A Fourier-Transform-Based Framework with Asymptotic Attention for Mobile Thermal InfraRed Object Detection

Zeyu Wang, Haibin Shen, Wenyu Jiang, Kejie Huang

May 8, 2024

{wangzeyu2020, shen_hb, huangkejie}@zju.edu.cn, wjiang@i2r.a-star.edu.sg

Abstract

Thermal InfraRed (TIR) technology has emerged as a significant tool in autonomous driving systems. Unlike natural images, TIR images are distinguished by their enriched thermal and illumination information while lacking chromatic contrast. Traditional object detection on natural images normally use deep neural networks based on convolutional layers or attention modules. However, TIR-based object detection necessitates high computational efficiency to eliminate the extraction of redundant chromatic features. Furthermore, the robust space-frequency perception and expansive receptive field are critical due to the distinct brightness and contour features of TIR images. In this paper, we propose a novel network, namely Lightweight-Fourier-Transform Detector (*LFTDet*), meticulously designed to strike a balance between computational efficiency and accuracy in TIR object detection. Specifically, our innovative Fourier Transform-Efficient Layer Aggregation Network (FT-ELAN) backbone takes advantage of Fourier Transform (FT) in synergy with deep neural networks. In addition, we propose the detection neck called Asymptotic Attention-based Feature Pyramid Network (AA-FPN), which integrates the SimA mechanism in the asymptotic structure to facilitate the FT-based operation. Extensive experiments conducted on FLIR and LLVIP datasets demonstrate that *LFTDet* surpasses all baselines while maintaining an extremely low computational cost. Code is available at https://github.com/zeyuwang-zju/LFTDet.

1 Introduction

Object detection is the process of precisely locating and identifying objects within images or videos. Current object detection methods have predominantly relied on Red-Green-Blue (RGB) images as their primary input source, capitalizing on the advancements in deep-learning technology [1–13]. Nevertheless,

conventional RGB images usually fall short in providing high-quality visual information for precise object detection under challenging environmental conditions, such as low-light settings, dense fog, or snowy landscapes. In such conditions, Thermal InfraRed (TIR) technology [14] emerges as an alternative solution, which is capable of capturing thermal radiation across the wavelength spectrum of 0.75–15μm. Significantly, TIR technology can detect objects with temperatures above absolute zero (-273.15°C), which is insensitive to illumination variations. Unlike visible RGB images, which are characterized by high chromatic contrast and visual realism, TIR images typically exhibit rich thermal and illumination information, accompanied by distinct object edge contours. Consequently, TIR technology has demonstrated its superiority across various domains, including driver-assistance systems [15, 16], remote sensing [17–19], person re-identification [20–22], intelligent detection [23–25], and an extensive array of applications under low-light or complex conditions [26–29].

Current research on TIR-based object detection typically involves the use of multi-spectral RGB-TIR image pairs as the input source. Recently, we have argued that multi-spectral object detection lacks robustness in modern driving systems under complex environmental conditions and proposed the prior T2V (Thermal-To-Visible) method for mono-modal TIR-based object detection [30]. Nonetheless, an inherent limitation remains: the T2V translation model still relies on RGB-TIR image pairs during the pre-training phase, which can restrict the scalability under challenging scenes.

On the other hand, applying conventional deep-learning methods to TIR images for mono-modal object detection presents several noteworthy challenges: **(1) Redundancy in Information Extraction**: Traditional deep convolutional or fully connected layers may inadvertently extract redundant features due to the limited chromatic information in TIR images, hindering the efficiency of information extraction. **(2) Computational Overhead**: Incorporating an excessive number of layers in detection networks can lead to a surge in computational cost and parameters, posing practical challenges, particularly in modern driving systems where real-time processing and resource efficiency are critical. **(3) Infrared Feature Prerequisites**: TIR images display prominent brightness and contour features, demanding object detectors with robust perception capability and global receptive fields to effectively capture these distinctive features.

In this paper, we address the above challenges by proposing a novel lightweight model called *LFTDet*, specifically designed for object detection in TIR images. To overcome the challenges of information redundancy and computational overhead, *LFTDet* adopts our FT-ELAN as the backbone, which leverages Fourier Transform (FT) in conjunction with the deep neural network, enabling robust spatial-frequency perception and global receptive field. To enhance the operation of the FT-based detector, we introduce a novel detection neck called Asymptotic Attention-based Feature Pyramid Network (AA-FPN). We conduct extensive experiments on two benchmark datasets, namely FLIR [31, 32] and LLVIP [33], for TIR-based object detection. The experimental results demonstrate that our *LFTDet* outperforms conventional object detectors while main-

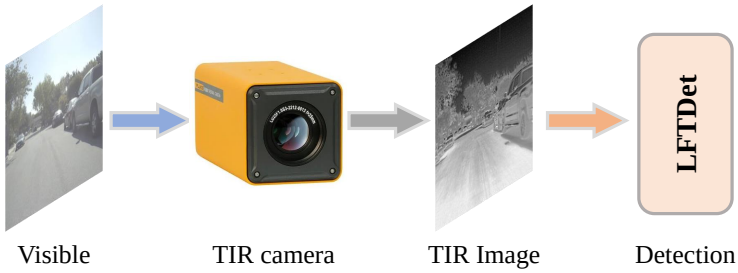


Figure 1: The workflow of the TIR-based object detection by our LFTDet.

taining a minimal number of model parameters and low computational cost. Notably, our *LFTDet* further surpasses the majority of multi-spectral methods employing RGB-TIR image pairs on FLIR dataset, despite solely relying on TIR images as the mono-modal source.

To summarize, our main contributions in this work can be outlined as follows:

- Identification of the challenges associated with conventional object-detection methods in TIR domain.
- Introduction of an efficient model called *LFTDet* for TIR object detection, exclusively utilizing TIR images.
- Utilization of the Fourier Transform (FT) to form the innovative FT-ELAN backbone, which enhances spatial-frequency perception and achieves the global receptive field without introducing additional parameters.
- Development of a novel detection head called AA-FPN to facilitate the operation of our FT-based object detector.
- Extensive experiments conducted on two benchmark datasets, illustrating the advantages of our *LFTDet* and the proposed innovative components.

2 Related Works

2.1 RGB Image Object Detection

Object detection on RGB images has been the focal point of computer vision, with benchmark datasets such as Pascal VOC [34], ADE20K [35], and COCO-Stuff [36] serving as the foundational resources. Classical algorithms like R-CNN [1], Fast R-CNN [37], and Faster R-CNN [38] played a pioneering role in the integration of deep learning into object detection. Subsequently, one-stage approaches, including SSD [3] and YOLO-series models [2, 4–8, 11–13], further facilitated the object detection process, delivering real-time performance while maintaining accuracy. In these methods, Convolutional Neural Networks

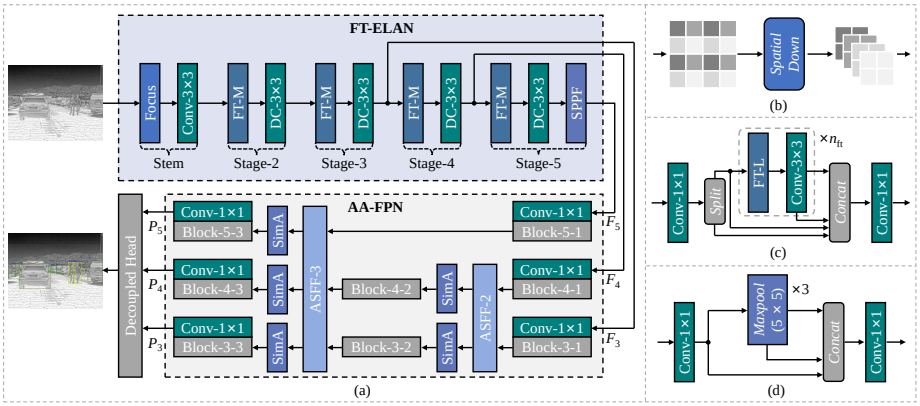


Figure 2: Illustration of our *LFTDet*. (a) The model pipeline of *LFTDet*, which consists of FT-ELAN backbone, AA-FPN neck, and decoupled head. DC -3×3 denotes the the 3×3 Down-Sampling Convolution. (b) The Focus module. (c) The FT-Module (FT-M). (d) The SPPF module.

(CNNs) were commonly employed as backbone models to extract semantic features from images. Recently, the rise of innovative Transformer-based algorithms has marked a paradigm shift in object detection. Notable works include Vision Transformer [39], DETR [40], Deformable DETR [41], and ViTDet [39], which explore the integration of self-attention mechanism [42] to enhance the global-context interaction. In addition, with the development of popular large language model, DetGPT [43] emerges as a multimodal tool which automatically locates the objects according to the user's interest.

2.2 Intelligent Detection

Intelligent detection refers to the use of advanced technologies, such as artificial intelligence, to detect, classify, and predict events or anomalies in various domains. In recent years, there have been significant advancements in data-driven methods for intelligent detection, driven by the availability of large amounts of data, more powerful computational resources, and improved algorithms [44, 45]. Deep learning models can automatically learn complex patterns and features from data, making them highly effective for intelligent detection tasks. Meanwhile, an important development in data-driven methods for intelligent detection is the integration of multiple data modalities. By combining data from different sensors and sources, researchers can improve detection performance and robustness. Recent research on intelligent detection can also be applied to various fault diagnosis [46, 47], energy estimation [48, 49], and thermal vision processing [23–25].

2.3 Object Detection in TIR Domain

Currently, object detection in TIR domain typically relies on multi-spectral RGB-TIR image pairs as the primary data source. It has shown promise in applications such as nighttime driver-assistance systems with restricted visibility. Conventional multi-spectral methods have typically employed CNN-based models, as seen in Zhang *et al.* and Fang *et al.*'s research on cross-modal feature fusion [32, 50–52]. Prominent CNN models like VGG [53] and ResNet [54] have commonly served as the foundational detection backbones for these methods. Recent SOTA approaches, including Transformer-based models such as CMX [55] and IGT [56], have demonstrated significant potential in effectively integrating cross-modality features from visible-thermal image pairs, which hold promise for enhancing the accuracy and robustness of multi-spectral object detection. We have argued that multi-spectral object detection faces limitations in modern driving systems operating under complex environmental conditions, and proposed the mono-modal method TIRDet which leverages prior Thermal-To-Visible (T2V) translation [30], However, it is notable that the prior T2V method still necessitates RGB-TIR image pairs during the training of the preliminary T2V translation model. Therefore, our objective is to develop an efficient mono-modal TIR-based object detector that exclusively requires TIR images for training and inference.

3 Methods

3.1 Overview

Fig. 2(a) shows the model pipeline of our proposed *LFTDet* for TIR image object detection. It is composed of three core components: the FT-ELAN backbone, the AA-FPN detection neck, and the decoupled detection head. Our FT-ELAN backbone is ingeniously designed based on the Discrete Fourier Transform (DFT), an approach that effectively exploits robust spatial-frequency perception and global-range receptive field. This architectural design plays a significant role in enhancing the detection performance of our FT-based detector. To facilitate the training and operation of our FT-based backbone, we introduce a novel detection head called the Asymptotic Attention-based Feature Pyramid Network (AA-FPN), which employs an asymptotic structure with the SimA mechanism [57] to enhance the multi-scale feature interaction. Finally, the decoupled detection head [58] is employed to obtain the detection results on TIR images.

3.2 FT-ELAN backbone

The structure of our FT-ELAN backbone is shown in Fig. 2(a). The FT-ELAN backbone is composed of five distinct stages, with Stage-1 referred to as the Stem. In each of these stages, a down-sampling step is adopted to systematically reduce the feature size. The Stem comprises a single Focus module

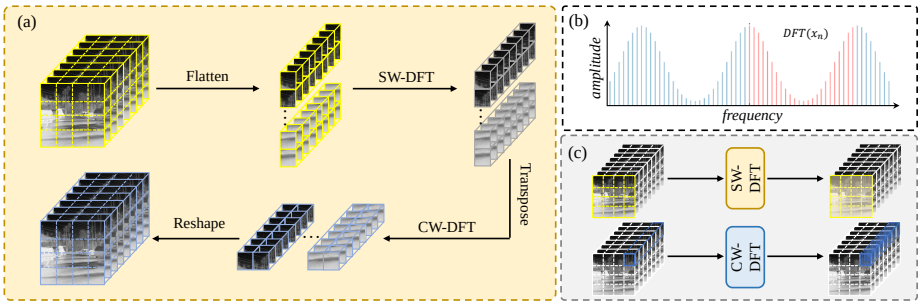


Figure 3: (a) Illustration of the workflow of our FT-Layer (FT-L). SW-DFT denotes Spatial-Wise Discrete Fourier Transform, while CW-DFT denotes Channel-Wise Discrete Fourier Transform. (b) Simple illustration of the Discrete Fourier Transform (DFT). (c) Illustration of the information interaction in SW-DFT and CW-DFT.

in combination with a 3×3 convolution. Every other stage includes an FT-Module (FT-M) and a 3×3 Down-sampling Convolution (DC $- 3 \times 3$). The FT-M is the key component of our FT-ELAN, utilizing cascaded FT-Layers (FT-Ls) to effectively explore robust spatial-frequency perception and expand the global receptive field. Stage-5 also contains a Spatial Pyramid Pooling - Fast (SPPF) module [7] at the end, which substantially enlarges the receptive field by employing a series of max-pooling layers.

3.2.1 Focus Module

As illustrated in Fig. 2(b), the Focus module is employed to process information from different pixels of the input TIR images, which has been widely utilized in the YOLO-series models [7, 58]. It extracts values of the adjacent pixels from the input TIR image, which conducts the spatial downsampling while expanding the channels four times. After that, a 3×3 convolutional layer is followed to extract the semantic information and increase the number of feature channels.

3.2.2 FT-Module (FT-M)

Fig. 2(c) provides the visual representation of our FT-M structure, which greatly augments the model's ability to capture global-context information while maintaining the computational efficiency. To harness the power of the Discrete Fourier Transform (DFT) in extracting long-range information of the features, we integrate the cascaded novel FT-Layers (FT-Ls) in our FT-Ms. In Stage-i ($i \in \{2, 3, 4, 5\}$), each FT-M is followed by one 3×3 Down-Sampling Convolution (DC $- 3 \times 3$) with a stride of 2, which also increases the number of channels twice.

The FT-M initiates with a 1×1 convolutional layer, maintaining the existing channel count. Subsequently, the feature map is divided into two branches,

denoted as f_{top} and f_{bot}, with an equal number of channels. It ensures that our model can effectively leverage global-range information at a low computational cost. The upper branch f_{top} is routed through a total of n_{ft} blocks, where each block consists of an FT-L and a 3×3 convolutional layer:

$$f_{top}^{l} = \phi_{3\times3}(\text{FT-L}(f_{top}^{l-1})), \quad 1 \leq l \leq n_{ft}, \tag{1}$$

where f_{top}^{0} is set equal to f_{top}. We use the symbol ϕ to denote the convolutional layer for brevity, the same as below. Then, the outputs of all blocks and the lower branch f_{bot} are concatenated and fed into a 1×1 convolutional layer to compress the channels:

$$f_{out} = \phi_{1\times1}(\mathcal{C}[f_{top}^{0}, f_{top}^{1}, ..., f_{top}^{n_{ft}}, f_{bot}]), \tag{2}$$

where $\mathcal{C}[\cdot]$ denotes the concatenation process along the channel axis. The output feature, denoted as f_{out}, maintains the same shape as the input feature f_{in}.

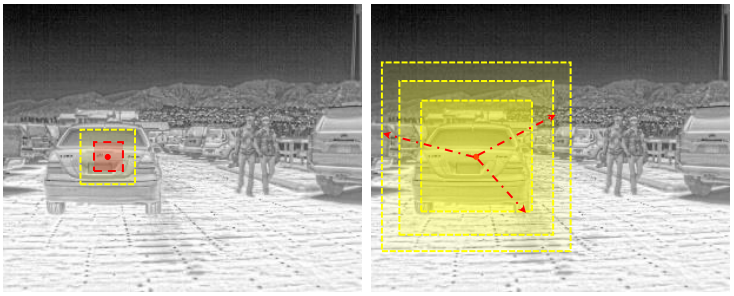
In specific, the workflow of our FT-L is illustrated in Fig. 3(a). As it processes an input feature with the shape of $h \times w \times c$, the following steps are carried out: It firstly reshapes the three-dimensional input feature into a two-dimensional format of $hw \times c$, which flattens the spatial dimensions. Next, we apply the Spatial-Wise Discrete Fourier Transform (SW-DFT) along the spatial direction to enhance the global-range interaction. Afterward, the two-dimensional feature is transposed to the dimension of $c \times hw$, and the Channel-Wise Discrete Fourier Transform (CW-DFT) is performed to extract the in-depth featural information. Finally, the feature is reshaped back to its original shape $h \times w \times c$. The information interaction methods in SW-DFT and CW-DFT are illustrated in Fig. 3(c).

In the process of the SW-DFT and CW-DFT operation, if the input two-dimensional feature x contains N discrete latent vectors, the operation of DFT [59] on it can be illustrated in Fig. 3(b) and formulated as:

$$X_k = \text{DFT}(x) = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} nk}, \quad 0 \leq k \leq N-1, \tag{3}$$

where n represents the number of sequential vectors, which ranges from 0 to $N-1$. For each value of k, the DFT operation computes a new representation X_k as the summation of all the original input tokens x_n, each weighted by the complex exponential factor. In contrast to the conventional convolutional layers, the DFT layers can extract long-range information from the input tensors without learnable parameters and additional computational cost.

By incorporating the cascaded FT-Ls into the FT-Ms, our FT-ELAN backbone gains the capability to effectively capture spatial-frequential information from all locations and depths of the feature maps. In contrast to the ELAN backbone [60], which relies on two consecutive 3×3 convolutions to perceive a 5×5 region, our FT-ELAN leverages the power of the DFT to reduce computational demands while achieving an even more extensive receptive field, as shown in Fig. 4. In addition, it not only improves the model's efficiency but also



(a) Two cascaded 3 × 3 Conv (b) FT-L + 3 × 3 Conv

Figure 4: Illustration of the information flow in feature extraction. Compared to two consecutive 3×3 convolutions, which perceive a 5×5 region, our FT-M leverages Fourier Transform to achieve a global-range receptive field.

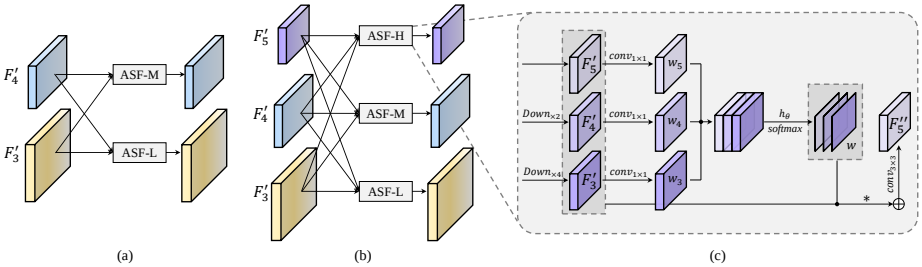


Figure 5: (a) Illustration of the structure of ASFF-2. (b) Illustration of the structure of ASFF-3. (c) Illustration of the workflow of ASF-H in ASFF-3. In (a) and (b), the upward arrows represent down-sampling steps, while the downward arrows represent up-sampling steps.

enables our *LFTDet* to process information from a more comprehensive view, ultimately enhancing its stable performance in TIR-domain applications.

3.2.3 SPPF Module

The SPPF module is illustrated in Fig. 2(d). The input feature f_{in} is firstly sent into a 1×1 convolutional layer to obtain $f_s^0 = \phi_{1\times1}(f_{in})$: Then, it is fed into a cascade of three max-pooling layers with the size of 5×5 and the stride of 1:

$$f_s^l = Maxpool_{5\times5}(f_s^{l-1}), \quad 1 \leq l \leq 3. \tag{4}$$

Afterwards, f_s^0 and the outputs of all max-pooling layers are concatenated and sent into a 1×1 convolutional layer:

$$f_{out} = \phi_{1\times1}(\mathcal{C}[f_s^l]), \quad l \in \{0, 1, 2, 3\}. \tag{5}$$

where the output f_{out} has the same shape as the input f_{in}.

The SPPF module allows for the efficient feature extraction at multiple scales. By performing the cascaded max-pooling operation, the SPPF module optimizes feature extraction while minimizing the computational cost.

3.3 AA-FPN Detection Neck

Our AA-FPN is illustrated in Fig. 2(a), which takes features from Stage-3, Stage-4, and Stage-5 of the FT-ELAN backbone as the input, denoted as $\{F_3, F_4, F_5\}$. The asymptotic structure incorporates ASFF modules [61] to seamlessly integrate and extract multi-scale information. Simultaneously, we leverage the SimA mechanism [57] (denoted as $\mathcal{S}_\alpha$) to enhance feature awareness of the detection neck, all without introducing any additional parameters.

3.3.1 Asymptotic Structure

We use $\mathcal{A}_2$ and $\mathcal{A}_3$ to denote ASFF-2 and ASFF-3, respectively. The AA-FPN begins by employing a 1×1 convolution and stacked convolution-based blocks (Block-i-j denoted as $\Phi_{i\text{-}j}$) to derive the latent features, denoted as $\{F_3^{'}, F_4^{'}, F_5^{'}\}$:

$$F_i^{'} = \Phi_{i\text{-}1}(\phi_{1\times1}(F_i)), \quad i \in \{3, 4, 5\}. \tag{6}$$

Subsequently, we utilize the ASFF-2 low-level feature aggregation module to integrate $F_3^{'}$ and $F_4^{'}$, namely $F_3^{'}, F_4^{'} = \mathcal{A}_2(F_3^{'}, F_4^{'})$. Following this step, both $F_3^{'}$ and $F_4^{'}$ are individually processed through the SimA modules and convolution-based blocks:

$$F_i^{'} = \Phi_{i\text{-}2}(\mathcal{S}_\alpha(F_i^{'})), \quad i \in \{3, 4\}. \tag{7}$$

The multi-scale features are then directed to the high-level feature aggregation module, ASFF-3, for adaptive spatial fusion: $F_3^{''}, F_4^{''}, F_5^{''} = \mathcal{A}_3(F_3^{'}, F_4^{'}, F_5^{'})$. Finally, we apply the SimA modules, convolution-based blocks, and 1×1 convolution to process the features, generating $\{P_3, P_4, P_5\}$:

$$P_i = \phi_{1\times1}(\Phi_{i\text{-}3}(\mathcal{S}_\alpha(F_i^{''}))), \quad i \in \{3, 4, 5\}, \tag{8}$$

where the resulting features are subsequently utilized in the detection head to produce the final results.

3.3.2 ASFF Module

The ASFF modules incorporate feature weighting mechanisms at various levels to enhance the feature fusion. The weighting scheme proves particularly efficient when reconciling potentially conflicting information in features with varying scales. The workflows of ASFF-2 and ASFF-3 are illustrated in Fig. 5(a) and Fig. 5(b), respectively. Both ASFF-2 and ASFF-3 consist of ASF-Low (ASF-L), ASF-Middle (ASF-M), and ASF-High (ASF-H) modules. In specific, ASF-H is only included in ASFF-3.

To elucidate the workflow, we take the workflow of ASF-H in ASFF-3 as an example (shown in Fig. 5(c)). Firstly, all features are resized to match the shape

of $F_5^{'}$, specifically, $F_3^{'} = Down_{\times 4}(F_3^{'})$ and $F_4^{'} = Down_{\times 2}(F_4^{'})$. Subsequently, we apply 1×1 convolutions separately on the features to derive the weights $w_i = \phi_{1\times 1}(F_i^{'})(i \in \{3, 4, 5\})$. Following this, the distinct weights w_i are concatenated and processed with a non-linear layer employing softmax activation:

$$w = \text{softmax}(h_\theta(\mathcal{C}[\phi_{1\times 1}(F_i^{'})])), \quad i \in \{3, 4, 5\}, \tag{9}$$

where $w \in \mathbb{R}^{h\times w\times 3}$, and h_θ represents a 3×3 convolutional layer to compress the number of channels into three. Finally, the three channels of w serve as the attention weight maps for the input features. An additional 3×3 convolutional layer generates the resultant output $F_5^{''}$:

$$F_5^{''} = \phi_{3\times 3}(F_3^{'} * w[0] + F_4^{'} * w[1] + F_5^{'} * w[2]). \tag{10}$$

A similar computational strategy is applied to ASF-L and ASF-M in both ASFF-2 and ASFF-3, where the distinct attention weights are computed and applied to the input features.

3.3.3 SimA Module

We employ SimA modules on the output features of ASFF-2 and ASFF-3 modules. The SimA module, as illustrated in Fig. 6, serves as a parameter-free attention mechanism that infers 3-D attention weights for a feature map within a given layer. For a given input feature $x_{in} \in \mathbb{R}^{h\times w\times c}$ and a scalar λ, the following steps are carried out:

We firstly calculate the channel-wise variance $\mathcal{V}(x_{in})$:

$$\mathcal{V}(x_{in}) = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \mu(x_{in})\right)^2, \quad n = h \times w, \tag{11}$$

where $\mathcal{V}(x_{in}) \in \mathbb{R}^{h\times w\times 1}$. In this equation, x_k represents the elements of x_{in}, and $\mu(x_{in})$ is the mean along spatial dimensions. Then we obtain the attention weight $Attn$:

$$Attn = \frac{\mathcal{T}(x_{in})}{4(\mathcal{V}(x_{in}) + \lambda)} + threshold, \tag{12}$$

where $\mathcal{T}(x_{in})$ is the squared difference tensor between x_{in} and its spatial mean. The scalar λ and threshold are set to $1e-4$ and 0.5 in this work, respectively. Finally, we obtain the weighted output feature by element-wise multiplication with the sigmoid form of $Attn$, namely $x_{out} = x_{in} \odot \sigma(Attn)$.

In all, the SimA module involves calculating channel-wise variances in the input feature and using them to compute attention weights, which are then applied to generate an attention-weighted feature. Combined with the asymptotic neck structure, the dynamic attention mechanism enhances the model's capacity to emphasize significant semantic information without the introduction of additional parameters.

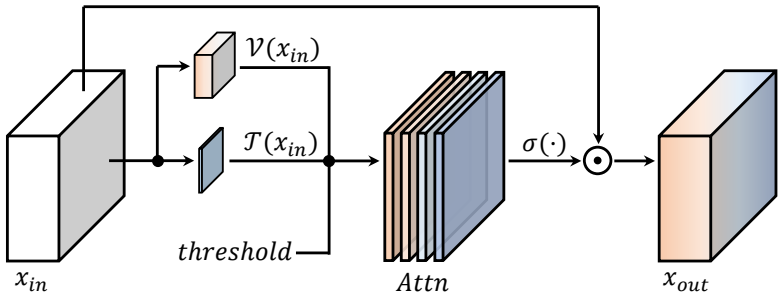


Figure 6: Illustration of the workflow of SimA module. The symbol $\sigma(\cdot)$ denotes the sigmoid function.

3.4 Decoupled Detection Head

The decoupled detection head, depicted in Fig. 7, is an anchor-free detection module. Incorporating features from the AA-FPN ($P3$, $P4$, and $P5$), it employs a 1×1 convolution operation to enhance the features and reduce the channel dimension to 256. Subsequently, the feature map is split into two distinct branches, each consisting of two consecutive 3×3 convolutional layers. The upper branch is dedicated to class determination (Cls.), responsible for assigning the feature point to a specific object category. On the other hand, the lower branch is accompanied by two parallel prediction layers: Reg. (Regression) and IoU. (Intersection over Union). Reg. predicts the center point coordinates of the detection box, along with the box width and height, whereas IoU. estimates the likelihood that the predicted box encompasses a target object. Once the outputs of all prediction branches are generated, we follow the training and inference strategies in [58] to derive the final detection results.

3.5 Model Configurations

We have built an *LFTDet-T* as a tiny-size network with a model size of 5.30 M and the computational cost of 11.15 GFLOPs. In addition, we also build *LFTDet-B* as a base-size network to make a competitive comparison, which has the model size of 13.52 M and the computational cost of 29.99 GFLOPs. The main differences in model configurations lie in: (1) The number of feature channels after the Stem module (denoted as C), which will be proportionally amplified at each stage of the model. (2) The number of cascaded FT-L and 3×3 convolutions in FT-Ms (denoted as n_{ft}). (3) The number of convolutional blocks in $\Phi_{3\text{-}2}$, $\Phi_{4\text{-}2}$, $\Phi_{3\text{-}3}$, $\Phi_{4\text{-}3}$, $\Phi_{5\text{-}3}$ in AA-FPN (denoted as *depth*). Once the number of model parameters increases, the detection results will become better, but at the same time, the computational efficiency will decrease. We also found that in the TIR object detection task, the performance improvement brought by increasing the calculation amount based on LFTDet-B is not substantial. Hence, based on our original design objectives, we choose these two parameter

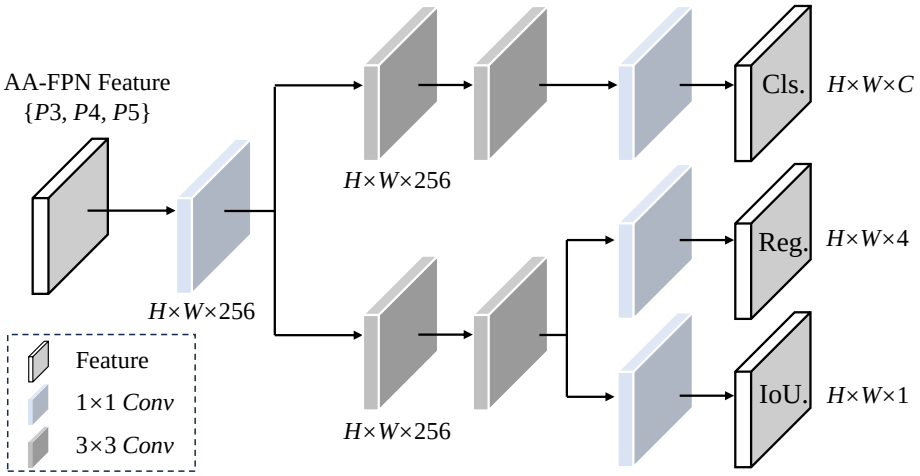


Figure 7: Illustration of the decoupled detection head. C represents the number of object classes in the dataset.

Table 1: Model configurations of *LFTDet-T* and *LFTDet-B*.

Model	Configs			Params	FLOPs
	C	n_{ft}	*depth*		
LFTDet-T	32	4	1	5.30 M	11.15 G
LFTDet-B	48	6	2	13.52 M	29.99 G

settings. The specific model configurations of *LFTDet-T* and *LFTDet-B* are listed in Table 1.

4 Experiments

4.1 Implementation

We employ PyTorch version 1.8.1 for the implementation of our proposed *LFTDet*, which is executed on an Intel(R) Xeon(R) Gold 5218 CPU operating at 2.30GHz, paired with NVIDIA RTX A6000 GPUs equipped with CUDA 11.4. Our experimentation process is conducted using the MMDetection framework [64, 65]. For training our *TIRDet*, we have utilized the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.01, a momentum value of 0.9, a weight decay of 0.0005. The batch size of 8 is configured. The training process spans 300 epochs, and we close the Mosaic and Mixup data augmentation after 200 epochs.

Table 2: Quantitative results of conventional object-detection methods and our LFTDet on FLIR and LLVIP datasets.

Model	FLIR				LLVIP				Params(M) (±0.10)	GFLOPs (±0.20)	FPS
	mAP_{50}	mAP_{75}	mAP	mAR	mAP_{50}	mAP_{75}	mAP	mAR			
Faster RCNN [38]	74.4	32.5	37.6	49.7	96.1	68.5	61.1	59.7	41.13	91.01	1.35
SSD [3]	65.5	22.4	29.6	44.3	90.2	57.9	53.5	57.3	24.01	137.53	1.54
RetinaNet [62]	64.5	20.3	28.3	44.4	93.7	49.3	50.9	60.6	36.15	82.04	1.14
YOLOv3 [5]	73.6	31.3	36.8	46.5	89.7	53.4	52.8	61.7	61.53	77.56	2.07
YOLOv5 [7]	73.9	35.7	39.5	47.3	94.6	72.2	61.9	59.5	46.52	109.14	1.87
YOLOF [8]	74.9	26.7	34.6	47.9	91.4	43.7	47.5	58.6	42.11	39.28	1.48
DDOD [63]	72.7	26.2	33.9	48.2	94.3	59.9	56.6	63.7	31.98	71.22	0.57
YOLOX [58]	80.9	37.5	42.0	52.2	95.7	71.5	62.3	68.3	54.15	77.66	1.20
YOLOv7 [60]	75.6	32.2	38.2	49.0	95.5	67.7	59.4	62.5	36.49	103.51	1.60
YOLOv9 [12]	74.9	35.8	39.6	50.1	95.1	68.2	61.0	65.4	20.05	76.35	1.98
LFTDet-T	79.3	37.8	42.3	52.2	95.7	71.4	63.4	68.0	5.30	11.15	4.79
LFTDet-B	81.1	41.3	44.3	54.0	96.2	72.7	63.8	69.0	13.52	29.99	2.32

GFLOPs are tested at 640×640 resolution with batch = 1. FPS is tested on an Intel(R) Xeon(R) Gold 5218 CPU operating at 2.30GHz. The values marked in red denote the best results, while the values marked in blue denote the second best results.

4.2 Datasets

We conduct extensive experiments on two benchmark datasets for TIR image object detection, namely FLIR [31, 32] and LLVIP [33] datasets. The detailed information of the two datasets is described as follows:

4.2.1 FLIR Dataset

FLIR is a multi-spectral road-scene dataset, which has three annotated categories: "person", "bicycle", and "car". Later, a "clean-version" FLIR dataset was proposed, which removed the mis-annotated thermal images [32]. In this work, we adopt the "clean-version" FLIR dataset to conduct the experiments, which contains 4,129 training images and 1,013 testing images.

4.2.2 LLVIP Dataset

LLVIP is a newly introduced multi-spectral dataset for nighttime pedestrian detection. It encompasses 12,025 training images and 3,463 testing images. The TIR images in the dataset are captured across the wavelength of 8—14μm in exceptionally low-light environments with high image clarity.

4.3 Evaluation Metrics

We adopt standard metrics in object detection for quantitative evaluation, including mean Average Precision (mAP), mAP_{50}, mAP_{75}, and mean Average

Table 3: Quantitative results of multi-spectral methods and our LFTDet on FLIR dataset.

Model	Source	Backbone	mAP_{50}	mAP_{75}	mAP	mAR
CFR_3 [32]	V+T	VGG16	72.4	-	-	-
GAFF [50]	V+T	ResNet18	72.9	32.9	37.5	-
GAFF [50]	V+T	VGG16	72.7	30.9	37.3	-
YOLOFusion [52]	V+T	VGG16	76.6	-	39.8	-
CFT [51]	V+T	CFB	78.7	35.5	40.2	52.5
InfusionNet [66]	V+T	Infusion	79.1	35.2	40.3	-
CMX [55]	V+T	SwinT	82.2	37.1	42.3	-
IGT [56]	V+T	SwinT	85.0	36.9	43.6	-
TIRDet [30]	T→V	CSPD53	81.4	41.1	44.3	54.0
LFTDet-T	**T**	FT-ELAN	79.3	37.8	42.3	52.2
LFTDet-B	**T**	FT-ELAN	81.1	41.3	44.3	54.0

V denotes the visible input, while T denotes the thermal input. T→V denotes the prior Thermal-To-Visible Translation method.

Table 4: Quantitative results of multi-spectral methods and our LFTDet on LLVIP dataset.

Model	Source	Backbone	mAP_{50}	mAP_{75}	mAP	mAR
YOLOv5-VT [7]	V+T	CSPD53	95.8	71.4	62.3	63.1
CFT [51]	V+T	CFB	97.5	72.9	63.6	68.4
InfusionNet [66]	V+T	Infusion	98.6	73.3	64.6	-
TIRDet [30]	T→V	CSPD53	96.3	73.1	64.2	69.4
LFTDet-T	**T**	FT-ELAN	95.7	71.4	63.4	68.0
LFTDet-B	**T**	FT-ELAN	96.2	72.7	63.8	69.0

Recall (mAR). mAP and mAR are computed as the average values across all categories, using Intersection over Union (IoU) thresholds ranging from 0.50 to 0.95, with an interval of 0.05. Meanwhile, mAP_{50} and mAP_{75} are calculated at IoU thresholds of 0.50 and 0.75, respectively. The mAP score can be calculated as:

$$\text{mAP} = \frac{AP}{N} = \frac{\int_0^1 p(r)dr}{N}, \tag{13}$$

where $p = TP/(TP + FP)$ denotes the precision, $r = TP/(TP + FN)$ denotes the recall, and N is the number of categories. Meanwhile, TP, FP, and FN represent True-Positive, False-Positive, and False-Negative elements, respectively.

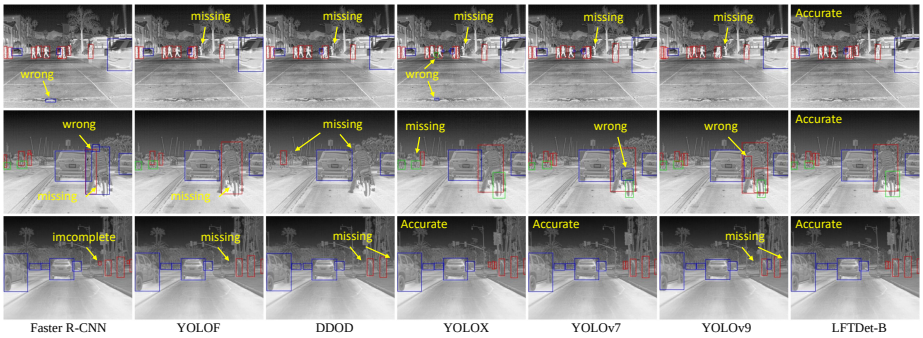


Figure 8: Qualitative comparison between the baseline methods and our *LFTDet-B* on FLIR dataset, where the red, blue, and green boxes denote the detected objects "person", "car", and "bicycle", respectively. Better viewed in color and zoomed in.

4.4 Quantitative Results

The quantitative results of conventional object-detection methods and our *LFTDet* on FLIR and LLVIP datasets are shown in Table 2[1]. All of them adopt TIR images as the input source for mono-modal object detection. For the YOLO-series models, we adopt the large-version variants for evaluation. Notably, the parameters (M) and GFLOPs exhibit a slight margin of error, approximately $\pm$ 0.10 and $\pm$ 0.20, respectively, due to the different object categories incorporated within the classification head. From the comparison, we can see that our *LFTDet-B* stands out by achieving the most outstanding results across all metrics, demonstrating superior performance on both the FLIR and LLVIP datasets. It achieves the remarkable 44.3% mAP and 54.0 % mAR on FLIR dataset, and 63.8% mAP and 69.0 % mAR on LLVIP dataset, respectively, with a minimal parameter count of 13.52 M and a computational cost of 29.99 GFLOPs. Furthermore, our *LFTDet-T* secures the second-best position in terms of mAP_{75}, mAP, and mAR on FLIR dataset, as well as the second-best mAP on the LLVIP dataset, at an extremely low number of parameters and GFLOPs required. These results underscore the efficacy of our *LFTDet*, which harnesses zero-parameter Fourier-Transform operation to extract global-range information. Meanwhile, we also test the Frames Per Second (FPS) on our Intel(R) Xeon(R) Gold 5218 CPU to verify our *LFTDet*'s ability in mobile low-light applications. Remarkably, our *LFTDet-T* and *LFTDet-B* models exhibit exceptional performance, which achieve the FPS rates of 4.79 and 2.32, respectively, showing the practicality and efficiency of our approach.

On the other hand, the quantitative results of multi-spectral methods and our *LFTDet* on FLIR and LLVIP datasets are presented in Table 3 and Table 4, respectively. The multi-spectral methods adopt the RGB-TIR image pairs as the input, while our *LFTDet* is a mono-modal TIR object detector.

[1] All quantitative results in the tables are recorded in (%).

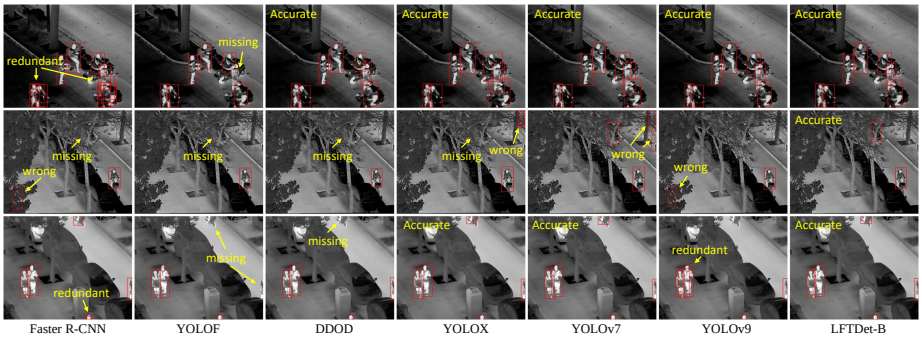


Figure 9: Qualitative comparison between the baseline methods and our *LFTDet-B* on LLVIP dataset, where the red boxes denote the detected objects "person". Better viewed in color and zoomed in.

From Table 3, it is evident that our *LFTDet-B* maintains its exceptional performance by achieving the highest mAP_{75} (41.3%), mAP (44.3%), and mAR (54.3%) on FLIR dataset, even in the absence of input visible images. Overall, TIRDet, which also adopts the TIR images as the mono-modal input, outperforms the multi-spectral methods, matching the best mAP and mAR achieved as our *LFTDet-B*. However, TIRDet still integrates a prior Thermal-To-Visible translation model, which necessitates access to RGB-TIR image pairs during the pre-training phase, consequently introducing additional parameters into the model. From Table 4, we observe that our *LFTDet-T* and *LFTDet-B* exhibit slightly lower performance compared to the multi-spectral methods, especially InfusionNet. This discrepancy can be attributed to the fact that LLVIP dataset contains high-quality visible images, whereas our method exclusively utilizes TIR images as the mono-modal input source.

Overall, our proposed *LFTDet* outperforms conventional object detectors that rely solely on TIR images. Although *LFTDet* eliminates visible images as the input source, it still surpasses the performance of most multi-spectral methods, underscoring its practical applicability in modern driving systems low-light remote sensing.

4.5 Qualitative Results

In Fig. 8 and Fig. 9, we present the qualitative detection results to evaluate the performance of our *LFTDet-B* and the competitive baseline methods, including Faster R-CNN, YOLOF, DDOD, YOLOX, YOLOv7, and YOLOv9. The results are meticulously obtained by employing a confidence score threshold of 0.5 [2]. Overall, the models' performance on LLVIP dataset is better than that on FLIR dataset, since FLIR dataset has more object categories and complex background. Obviously, our *LFTDet-B* stands out by virtue of its exemplary performance in accurately detecting pedestrians, cars, and bicycles within TIR images. Its precision in identifying objects in challenging TIR scenarios demonstrates its

Table 5: Ablation study on our *LFTDet-T* and *LFTDet-B*.

Model		FLIR		LLVIP	
		mAP	mAR	mAP	mAR
LFTDet-T	**Complete**	**42.3**	**52.2**	**63.4**	**68.0**
	w/o FT-M	38.5	49.1	59.2	62.7
	w/o FT-L	39.9	48.8	60.8	63.7
	w/o SimA	40.0	51.1	61.7	65.9
	w/o SPPF	41.7	51.5	63.0	66.9
	FT-L$\rightarrow \phi_{3\times3}$	40.5	51.4	61.9	66.8
LFTDet-B	**Complete**	**44.3**	**54.0**	**63.8**	**69.0**
	w/o FT-M	41.2	50.1	60.4	64.5
	w/o FT-L	42.4	51.0	61.3	64.8
	w/o SimA	43.0	52.9	62.8	68.3
	w/o SPPF	43.3	53.5	63.2	68.6
	FT-L$\rightarrow \phi_{3\times3}$	42.9	53.0	62.5	68.0

"FT-L$\rightarrow \phi_{3\times3}$" denotes the variant model which replaces FT-Ls with 3×3 convolution.

applicability. In contrast, several of the methods encounter significant difficulties in identifying minuscule objects, such as the diminutive bicycles and pedestrians in the second row of TIR images on FLIR dataset. Furthermore, YOLOv7 tend to produce redundant detection boxes in intricate scenes, especially on LLVIP dataset, where pedestrians are in close proximity to one another. In addition, the latest method YOLOv9 still generates some wrong results or miss the key objects on the two datasets. The noteworthy advantage of our *LFTDet-B* is its efficient use of model parameters and computational resources while maintaining a high level of precision and delivering consistently accurate detection results. The qualitative advantages position our model's ability in the domain of TIR-based object detection, particularly for applications that demand both efficiency and reliability, such as nighttime driving and remote sensing.

4.6 Ablation Study

To demonstrate the effectiveness of the core components within our *LFTDet* model, we conduct a comprehensive ablation study and present the experimental results in Table 5. The results obviously highlight the superior performance of the complete *LFTDet-T* and *LFTDet-B* models compared to their variant counterparts. Among the variants, the most notable performance drop is observed in the model referred to as "w/o FT-M", which exhibits a significant decrease of 5.3% in mAR on the LLVIP dataset compared with *LFTDet-T*. This striking decrease underscores the critical importance of the FT-Ms within the model backbone. Additionally, variants that remove the FT-Ls or replace them with 3×3 convolution also demonstrate varying degrees of performance

drops. This observation underscores the pivotal role that FT-Ls play in enabling the model to capture long-range interaction effectively. Furthermore, the model denoted as "w/o SimA" still displays a substantial performance decrease, emphasizing the significant contribution of SimA modules in enhancing the *LFTDet* model's performance. The lack of SPPF module exerts a relatively minor influence on the performance, leading to a marginal drop in mAP from 0.4% to 1.0% and a drop in mAR from 0.4% to 1.1%.

In summary, it is evident that each of the key components within our *LFTDet* model plays a vital role in achieving its high performance, with the FT-L and FT-M standing out as the most crucial elements.

4.7 Effectiveness of FT-ELAN and AA-FPN

To provide a more comprehensive illustration of the advantages of our FT-ELAN and AA-FPN compared to conventional detection backbones and necks, we have conducted a series of experiments as follows: We have trained and tested our detectors using three different backbones: ResNet50 [54], CSPDarknet53 [7], and E-ELAN [60], all integrated with our AA-FPN neck. Simultaneously, we have replaced the AA-FPN neck with three alternatives, namely FPN [67], BiFPN [68], and PAFPN [58], in our *LFTDet-B* model.

The results of these experiments are summarized in Table 6. From the comparison, we can see that our *LFTDet-B* (FT-ELAN + AA-FPN) exhibits the most compelling quantitative results on the two datasets. Notably, the detector employing the CSPDarknet53 backbone in combination with our AA-FPN neck has also shown impressive performance, achieving a mAP of 43.5% and a mAR of 53.4% on the FLIR dataset, as well as a mAP of 63.3% and a mAR of 68.6% on the LLVIP dataset. In contrast, when we replace the AA-FPN neck with the standard FPN, there is a noticeable decline in performance, with mAP dropping by 1.6% on the FLIR dataset and 1.0% on the LLVIP dataset. In summary, our FT-ELAN backbone consistently outperforms conventional backbone models across both datasets, and the AA-FPN neck plays a vital role in enhancing its performance further. These results highlight the efficacy of our FT-ELAN and AA-FPN components in achieving superior quantitative outcomes in comparison to traditional detection architectures.

4.8 Effectiveness of FT-L

The ablation study clearly demonstrates that the absence of FT-L or FT-M results in a significant decline in the performance of our *LFTDet* model. To further investigate the effectiveness of our proposed FT-L, we conducted a visualization analysis of the features within the FT-ELAN backbone. Fig. 10 showcases the visualized features after the Stem, Stage-2, and Stage-3 for illustrative purposes.

Upon observing the features after the Stem, it becomes evident that they predominantly focus on the illuminated regions of the TIR images, such as the

Table 6: Quantitative results of the variant models with different detection backbones or necks.

Backbone	Neck	FLIR		LLVIP	
		mAP	mAR	mAP	mAR
ResNet50 [54]	AA-FPN	43.1	53.5	62.9	68.2
CSPD53 [7]	AA-FPN	43.5	53.4	63.3	68.6
E-ELAN [60]	AA-FPN	43.0	53.1	62.8	68.4
FT-ELAN	FPN [67]	42.7	52.9	62.8	67.9
	BiFPN [68]	43.1	53.0	63.2	67.7
	PAFPN [58]	43.3	53.5	63.0	68.1
	AA-FPN	**44.3**	**54.0**	**63.8**	**69.0**

The models keep the same model configurations as LFTDet-B.

ground in the second TIR image from the FLIR dataset. This observation indicates that the Focus module in the shallow layer has not sufficiently learned the target regions and objects. In contrast, the features after Stage-2 exhibit a reduced emphasis on illumination and a heightened attention towards the target objects and their edges. This shift in focus suggests that the model is starting to learn and capture relevant information about the objects of interest. Moreover, the features after Stage-3 clearly demonstrate a stronger focus on the target objects and a better establishment of global-range interactions. This indicates that our FT-L effectively improves the detection ability of the model while still maintaining its low computational cost. In summary, the gradual enhancement of target attention and long-range perception through the FT-L highlights its ability to efficiently improve the detection performance of our *LFTDet* model. This is achieved without compromising on computational efficiency. The continual refinement of target attention and long-range interactions is crucial for advancing the capabilities of our model.

4.9 Effectiveness of SimA

To thoroughly investigate the advantages of the SimA modules over conventional attention modules in our *LFTDet*, we replace the SimA modules in our *LFTDet-B* with various attention modules, including CA (Channel Attention) [69], SA (Spatial Attention) [70], CBAM (Convolutional Block Attention Module) [70], and SK (Selective Kernel) [71]. The experimental results are presented in Table 7.

Analyzing the results, it is evident that our model with the integrated SimA modules consistently outperforms the variant models with other attention modules. Notably, the CBAM and SK modules demonstrate superior performance compared to the CA and SA modules. In fact, the model incorporating the CBAM module achieves the same highest mean Average Precision (mAP) of 63.8% as our *LFTDet-B* on the LLVIP dataset. However, there is an exception where the model with SA attention modules performs even worse than the

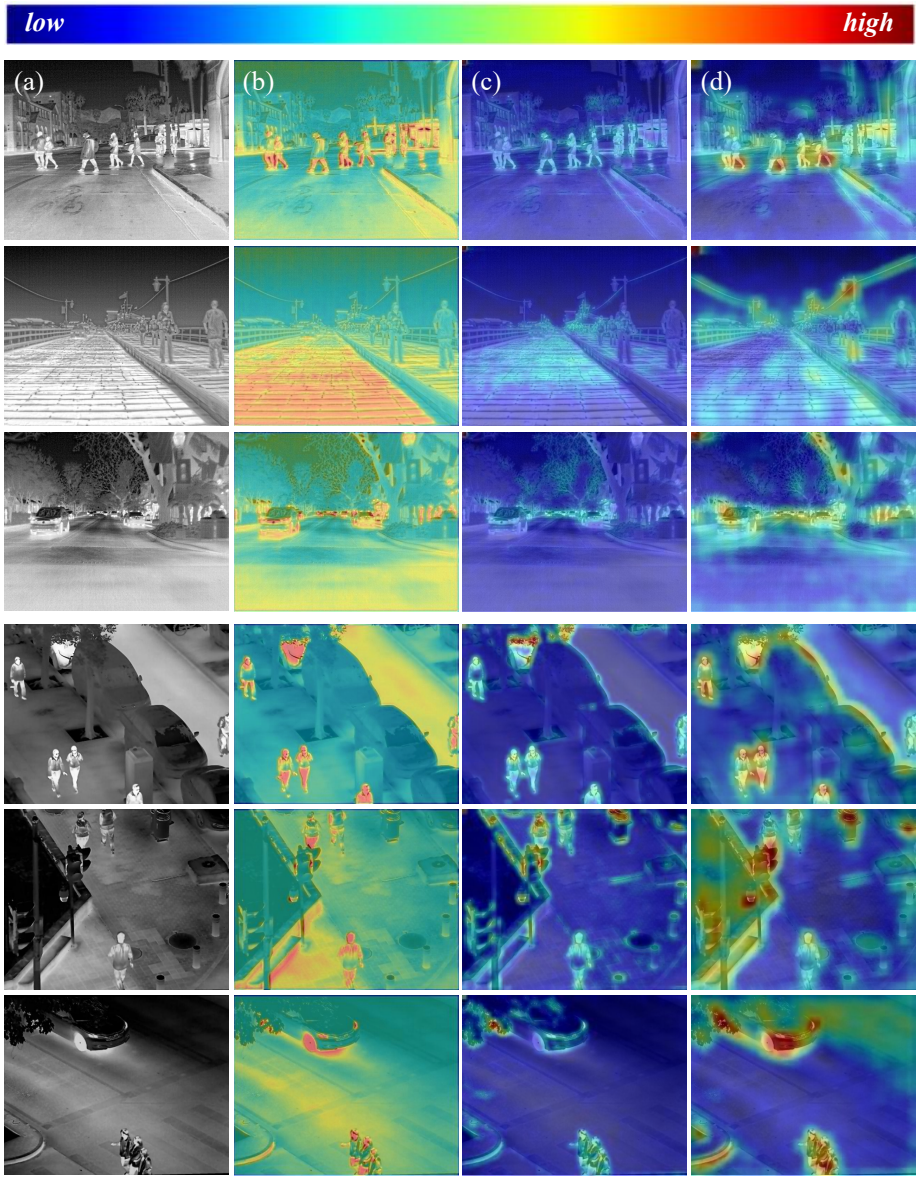


Figure 10: (a) The input TIR images. (b) Visualization of the feature maps after Stem. (c) Visualization of the feature maps after Stage-2. (d) Visualization of the feature maps after Stage-3.

model without any attention modules, particularly on the LLVIP dataset. It suggests that the SA module's attention mechanism might not appropriately capture the in-depth relevant information. On the contrary, the SimA modules

Table 7: Comparison between SimA and other attention modules.

Module	FLIR		LLVIP	
	mAP	mAR	mAP	mAR
None	43.0	52.9	62.8	68.3
SimA	**44.3**	**54.0**	**63.8**	**69.0**
CA [69]	43.0	53.1	62.9	68.3
SA [70]	43.3	53.4	62.6	67.9
CBAM [70]	43.9	53.6	**63.8**	68.7
SK [71]	44.1	53.6	63.5	68.6

The variant models keep the same model configurations as LFTDet-B.

consistently achieve the best performance due to their dynamic attention mechanism within the asymptotic neck structure. Remarkably, the SimA achieves this without introducing any additional learnable parameters, thus maintaining efficiency of our model. Overall, these experimental results clearly demonstrate that the SimA modules in our *LFTDet* offer distinct advantages over conventional attention modules.

4.10 User Study

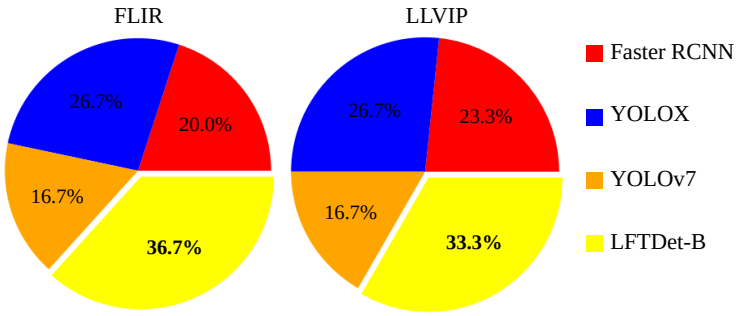


Figure 11: Results of user study: voting percentages from 30 volunteers.

To demonstrate the practical effectiveness of our method, we conducted a user study involving 30 volunteers. Prior to the experiments, we gathered the inference results and chose 120 consecutive images to create 30-second videos from the FLIR and LLVIP datasets. To compare our *LFTDet-B* with other methods, namely Faster R-CNN, YOLOX, and YOLOv7, each volunteer simultaneously watched the four videos and selected the one they believed to be the best. The results of voting percentages are illustrated in Fig. 11. The results indicate that our *LFTDet-B* received the highest number of votes on both datasets, with 36.7% on FLIR and 33.3% on LLVIP. YOLOX also achieved sig-

nificant popularity, while Faster R-CNN performed well on LLVIP but poorly on FLIR. Overall, the user study further validates the practical effectiveness of our proposed *LFTDet* method.

5 Discussion

In this section, we aim to discuss the strengths and weaknesses of our proposed LFTDet.

5.1 Strengths

To our knowledge, this is the first work that applies Fourier-Transform (FT) to Thermal InfraRed (TIR) vision tasks. In this work, we propose the FT-ELAN network as the backbone model using the FT strategy to enhance efficiency and performance. The proposed AA-FPN aims to facilitate the operation of the FT operation. In contrast to previous algorithms that utilized the Fourier Transform in deep learning, our FT-ELAN method delves into the efficacy of the Discrete Fourier Transform (DFT) in facilitating interaction between spatial-wise and channel-wise information, which effectively explores robust spatial-frequency perception and global-range receptive field. Earlier algorithms, including FNet [59], Fast Fourier Convolution (FFC) in LaMa [72], and EFDMs [73], primarily focused on one-dimensional signal processing or concentrated on one-dimensional aspects of multidimensional features.

5.2 Weaknesses

Due to its utilization of the Fourier Transform (FT) strategy, there may be challenges related to efficiency when running on certain specialized hardware setups. Notably, our LFTDet-T and LFTDet-B models have showcased impressive performance, achieving frame-per-second (FPS) rates of 4.79 and 2.32, respectively, on the Intel(R) Xeon(R) Gold 5218 CPU. However, when deployed on unconventional hardware devices such as those utilized in aerospace or satellite systems, the efficiency of FT operations may diminish. Therefore, addressing the issue of hardware adaptability for Fourier Transform becomes imperative in ensuring optimal performance across various hardware environments.

6 Conclusion

In this work, we propose a lightweight TIR object detector called *LFTDet*. It leverages our innovative FT-ELAN backbone, which employs Fourier Transform to capture in-depth information and establish long-range interaction within TIR images. In addition, we propose the AA-FPN neck, featuring an asymptotic structure with the SimA mechanism, designed to enhance the detection performance of *LFTDet*. The extensive experiments on FLIR and LLVIP datasets

demonstrate that our *LFTDet-T* and *LFTDet-B* significantly outperform conventional object detectors, such as Faster R-CNN, YOLOX, YOLOv7, and YOLOv9, while also maintaining a minimal number of parameters and low computational cost. Meanwhile, the ablation studies and additional experiments further validate the effectiveness of our FT-ELAN backbone and AA-FPN neck. In all, these results highlight the suitability of *LFTDet* for modern driving systems and remote sensing applications. Moving forward, we plan to explore potential enhancements to the training methods of *LFTDet*, such as incorporating contrastive learning, to further improve its detection performance.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[4] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[5] ——, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[6] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[7] G. J. et. al., "ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support," Oct. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5563715

[8] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 039–13 048.

[9] R. Ravindran, M. J. Santora, and M. M. Jamali, "Multi-object detection and tracking, based on dnn, for autonomous vehicles: A review," *IEEE Sensors Journal*, vol. 21, no. 5, pp. 5668–5677, 2020.

[10] C. Jiang, Z. Wang, H. Liang, and S. Tan, "A fast and high-performance object proposal method for vision sensors: Application to object detection," *IEEE Sensors Journal*, vol. 22, no. 10, pp. 9543–9557, 2022.

[11] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[12] C.-Y. Wang and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024.

[13] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," *arXiv preprint arXiv:2401.17270*, 2024.

[14] C. Kuenzer and S. Dech, "Thermal infrared remote sensing," *Remote Sensing and Digital Image Processing. doi*, vol. 10, no. 1007, pp. 978–94, 2013.

[15] J. Ge, Y. Luo, and G. Tei, "Real-time pedestrian detection and tracking at nighttime for driver-assistance systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 2, pp. 283–298, 2009.

[16] M. Ding, X. Zhang, W.-H. Chen, L. Wei, and Y.-F. Cao, "Thermal infrared pedestrian tracking via fusion of features in driving assistance system of intelligent vehicles," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 233, no. 16, pp. 6089–6103, 2019.

[17] J. A. Sobrino, J. C. Jiménez-Muñoz, G. Sòria, M. Romaguera, L. Guanter, J. Moreno, A. Plaza, and P. Martínez, "Land surface emissivity retrieval from different vnir and tir sensors," *IEEE transactions on geoscience and remote sensing*, vol. 46, no. 2, pp. 316–327, 2008.

[18] J. A. Sobrino, F. Del Frate, M. Drusch, J. C. Jiménez-Muñoz, P. Manunta, and A. Regan, "Review of thermal infrared applications and requirements for future high-resolution sensors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 2963–2972, 2016.

[19] E. Neinavaz, M. Schlerf, R. Darvishzadeh, M. Gerhards, and A. K. Skidmore, "Thermal infrared remote sensing of vegetation: Current status and perspectives," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102415, 2021.

[20] M. Ye, X. Lan, and Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 347–355.

[21] Y. Ling, Z. Zhong, Z. Luo, P. Rota, S. Li, and N. Sebe, "Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 889–897.

[22] B. Wu, Y. Feng, Y. Sun, and Y. Ji, "Feature aggregation via attention mechanism for visible-thermal person re-identification," *IEEE Signal Processing Letters*, 2023.

[23] M. Moradi, A. Chaibakhsh, and A. Ramezani, "An intelligent hybrid technique for fault detection and condition monitoring of a thermal power plant," *Applied Mathematical Modelling*, vol. 60, pp. 34–47, 2018.

[24] X. Hou *et al.*, "Thermal and wet comfort of clothing in different environments based on multidimensional sensor data fusion and intelligent detection," *Journal of Sensors*, vol. 2022, 2022.

[25] M. Kumar, S. Ray, and D. K. Yadav, "Moving human detection and tracking from thermal video through intelligent surveillance system for smart applications," *Multimedia Tools and Applications*, vol. 82, no. 25, pp. 39 551–39 570, 2023.

[26] H. Tonooka and F. D. Palluconi, "Validation of aster/tir standard atmospheric correction using water surfaces," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 12, pp. 2769–2777, 2005.

[27] B. Pan and S. Wang, "Facial expression recognition enhanced by thermal images through adversarial learning," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1346–1353.

[28] S. Li, B. Han, Z. Yu, C. H. Liu, K. Chen, and S. Wang, "I2v-gan: Unpaired infrared-to-visible video translation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3061–3069.

[29] Z. Wang, H. Shen, C. Men, Q. Sun, and K. Huang, "Thermal infrared image inpainting via edge-aware guidance," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[30] Z. Wang, F. Colonnier, J. Zheng, J. Acharya, W. Jiang, and K. Huang, "Tirdet: Mono-modality thermal infrared object detection based on prior thermal-to-visible translation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2663–2672.

[31] F.A.Group, "Flir thermal dataset for algorithm training," *https://www.flir.co.uk/oem/adas/adas-dataset-form/*, 2019.

[32] H. Zhang, E. Fromont, S. Lefevre, and 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.

[33] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3496–3504.

[34] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

[35] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[36] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1209–1218.

[37] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[39] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 280–296.

[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[41] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[43] R. Pi, J. Gao, S. Diao, R. Pan, H. Dong, J. Zhang, L. Yao, J. Han, H. Xu, and L. K. T. Zhang, "Detgpt: Detect what you need via reasoning," *arXiv preprint arXiv:2305.14167*, 2023.

[44] K. Worden and J. M. Dulieu-Barton, "An overview of intelligent fault detection in systems and structures," *Structural Health Monitoring*, vol. 3, no. 1, pp. 85–98, 2004.

[45] Y. Chen, M. Rao, K. Feng, and G. Niu, "Modified varying index coefficient autoregression model for representation of the nonstationary vibration from a planetary gearbox," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[46] T. Han, W. Xie, and Z. Pei, "Semi-supervised adversarial discriminative learning approach for intelligent fault diagnosis of wind turbine," *Information Sciences*, vol. 648, p. 119496, 2023.

[47] Y. Chen, M. Rao, K. Feng, and M. J. Zuo, "Physics-informed lstm hyperparameters selection for gearbox fault detection," *Mechanical Systems and Signal Processing*, vol. 171, p. 108907, 2022.

[48] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," *Applied Energy*, vol. 287, p. 116601, 2021.

[49] J. Yao and T. Han, "Data-driven lithium-ion batteries capacity estimation based on deep transfer learning using partial segment of charging/discharging data," *Energy*, vol. 271, p. 127033, 2023.

[50] H. Zhang, E. Fromont, S. Lefèvre, and 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.

[51] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv:2111.00273*, 2021.

[52] F. Qingyun and W. Zhaokui, "Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery," *Pattern Recognition*, vol. 130, p. 108786, 2022.

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[55] H. Liu, J. Zhang, K. Yang, X. Hu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *arXiv preprint arXiv:2203.04838*, 2022.

[56] K. Chen, J. Liu, and H. Zhang, "Igt: Illumination-guided rgb-t object detection with transformers," *Knowledge-Based Systems*, p. 110423, 2023.

[57] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *International conference on machine learning*. PMLR, 2021, pp. 11 863–11 874.

[58] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[59] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "Fnet: Mixing tokens with fourier transforms," *arXiv preprint arXiv:2105.03824*, 2021.

[60] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.

[61] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," *arXiv preprint arXiv:1911.09516*, 2019.

[62] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[63] Z. Chen, C. Yang, Q. Li, F. Zhao, Z.-J. Zha, and F. Wu, "Disentangle your dense object detector," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4939–4948.

[64] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[65] M. Contributors, "MMCV: OpenMMLab computer vision foundation," https://github.com/open-mmlab/mmcv, 2018.

[66] J.-S. Yun, S.-H. Park, and S. B. Yoo, "Infusion-net: Inter-and intra-weighted cross-fusion network for multispectral object detection," *Mathematics*, vol. 10, no. 21, p. 3966, 2022.

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

[68] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

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

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

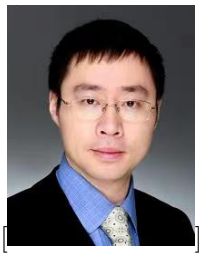
[71] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[72] T. Chu, J. Chen, J. Sun, S. Lian, Z. Wang, Z. Zuo, L. Zhao, W. Xing, and D. Lu, "Rethinking fast fourier convolution in image inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 195–23 205.

[73] F. Wang, S. Wu, W. Zhang, Z. Xu, Y. Zhang, C. Wu, and S. Coleman, "Emotion recognition with convolutional neural network and eeg-based efdms," *Neuropsychologia*, vol. 146, p. 107506, 2020.

Zeyu Wang (S'21) is currently pursuing the Ph.D. degree in College of Information Science & Electronic Engineering at Zhejiang University (ZJU), Hangzhou, China, since 2020. From Dec. 2022 to Jun. 2023, he was a visiting research assistant at the Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include machine learning, thermal infrared vision, remote sensing, and human-computer interface. He has published seven papers in international academic journals and conference proceedings. He also serves as the reviewer of international journals and conferences such as AAAI, ECCV, MICCAI, ICASSP, and IEEE TCASII.

Haibin Shen is currently a Professor with Zhejiang University, a member of the second level of 151 talents project of Zhejiang Province, and a member of the Key Team of Zhejiang Science and Technology Innovation. His research interests include learning algorithms, processor architecture, and modeling. His research achievement has been used by many authority organizations. He has published more than 100 papers on academic journals, and he has been granted more than 30 patents of invention. He was a recipient of the First Prize of Electronic Information Science and Technology Award from the Chinese Institute of Electronics, and has won a second prize at the provincial level.

 Wenyu Jiang (Senior Member, IEEE) received the Ph.D. degree in computer science from Columbia University, New York City, NY, USA, in 2003. From 2003 to 2011, he was with Dolby Laboratories San Francisco and then Dolby Laboratories International Services (Beijing), as a Staff Engineer, and researched on quality of service for streaming media over wireless, P2P-based digital rights management (DRM) friendly content distribution, and multimedia fingerprinting and retrieval. Since 2011, he has been with the Institute for Infocomm Research (I^2R), Agency for Science, Technology, and Research (A*STAR), Singapore. He is currently a Principal Scientist. His research areas span from memory computing-based multimedia retrieval and fuzzy pattern search, signal processing and applications for fiber Bragg gratings (FBG) sensors and thermal imaging sensors to neuromorphic computing, including spiking neural network algorithms and in-memory computing. Now, Dr. Jiang is the Chair of IEEE SSCS Singapore Chapter.

 Kejie Huang (Senior Member, IEEE) received the Ph.D degree from the Department of Electrical Engineering, National University of Singapore, Singapore, in 2014. He has been a principal investigator at College of Information Science & Electronic Engineering, Zhejiang University (ZJU) since 2016. Prior to joining ZJU, he spent five years in the industry including Samsung and Xilinx, two years in the Data Storage Institute, Agency for Science Technology and Research (A*STAR), and another three years in Singapore University of Technology and Design (SUTD), Singapore. He has authored or co-authored more than 80 scientific papers in international peer-reviewed journals and conference proceedings. He holds more than 10 granted patents, and another 20 pending ones. His research interests include hardware acceleration for deep learning, architecture development for post Moore's law era, and computer vision. He is an IEEE Senior Member, the Associate Editor of IEEE TCASII, and the reviewer of many international journals such as Nature Communications, IEEE TCAS, TVLSI, EDL.