akolekar, M. M Sharma, V. N Balasubramanian, Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[50] M. Kieu, A. D. Bagdanov, M. Bertini, Bottom-up and layerwise domain adaptation for pedestrian detection in thermal images, ACM Transactions

on Multimedia Computing, Communications, and Applications (TOMM) 17 (1) (2021) 1–19.

[51] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, Q. Du, Superyolo: Super resolution assisted object detection in multimodal remote sensing imagery, IEEE Transactions on Geoscience and Remote Sensing 61 (2023) 1–15.