TIRDet: Mono-Modality Thermal InfraRed Object Detection Based on Prior Thermal-To-Visible Translation

Zeyu Wang
wangzeyu2020@zju.edu.cn
Zhejiang University
Hangzhou, China

Fabien Colonnier
fabien.colonnier@gmail.com
Agency for Science, Technology and Research, Singapore

Jinghong Zheng
jzheng@i2r.a-star.edu.sg
Agency for Science, Technology and Research, Singapore

Jyotibdha Acharya
acharyaj@i2r.a-star.edu.sg
Agency for Science, Technology and Research, Singapore

Wenyu Jiang*
wjiang@i2r.a-star.edu.sg
Agency for Science, Technology and Research, Singapore

Kejie Huang*
huangkejie@zju.edu.cn
Zhejiang University
Hangzhou, China

ABSTRACT

Cross-modality images that combine visible-infrared spectra can provide complementary information for object detection. In particular, they are well-suited for autonomous vehicle applications in dark environments with limited illumination. However, it is time-consuming to acquire a large number of pixel-aligned visible-thermal image pairs, and real-time alignment is challenging in practical driving systems. Furthermore, the quality of visible-spectrum images can be adversely affected by complex environmental conditions. In this paper, we propose a novel neural network called *TIRDet*, which only utilizes Thermal InfraRed (TIR) images for mono-modality object detection. To compensate for the lacked visible-band information, we adopt a prior Thermal-To-Visible (T2V) translation model to obtain the translated visible images and the latent T2V codes. In addition, we introduce a novel attention-based Cross-Modality Aggregation (CMA) module, which can augment the modality-translation awareness of *TIRDet* by preserving the T2V semantic information. Extensive experiments on FLIR and LLVIP datasets demonstrate that our *TIRDet* significantly outperforms all mono-modality detection methods based on thermal images, and it even surpasses most State-Of-The-Art (SOTA) multispectral methods using visible-thermal image pairs. Code is available at https://github.com/zeyuwang-zju/TIRDet.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**; • **Information systems** → *Information systems applications.*

KEYWORDS

Object detection, mono-modality, Thermal InfraRed (TIR) images, *TIRDet*, Cross-Modality Aggregation (CMA).

*Corresponding authors.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613849

ACM Reference Format:
Zeyu Wang, Fabien Colonnier, Jinghong Zheng, Jyotibdha Acharya, Wenyu Jiang, and Kejie Huang. 2023. TIRDet: Mono-Modality Thermal InfraRed Object Detection Based on Prior Thermal-To-Visible Translation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3613849

1 INTRODUCTION

Object detection, a fundamental task in computer vision, denotes the process of locating object instances in images or videos. Nowadays, the majority of image object detection methods adopt visible-spectrum Red-Green-Blue (RGB) images as the source data [4, 9, 10, 13, 17–19, 31, 45–48]. However, in some adverse environmental conditions (*e.g.*, nighttime, foggy, snowy), RGB images cannot provide high-quality visual information for accurate object detection. In contrast, Thermal InfraRed (TIR) sensors [21, 23] can capture thermal radiation at the wavelength of $0.75-15\mu m$, enabling them to observe objects with temperatures above absolute zero. Consequently, thermal images have shown superior abilities in low-light detection [30, 52], driver-assistance systems [16], person re-identification [27, 54, 63, 64], and other applications [24, 42, 60, 61]. Currently, many multispectral object detection methods utilizing visible-thermal image pairs have been proposed [29, 40, 43, 44, 53, 66, 67]. Visible images are characterized by high chromatic contrast and visual fidelity, while thermal images exhibit rich thermal information and sharp edge contours. The cross-modality source images can provide sufficient and complementary visual information, enhancing the robustness and reliability of multispectral detection systems.

However, current multispectral object detection methods still adopt visible images as part of the source data, which presents potential limitations in the following aspects: (1) The visible images are usually unavailable in rural areas at nighttime due to the limited illumination. (2) Some complex environmental conditions can significantly influence the quality of visible images. (3) It is difficult and time-consuming to collect large-scale aligned visible-thermal image pairs for training multispectral detection algorithms. For example, Zhang *et al.* [66] reported that the FLIR dataset [15] contained many misaligned visible-thermal image pairs despite manual calibration. (4) Although several visible-thermal image registration methods [2, 55, 59] have been proposed, achieving real-time and

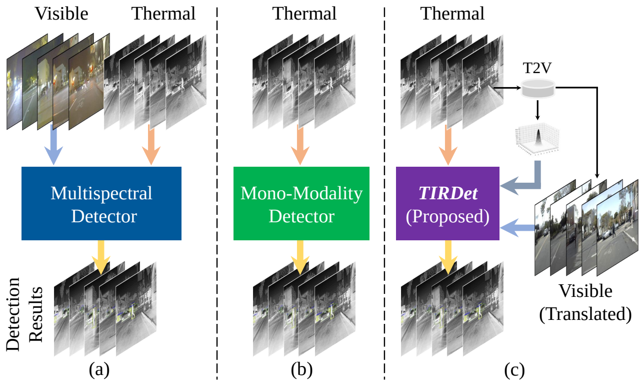

Figure 1: (a) Multispectral object detection methods based on visible-thermal image pairs. (b) Mono-modality object detection methods based on thermal images. (c) The proposed *TIRDet* model based on thermal images, which integrates a prior Thermal-To-Visible (T2V) translation model.

robust alignment remains a significant challenge, which hinders the practical application in modern driving systems. (5) The pre-trained multispectral detection systems may become inefficient once the visible scene changes, such as when switching from daytime to nighttime. Due to these factors, despite their potential to leverage complementary information, current multispectral methods may not be a feasible option for practical night-driving systems.

In this paper, we address the above problems by proposing a novel model called *TIRDet*, which stands for mono-modality Thermal InfraRed object Detector. It only requires thermal images as input and eliminates the need for visible images. In order to compensate for the absence of visible-band information, we employ a pre-trained Thermal-To-Visible (T2V) translation model, *Pearl-GAN* [34], to generate the translated visible images. The generated visible images are concatenated with input thermal images and processed by the *CSPDarknet* backbone [4]. Additionally, we leverage the latent T2V codes by implementing the Cross-Modality Aggregation (CMA), an attention-based module, which enhances the modality-translation awareness of *TIRDet*. Extensive experiments on FLIR [15, 66] and LLVIP [22] datasets demonstrate that *TIRDet* exhibits significant advantages over current mono-modality methods based on thermal images and even outperforms most multispectral methods using visible-thermal image pairs. The ablation study and further investigation confirm the effectiveness of our proposed method based on prior T2V translation and cross-modality fusion.

The contributions of this work can be summarized as follows:

- We point out the main limitations of current multispectral object detection methods and illustrate that mono-modality methods based on thermal images offer greater robustness for low-light autonomous vehicle systems.
- We propose a novel model called *TIRDet* for mono-modality object detection utilizing only thermal images, which incorporates a prior Thermal-To-Visible (T2V) translation model to compensate for the lack of visible-band information.
- Based on the latent T2V codes, we propose an attention-based Cross-Modality Aggregation (CMA) module to augment the modality-translation awareness of our *TIRDet*.
- We conduct experiments on public FLIR and LLVIP datasets, which demonstrate our significant advantages over State-Of-The-Art (SOTA) object detection methods.
- Finally, we discuss the drawbacks of our proposed method and suggest future directions for improvement.

2 RELATED WORKS

2.1 RGB Object Detection

Object detection on RGB images has been extensively studied, with benchmark datasets including Pascal VOC [14], ADE20K [68], and COCO-Stuff [5]. Classical algorithms such as R-CNN [19], Fast R-CNN [18], and Faster R-CNN [48] pioneered the use of deep learning for object detection, followed by one-stage methods including SSD [31] and YOLO-series algorithms [4, 9, 10, 13, 17, 45–47]. In these methods, Convolutional Neural Networks (CNN) are typically used to construct the backbones. Recently, some novel Transformer-based algorithms [6, 25, 69] have explored the use of self-attention [57] in object detection for its global-context interaction.

2.2 Multispectral Object Detection

Multispectral object detection based on visible-thermal image pairs has become a promising research field, particularly in nighttime driver-assistance systems where low-light conditions can impair visibility. Conventional multispectral methods relied on CNN-based models [44, 66, 67], such as VGG [49] and ResNet [20], as their detection backbones. Recent SOTA algorithms, such as the Transformer-based models CMX [28] and IGT [7], have shown potential in fusing cross-modality features of visible-thermal images. Nonetheless, current multispectral methods still adopt visible images as the input source due to their high chromatic contrast under high illumination, which may limit their applicability and efficiency.

2.3 Thermal-To-Visible Translation

Thermal-To-Visible (T2V) translation, also known as thermal image colorization, has recently garnered considerable attention for its capacity in user-friendly driving systems and human-computer interaction [37]. Most studies focus on traffic-scene images [3, 33–35, 41, 50, 51] and human-face images [36, 38, 39]. Early works treated T2V translation as a pixel-to-pixel mapping problem and used deep neural networks for supervised learning [3, 50, 51]. However, subsequent studies pointed out that the visible-thermal image pairs in public datasets usually lack precise pixel-level alignment, necessitating the use of unsupervised learning for T2V translation [41]. Recent works have also explored the connection between T2V translation and night-to-day translation [33–35].

3 METHODS

3.1 TIRDet—Overview

Figure 2 (a) illustrates the pipeline of our proposed *TIRDet*, which only uses the thermal image I_t as the input source. To complement the information in the visible band, we use a pre-trained T2V translation model, *Pearl-GAN* [34], to convert I_t into a fake but realistic

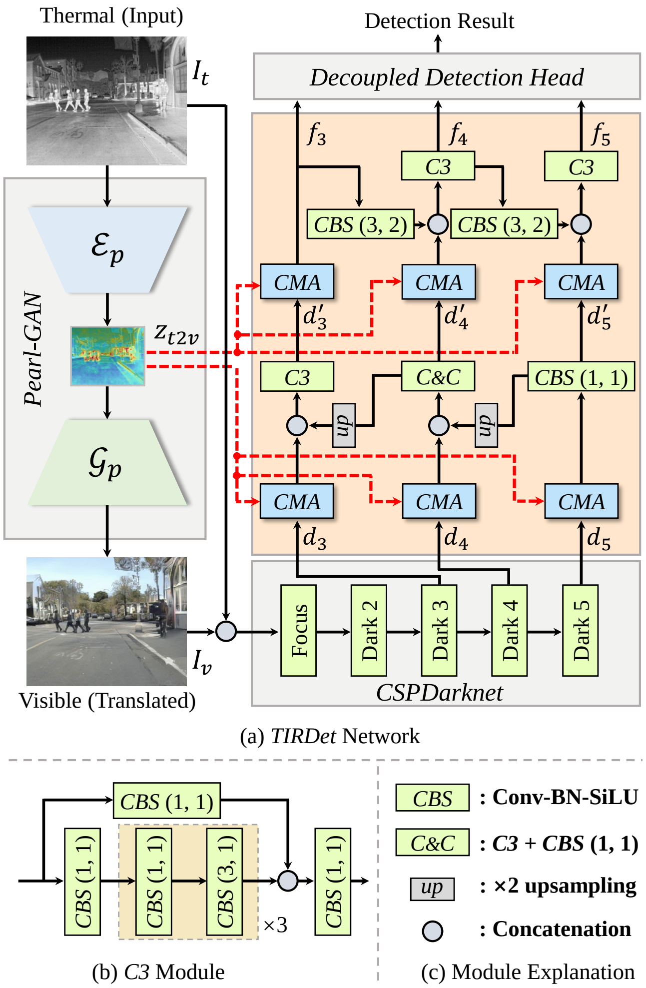

Figure 2: (a) Illustration of the pipeline of our *TIRDet* network. (b) Illustration of the *C3* module. (c) Explanations of the modules in *TIRDet*. "*CBS* (*m*, *n*)" denotes the Conv-BN-SiLU [12] with the kernel size of *m* and the stride of *n*.

visible image I_v. Meanwhile, we extract the latent T2V code z_{t2v}, which corresponds to a deep feature map in *Pearl-GAN*.

$$I_v, z_{t2v} = \textit{Pearl-GAN}(I_t), \tag{1}$$

where $I_t \in \mathbb{R}^{H\times W\times 1}$, $I_v \in \mathbb{R}^{H\times W\times 3}$, and $z_{t2v} \in \mathbb{R}^{h_z\times w_z\times c_z}$.

After that, we adopt *CSPDarknet* [4] as the backbone to extract multiscale features from the concatenated I_t and I_v.

$$d_3, d_4, d_5 = \textit{CSPDarknet}([I_t, I_v]), \tag{2}$$

where $[\cdot]$ denotes the concatenation process. The output feature $d_i (i \in \{3, 4, 5\})$ holds the spatial size of $H/2^i \times W/2^i$.

To augment the cross-modality awareness of *TIRDet*, we implement the novel Cross-Modality Aggregation (CMA) modules, which use the fusion-attention mechanism based on the latent T2V code z_{t2v}. The detection neck, a convolution-based network that incorporates our CMA modules, utilizes d_i $(i \in \{3, 4, 5\})$ to generate the corresponding output features f_i $(i \in \{3, 4, 5\})$.

$$f_3, f_4, f_5 = \textit{Detection-Neck}(d_3, d_4, d_5, z_{t2v}). \tag{3}$$

Finally, the Decoupled Detection Head [17] is employed to obtain the detection result based on $f_i (i \in \{3, 4, 5\})$.

$$\textit{Result} = \textit{Detection-Head}(f_3, f_4, f_5). \tag{4}$$

In this work, we implement our proposed model in three variations, namely *TIRDet-S*, *TIRDet-M*, and *TIRDet-L*, according to the varying depths and widths of the *CSPDarknet* backbones.

3.2 Prior Thermal-To-Visible Translation

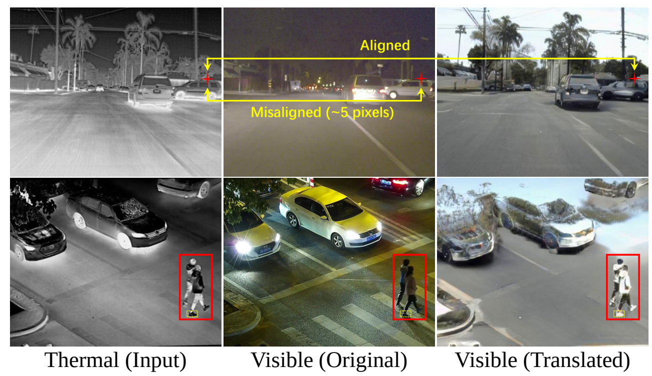

Figure 3: Examples of Thermal-To-Visible (T2V) translation on FLIR (top) and LLVIP (bottom) by *Pearl-GAN*.

As shown in Figure 3, although the public "aligned-version" FLIR dataset [66] has undergone manual registration, a misalignment of around 5 pixels still exists in the visible-thermal image pair. To address this issue, we aim to remove the need for visible images in thermal infrared object detection. In specific, our *TIRDet* integrates the pre-trained T2V translation model, *Pearl-GAN*, which aims to transform a nighttime thermal image I_t into a daytime visible image I_v (illustrated in the last column of Figure 3). The pre-trained model weights of *Pearl-GAN* on FLIR dataset are publicly available, which we also adopt for use on LLVIP dataset. Compared with current multispectral detection methods requiring visible-thermal image pairs, our proposed method eliminates the need for image registration and avoids misalignment through the prior T2V translation.

The adopted *Pearl-GAN* is based on the backbone of ToDayGAN model [1], which also incorporates top-down attention and gradient alignment. It comprises the convolution-based encoder $\mathcal{E}_p$ and decoder $\mathcal{G}_p$. In detail, $\mathcal{E}_p$ takes I_t as the input to generate the latent T2V code z_{t2v}, which is then decoded by $\mathcal{G}_p$ to derive I_v.

$$z_{t2v} = \mathcal{E}_p(I_t), \quad z_{t2v} \in \mathbb{R}^{h_z\times w_z\times c_z}, \tag{5}$$

$$I_v = \mathcal{G}_p(z_{t2v}), \quad I_v \in \mathbb{R}^{H\times W\times 3}. \tag{6}$$

After the prior T2V translation, the translated I_v is concatenated with I_t and fed as input to the *CSPDarknet* backbone. Meanwhile, the encoder $\mathcal{E}_p$ also operates as a deep-level feature extractor during the T2V translation process. Therefore, we employ the latent z_{t2v} to modulate the internal features in our novel CMA modules. In addition, it is worth noting that the weights of *Pearl-GAN* were quantized from *fp32*-precision to *fp16*-precision during the experiments to decrease additional memory cost.

3.3 CSPDarknet Backbone

Due to the strong feature-extraction capability, *CSPDarknet* has been widely adopted as the backbone of many YOLO-series detection models [4, 13, 17]. It consists of one Focus module and four convolution-based Dark modules (details can be found in [4]), which effectively extract in-depth information from the fused images I_t and I_v. The output features of Dark3, Dark4, and Dark5 (represented as d_3, d_4, and d_5, respectively) are fed into the detection neck, as formulated in Eq. (3). The model size of *CSPDarknet* can be adjusted by changing the depths and widths of the four Dark modules.

Specifically, the Focus acts as the stem module of *CSPDarknet*, which conducts the spatial-order down-sampling with the stride of 2. In our method, the Focus module takes the concatenated I_t and I_v with the form of "R-G-B-T" as input, as illustrated in Figure 4. In this way, the cross-modality Focus module can fuse and correlate low-level information in visible and thermal domains.

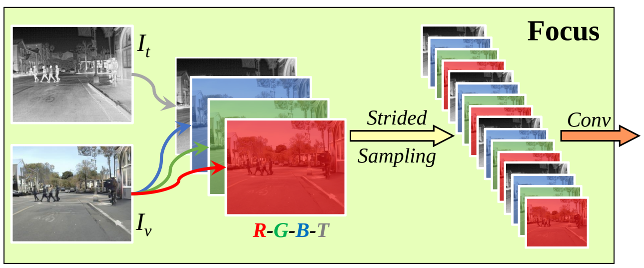

Figure 4: Illustration of the Focus module, which takes the concatenated I_t and I_v with the form of "R-G-B-T" as input.

3.4 Cross-Modality Aggregation (CMA)

As mentioned in Sect. 3.2, the encoder $\mathcal{E}_p$ of *Pearl-GAN* performs the deep feature extraction during the T2V translation. To preserve in-depth information on this process, we propose the novel CMA module to enhance the cross-modality awareness of our model. It leverages the latent T2V code z_{t2v} obtained from *Pearl-GAN* to modulate the feature F_{in} via the fusion-attention mechanism.

$$F_{out} = \mathrm{CMA}(F_{in}, z_{t2v}), \tag{7}$$

where the input F_{in} could be the output feature d_i ($i \in \{3,4,5\}$) of the *CSPDarknet* backbone or the internal feature $d_i^{'}$ ($i \in \{3,4,5\}$) inside the detection neck.

The detailed workflow of the CMA module is illustrated in Figure 5 (a) and can be described as follows: Based on the latent T2V code $z_{t2v} \in \mathbb{R}^{\mathrm{h_z \times w_z \times c_z}}$ extracted from *Pearl-GAN*, we apply a bicubic-interpolation *Resize* process with a 1×1 convolution (denoted as ω_c as a whole) to obtain a feature with the same shape as $F_{in} \in \mathbb{R}^{\mathrm{h_f \times w_f \times c_f}}$. Then, we fuse the obtained feature $\omega_c(z_{t2v})$ with F_{in} via the element-wise multiplication to generate the cross-modality feature F_{cross}, which can be formulated as:

$$F_{cross} = F_{in} \otimes (\omega_c(z_{t2v})), \tag{8}$$

where the fused feature $F_{cross} \in \mathbb{R}^{\mathrm{h_f \times w_f \times c_f}}$.

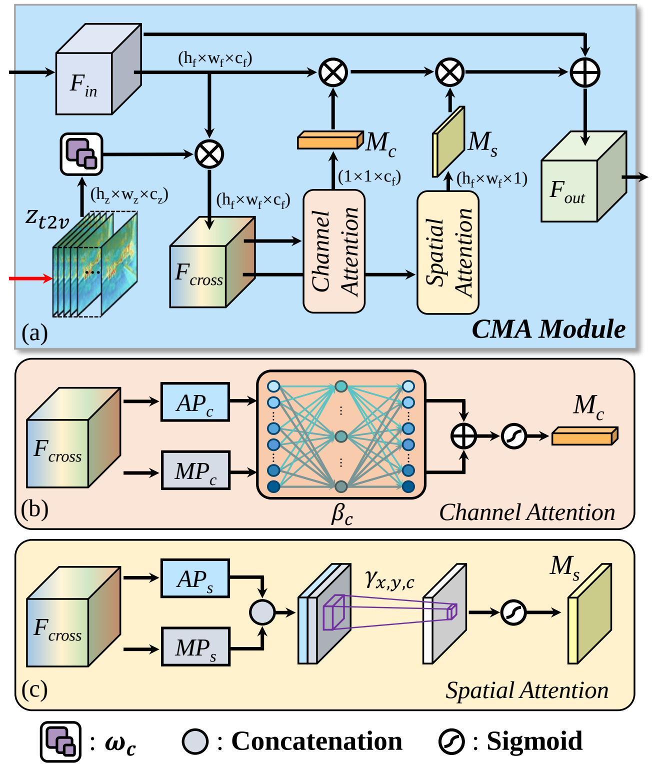

Figure 5: (a) Illustration of the Cross-Modality Aggregation (CMA) module. (b) Illustration of Channel Attention. (c) Illustration of Spatial Attention. $\oplus$ and $\otimes$ denote element-wise addition and element-wise multiplication, respectively.

Later, we apply the fusion-attention mechanism on F_{cross}, which consists of the Channel Attention (CA) (Figure 5 (b)) and the Spatial Attention (SA) (Figure 5 (c)), to obtain the attention maps $M_c \in \mathbb{R}^{1 \times 1 \times \mathrm{c_f}}$ and $M_s \in \mathbb{R}^{\mathrm{h_f \times w_f} \times 1}$, respectively. This is inspired by Convolutional Block Attention Module (CBAM) [62], a simple but effective attention module in computer vision. In our proposed CMA modules, we adopt the CA and SA structures to fully preserve the cross-modality information in F_{cross}. The generation of M_c and M_s can be formulated in Eq. (9) and Eq. (10), respectively.

$$M_c = \sigma(\beta_c(AP_c(F_{cross})) + \beta_c(MP_c(F_{cross}))), \tag{9}$$

$$M_s = \sigma(\gamma_{x,y,c}([AP_s(F_{cross}), MP_s(F_{cross})])), \tag{10}$$

where σ denotes the sigmoid activation; β_c represents a channel-wise Multi-Layer Perceptron (MLP); $\gamma_{x,y,c}$ refers to a 7×7 convolutional layer; $AP_{c/s}$ and $MP_{c/s}$ denote the Average-Pooling and Max-Pooling along the channel/spatial direction, respectively. In Eq. (9), the adopted β_c applied on $AP_c(F_{cross})$ and $MP_c(F_{cross})$ shares the model weights, which consists of two fully connected layers with the intermediate ReLU activation function.

After that, we fuse the input feature F_{in} with the generated attention maps M_c and M_f via element-wise multiplication in sequential order. Finally, we obtain F_{out} with a skip connection from F_{in}.

$$F_{out} = (F_{in} \otimes M_c) \otimes M_s + F_{in}, \tag{11}$$

where the output feature $F_{out} \in \mathbb{R}^{\mathrm{h_f} \times \mathrm{w_f} \times \mathrm{c_f}}$.

As shown in Figure 2 (a), through the multiscale attention-based CMA modules in the detection neck, our *TIRDet* efficiently preserves the in-depth semantic information of the T2V translation process. Furthermore, the consecutive modulation with the latent T2V code z_{t2v} greatly enhances the cross-modality awareness of our model in the absence of real visible images.

4 EXPERIMENTS

4.1 Implementation Details

We implement the models using Pytorch 1.9.1 on Intel Xeon CPU E5-2696 v4 @ 2.20GHz and NVIDIA RTX 2080Ti GPUs with CUDA 11.5. We reproduce all mono-modality baseline methods using MMDetection [8, 11] or their official public repositories. For the multispectral baseline methods, we also reproduce them if the implementation codes are publicly available. To train our *TIRDet*, we use the Stochastic Gradient Descent (SGD) optimizer with the initial learning rate of 0.01, the momentum of 0.9, and the weight decay of 0.0005 for 300 epochs. Meanwhile, we adopt the loss functions and data augmentation strategies used in [17], while we close the data augmentation in the last 15 epochs. The batch sizes are set to 8, 4, and 2 for training our *TIRDet-S*, *TIRDet-M*, and *TIRDet-L*, respectively. The manual seed is set to 0 for all experiments in this work.

4.2 Datasets

Table 1: Dataset characteristics of FLIR and LLVIP.

Dataset	Label			Image	
	person	car	bicycle	train	test
FLIR [15]	✓	✓	✓	8,862	1,366
"Aligned" FLIR [66]	✓	✓	✓	4,129	1,013
LLVIP [22]	✓	-	-	12,025	3,463

FLIR is a multispectral road-scene dataset that contains 8,862 training images and 1,366 testing images. It has three annotated categories: "person", "bicycle", and "car". Zhang *et al.* [66] reported that it contained many misaligned visible-thermal image pairs and manually removed them. To ensure a fair comparison, we adopt the "aligned-version" FLIR dataset proposed in [66] to conduct the experiments like previous works [7, 28, 43, 44].

LLVIP is a recently released multispectral dataset for low-light pedestrian detection, which contains 12,025 training images and 3,463 testing images. The majority of the images were captured under extremely low-light conditions. Additionally, each image pair has been aligned in space to ensure precise registration.

4.3 Evaluation Metrics

We adopt the standard quantitative metrics in object detection for evaluation, including mean Average Precision (mAP), mAP_{50}, mAP_{75} and mean Average Recall (mAR). The metrics mAP and mAR are evaluated as the mean values of all categories at Intersection over Union (IoU) = 0.50 : 0.05 : 0.95, while mAP_{50} and mAP_{75} are calculated at the IoU thresholds of 0.50 and 0.75, respectively. In most studies, mAP is considered the primary evaluation metric.

Table 2: Quantitative results (%) on FLIR dataset.

Model	Data	Backbone	mAP_{50}	mAP_{75}	mAP	mAR
Comparison with Mono-Modality (Thermal) Methods						
Faster RCNN [48]	T	ResNet50	74.4	32.5	37.6	49.7
SSD [31]	T	VGG16	65.5	22.4	29.6	44.3
RetinaNet [26]	T	ResNet50	64.5	20.3	28.3	44.4
YOLOv3 [47]	T	Darknet53	73.6	31.3	36.8	46.5
YOLOv5 [13]	T	CSPD53	73.9	35.7	39.5	47.3
YOLOF [9]	T	ResNet50	74.9	26.7	34.6	47.9
DDOD [10]	T	ResNet50	72.7	26.2	33.9	48.2
YOLOX-L [17]	T	CSPD53	80.9	37.5	42.0	52.2
YOLOv7 [58]	T	E-ELAN	75.6	32.2	38.2	49.0
TIRDet-L	T	CSPD53	**81.4**	**41.1**	**44.3**	**54.0**
Comparison with Multispectral (Visible+Thermal) Methods						
CFR_3 [66]	V+T	VGG16	72.4	-	-	-
GAFF [67]	V+T	ResNet18	72.9	32.9	37.5	-
GAFF [67]	V+T	VGG16	72.7	30.9	37.3	-
YOLOFusion [44]	V+T	VGG16	76.6	-	39.8	-
CFT [43]	V+T	CFB	78.7	35.5	40.2	52.5
InfusionNet [65]	V+T	Infusion	79.1	35.2	40.3	-
CMX [28]	V+T	SwinT	82.2	37.1	42.3	-
IGT [7]	V+T	SwinT	**85.0**	36.9	43.6	-
TIRDet-L	T	CSPD53	81.4	**41.1**	**44.3**	**54.0**

Abbreviations in this table (also used in Table 3): CSPD53 (CSPDarknet-53) [4], CFB (Cross-Modality Fusion Backbone) [43], SwinT (Swin Transformer) [32].

Table 3: Quantitative results (%) on LLVIP dataset.

Model	Data	Backbone	mAP_{50}	mAP_{75}	mAP	mAR
Comparison with Mono-Modality (Thermal) Methods						
Faster RCNN [48]	T	ResNet50	96.1	68.5	61.1	59.7
SSD [31]	T	VGG16	90.2	57.9	53.5	57.3
RetinaNet [26]	T	ResNet50	93.7	49.3	50.9	60.6
YOLOv3 [47]	T	Darknet53	89.7	53.4	52.8	61.7
YOLOv5 [13]	T	CSPD53	94.6	72.2	61.9	59.5
YOLOF [9]	T	ResNet50	91.4	43.7	47.5	58.6
DDOD [10]	T	ResNet50	94.3	59.9	56.6	63.7
YOLOX-L [17]	T	CSPD53	95.7	71.5	62.3	68.3
YOLOv7 [58]	T	E-ELAN	95.5	67.7	59.4	62.5
TIRDet-L	T	CSPD53	**96.3**	**73.1**	**64.2**	**69.4**
Comparison with Multispectral (Visible + Thermal) Methods						
YOLOv5-VT [13]	V+T	CSPD53	95.8	71.4	62.3	63.1
CFT [43]	V+T	CFB	97.5	72.9	63.6	68.4
InfusionNet [65]	V+T	Infusion	**98.6**	**73.3**	**64.6**	-
TIRDet-L	T	CSPD53	96.3	73.1	64.2	**69.4**

4.4 Quantitative Comparison

We choose *TIRDet-L* to compare with the baseline methods. As shown in Table 2 and Table 3, our *TIRDet-L* achieves 44.3% mAP and 54.0% mAR on FLIR dataset, and achieves 64.2% mAP and 69.4% mAR on LLVIP dataset. It outperforms all mono-modality methods based on thermal images, especially on the challenging FLIR dataset. The results demonstrate the significance of the visible-band information obtained from the prior T2V translation. In addition,

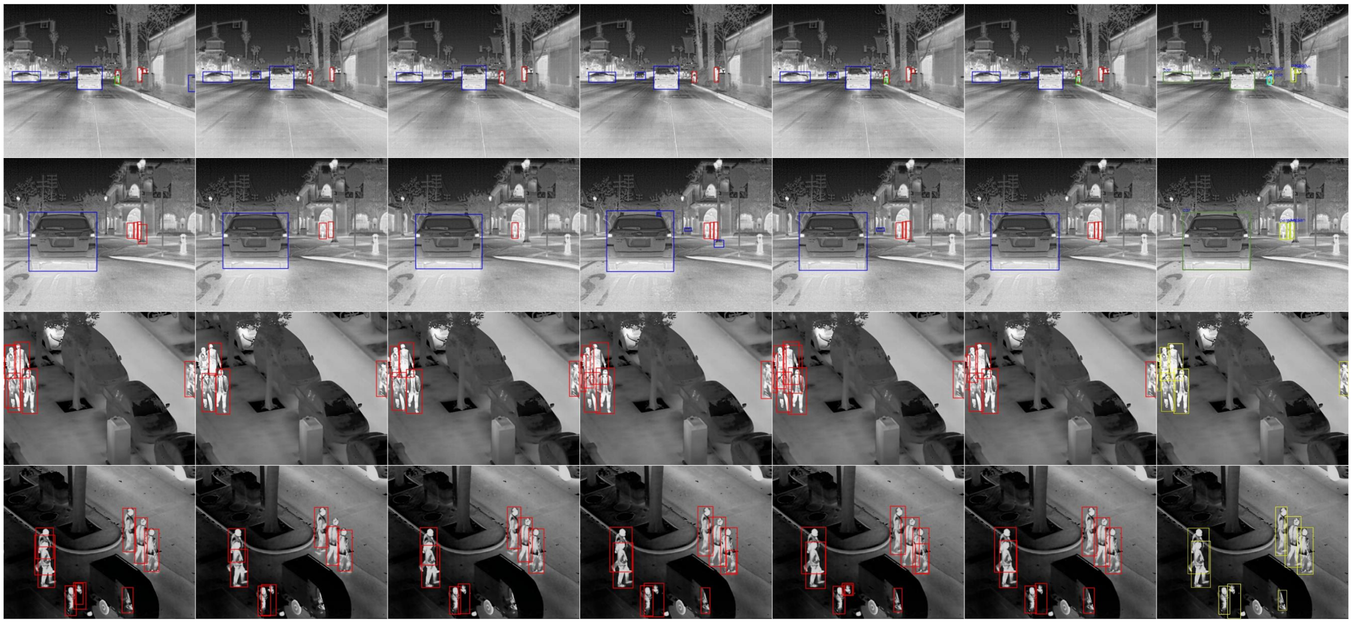

Figure 6: Qualitative comparison between the mono-modality baseline methods and our *TIRDet-L*, where the red, blue, and green boxes denote the detected objects "person", "car", and "bicycle", respectively. We use different colors for the bounding boxes in Label to distinguish them from the detection results. Better viewed in color and zoomed in.

compared with the multispectral methods, our method still achieves the best mAP_{75} and mAP on FLIR, and the second-best mAP_{75} and mAP on LLVIP, even though the visible images are absent from the input. Although the SOTA multispectral method IGT [7] achieves 85.0% mAP_{50} on FLIR dataset, our *TIRDet-L* obtains higher mAP_{75} and mAP scores, outperforming IGT by 4.2% and 0.7%, respectively. This suggests that additional visible images may become interfering factors for multispectral detection under certain conditions. For instance, if the visible image is captured under extremely dark conditions, it cannot provide valuable visual information and may even negatively influence the detection performance. In contrast, our approach based on prior T2V translation compensates for the absence of visible-band information while avoiding the misalignment of visible-thermal image pairs, ensuring its practicability in modern driver-assistance systems.

4.5 Qualitative Comparison

Figure 6 displays the qualitative results of our *TIRDet-L* and mono-modality baseline methods, all of which only employ thermal images for object detection. The results are obtained at the confidence score [45] threshold of 0.5. The comparison indicates that all methods can detect common objects, such as the cars close to the cameras and the pedestrians in the center of the images. However, our *TIRDet-L* outperforms most of the baselines in identifying challenging instances, such as the tiny bicycle shown in the first row of images. Conversely, some of the baseline methods, like YOLOX-L and YOLOv7, fail to detect complex objects and misclassify small instances as cars, as seen in the second row of images. Although the latest method YOLOv7 achieves good quantitative results on LLVIP dataset, it generates some redundant boxes for pedestrian detection. In contrast, our *TIRDet-L* distinguishes complex objects with precision and produces unambiguous detected bounding boxes on the two datasets. In all, our proposed method incorporating prior T2V translation exhibits better performance and robustness for mono-modality thermal infrared object detection.

On the other hand, we select CFT [43] as the representative multispectral method for qualitative comparison, as it has released the implementation codes that enable reproducibility. The comparison in Figure 8 reveals that *TIRDet-L* is more effective in detecting uncommon objects, such as the bicycle wheel in the first image. Conversely, CFT excels at detecting distant objects, such as the two small cars, which are not even labeled. This advantage of CFT could be attributed to the additional use of visible images, which provide high chromatic contrast even for small objects. Meanwhile, both methods perform well for pedestrian detection in the second image on LLVIP dataset. To summarize, both CFT and *TIRDet-L* are capable of detecting challenging objects and exhibit their unique advantages. Nevertheless, our *TIRDet-L* stands out in terms of practicality, as it solely uses thermal images as its input source.

4.6 Ablation Study

We conduct the ablation study on our *TIRDet-L* to demonstrate the effectiveness of T2V translation and CMA modules. The results in Table 4 show that the lack of I_v has the largest negative impact on the performance, as evidenced by the drop of 2.1% and 1.7% in mAP scores on FLIR and LLVIP, respectively. It illustrates that the translated visible image takes a critical role in feature extraction within the deep backbone model. Furthermore, the multiscale CMA

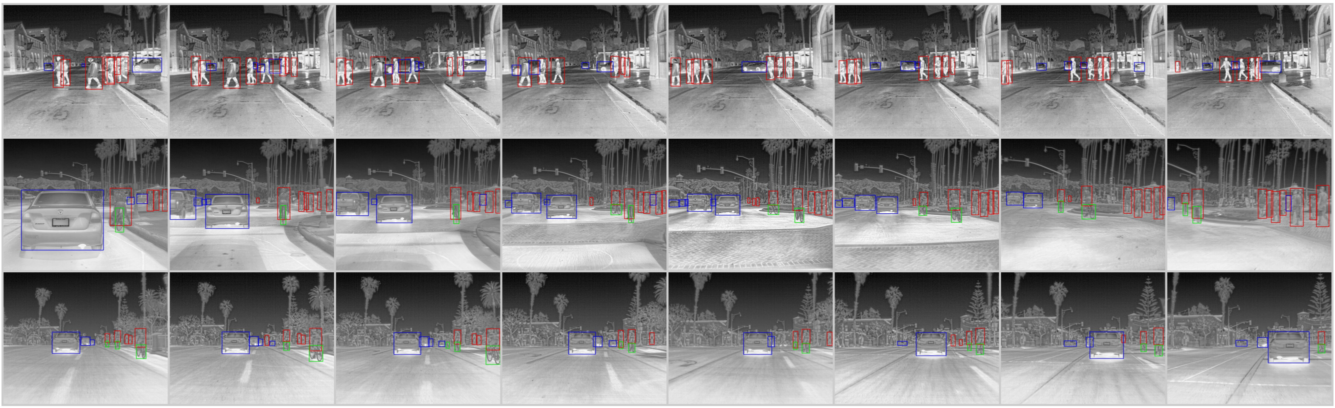

Figure 7: Qualitative results of our proposed *TIRDet-L* on FLIR dataset [15]. Better viewed in color and zoomed in.

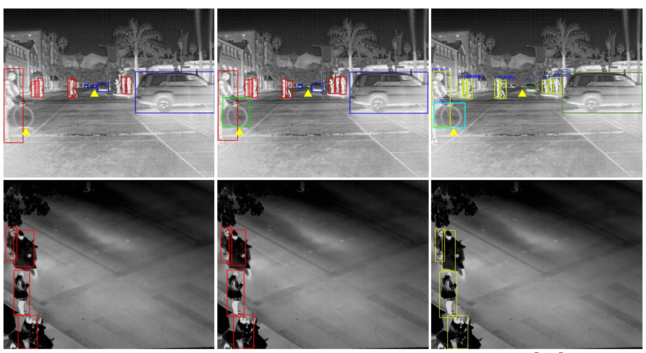

Figure 8: Qualitative comparison between the multispectral method CFT [43] and *TIRDet-L*. The yellow triangles denote places of emphasis. Better viewed in color and zoomed in.

Table 4: Ablation study on FLIR and LLVIP datasets.

Dataset	Model	mAP (%)	mAR (%)
FLIR	***TIRDet-L***	**44.3**	**54.0**
	w/o I_v (CSPD)	42.2 (↓2.1)	52.7 (↓1.3)
	w/o CMA [1-3]	42.8 (↓1.5)	53.1 (↓0.9)
	w/o CMA [4-6]	43.3 (↓1.0)	53.5 (↓0.5)
	w/o CMA [All]	42.5 (↓1.8)	53.0 (↓1.0)
	w/o CA (CMA)	43.2 (↓1.1)	53.3 (↓0.7)
	w/o SA (CMA)	42.9 (↓1.4)	53.2 (↓0.8)
LLVIP	***TIRDet-L***	**64.2**	**69.4**
	w/o I_v (CSPD)	62.5 (↓1.7)	68.5 (↓0.9)
	w/o CMA [1-3]	63.1 (↓1.1)	68.6 (↓0.8)
	w/o CMA [4-6]	63.4 (↓0.8)	68.8 (↓0.6)
	w/o CMA [All]	62.9 (↓1.3)	68.6 (↓0.8)
	w/o CA (CMA)	63.8 (↓0.4)	69.0 (↓0.4)
	w/o SA (CMA)	63.0 (↓1.2)	68.4 (↓1.0)

CMA [1-3] denote the three CMA modules which take d_3, d_4, and d_5 as input, respectively, while CMA [4-6] denote the other three in Figure 2 (a).

modules exhibit varying degrees of significance on *TIRDet*, with a greater influence on FLIR dataset. The absence of all CMA modules results in a reduction of mAP and mAR scores by 1.8% and 1.0% on FLIR dataset, respectively, indicating that the CMA modules are more effective in multiple-category detection. In addition, from the results of "w/o CA" and "w/o SA", we can see that SA has a larger influence than CA, potentially due to the preservation of rich semantic information from the cross-modality feature maps. There is an exception that on LLVIP dataset, the variant "w/o SA" achieves the mAR score of 68.4%, which is 0.2% lower than that of "w/o CMA [All]". To conclude, both the T2V translation and cross-modality fusion techniques are effective in improving the performance of our model, particularly in multiple-category detection scenarios.

5 DISCUSSION

5.1 Cross-Modality Feature Visualization

To further investigate the effectiveness of our proposed CMA modules, we conduct inference on the two images shown in Figure 9. Meanwhile, we visualize the latent T2V codes z_{t2v} and the cross-modality features F_{cross} in the six CMA modules. Figure 9 (c), (e), and (g) denote F_{cross} in CMA modules which take d_3, d_4, and d_5 as input, respectively; (d), (f), and (h) denote F_{cross} in CMA modules which take d_3', d_4', and d_5' as input, respectively. The visualization of z_{t2v} in Figure 9 (b) reveals that the pre-trained *Pearl-GAN* encoder can extract critical semantic information from thermal images, including the prominent object instances and the edge contours. Additionally, the cross-modality features F_{cross} display their distinct areas of focus. In specific, F_{cross} in (c-d) highlight the thermal information across entire images, while those in (e-h) concentrate on the target object regions, such as the people and cars. By utilizing the multiscale CMA modules, our proposed *TIRDet* successfully leverages global and local information in thermal images to enhance its modality-translation awareness.

5.2 t-SNE Visualization

We employ t-SNE [56] to visualize the feature distribution inside the CMA modules. Specifically, we randomly select 100 images from FLIR dataset and map the features F_{in}, F_{cross}, F_{out}, and z_{t2v} onto the two-dimensional space, as illustrated in Figure 10. The

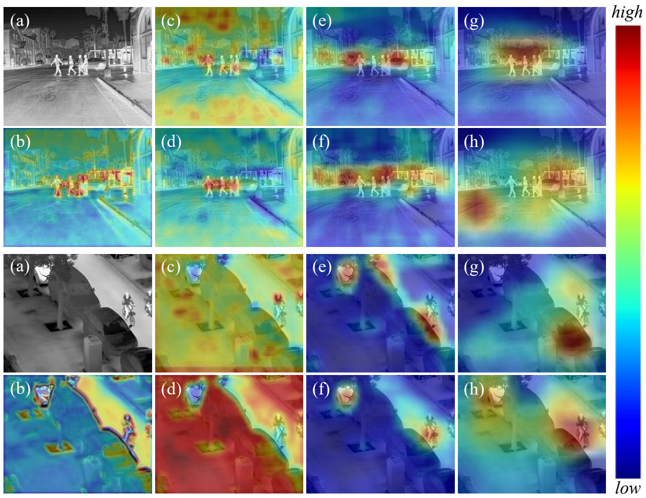

Figure 9: Visualization of the feature maps in CMA modules. (a) Input thermal images I_t. (b) Latent T2V codes z_{t2v}. (c-h) Cross-Modality Features F_{cross} in the six CMA modules.

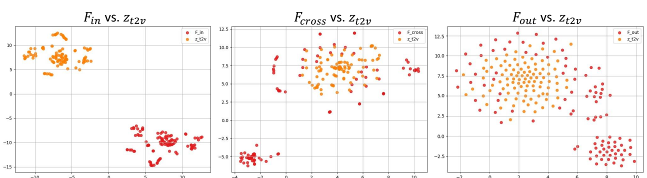

Figure 10: t-SNE visualization of $\{F_{in}, F_{cross}, F_{out}\}$ vs. z_{t2v}.

visualization indicates that F_{in} and z_{t2v} are located in two separate clusters, and the element-wise fusion substantially enhances the proximity between F_{cross} and z_{t2v}. Furthermore, the CMA modules with fusion-attention mechanism produce F_{out} with a cluster that closely aligns with the distribution of z_{t2v}. Nevertheless, we also observe two clusters in F_{cross} and F_{out} at a significant distance from z_{t2v}, which may suggest the information branch from F_{in}.

5.3 Selection of T2V Translation Model

Although there have been previous works on Thermal-To-Visible (T2V) translation, we select *Pearl-GAN* as the prior T2V translation model for the following reasons. Firstly, it employs an unpaired training method (unsupervised learning), eliminating the need for pixel-level aligned visible-thermal image pairs. Secondly, it is primarily trained on traffic-scene images, making it particularly effective for object detection in this work. Lastly, *Pearl-GAN* is the latest T2V translation model that offers open-source code and model weights, making it an accessible choice for our research.

5.4 Original vs. Translated

We analyze the characteristics of translated visible images (I_v) and the original visible images, as shown in Table 5. It reveals that the translated visible images generally exhibit lower illumination levels but a 53.0% increase in chromatic contrast. On the other hand, the entropy of the translated images is larger on the three color channels (R-G-B), indicating that they contain richer chromatic information, especially when compared to the nighttime low-light visible images in the original dataset.

Table 5: Visible image characteristics on FLIR Dataset.

Visible	Chromatic Features		Entropy		
	Luminance	Contrast	Red	Green	Blue
Original	**158.52**	37.29	6.77	6.88	6.85
Translated	125.88	**57.06**	**7.17**	**7.13**	**7.01**

5.5 Limitations and Future Expectations

Table 6: Comparison between YOLOX and *TIRDet*.

Model	Results [FLIR]		Results [LLVIP]		Efficiency	
	mAP	mAR	mAP	mAR	Params	FPS
YOLOX-S	40.7%	51.0%	60.8%	66.4%	8.94M	86.72
YOLOX-M	41.9%	51.8%	61.6%	67.0%	25.28M	69.10
YOLOX-L	42.0%	52.2%	62.3%	68.3%	54.15M	51.40
TIRDet-S	41.7%	51.4%	63.4%	68.1%	9.53M	47.31
TIRDet-M	43.9%	53.3%	63.8%	68.8%	26.33M	35.28
TIRDet-L	**44.3%**	**54.0%**	**64.2%**	**69.4%**	55.77M	28.31

FPS is tested at 640×640 resolution with batch = 1 on NVIDIA RTX 2080Ti GPUs.

As shown in Table 6, although our *TIRDet* variants comprehensively outperform *YOLOX* variants using the same *CSPDarknet* backbone, our model sizes and efficiency pose a disadvantage. Specifically, we observe the decrease in Frames Per Second (FPS) by 44.9% - 48.9% due to the complex computations required by the *Pearl-GAN* model. Therefore, we expect the future development of lightweight and robust T2V translation models, as well as more efficient techniques for cross-modality fusion.

6 CONCLUSION

This paper introduces a novel neural network, *TIRDet*, which only uses thermal images for object detection. Compared with current multispectral methods, our approach eliminates the dependence on visible images and compensates for the lack of visible-band information through the prior T2V translation. To enhance its cross-modality awareness, we introduce an attention-based CMA module that fully preserves the T2V semantic information. The quantitative comparison shows that our *TIRDet-L* outperforms all mono-modality methods and approaches the performance of SOTA multispectral methods. Furthermore, the qualitative results demonstrate that it can accurately detect complex instances in thermal images. Additionally, the ablation study highlights the contributions of the prior T2V translation and proposed CMA modules. In the future, we look forward to developing lightweight and robust T2V translation models and efficient modality-modulation methods.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program of China, Grant Number: 2020AAA0109002, and National Natural Science Foundation of China, Grant Number: 62274142 & U19B2043.

REFERENCES

[1] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. 2019. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 5958–5964.

[2] P Bamrungthai and P Wongkamchang. 2020. A novel method for camera calibration and image alignment of a thermal/visible image fusion system. In *Fourth International Conference on Photonics Solutions (ICPS2019)*, Vol. 11331. SPIE, 110–116.

[3] Amanda Berg, Jorgen Ahlberg, and Michael Felsberg. 2018. Generating visible spectrum images from thermal infrared. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1143–1152.

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

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1209–1218.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. 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, 213–229.

[7] Keyu Chen, Jinqiang Liu, and Han Zhang. 2023. IGT: Illumination-guided RGB-T object detection with transformers. *Knowledge-Based Systems* (2023), 110423.

[8] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155* (2019).

[9] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. 2021. You only look one-level feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13039–13048.

[10] Zehui Chen, Chenhongyi Yang, Qiaofei Li, Feng Zhao, Zheng-Jun Zha, and Feng Wu. 2021. Disentangle your dense object detector. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4939–4948.

[11] MMCV Contributors. 2018. MMCV: OpenMMLab Computer Vision Foundation. https://github.com/open-mmlab/mmcv.

[12] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks* 107 (2018), 3–11.

[13] Glenn Jocher et. al. 2021. *ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support.* https://doi.org/10.5281/zenodo.5563715

[14] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision* 111 (2015), 98–136.

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

[16] Junfeng Ge, Yupin Luo, and Gyomei Tei. 2009. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems. *IEEE Transactions on Intelligent Transportation Systems* 10, 2 (2009), 283–298.

[17] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. YOLOX: Exceeding YOLO Series in 2021. *arXiv preprint arXiv:2107.08430* (2021).

[18] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[21] Stephanos Ioannou, Vittorio Gallese, and Arcangelo Merla. 2014. Thermal infrared imaging in psychophysiology: potentialities and limits. *Psychophysiology* 51, 10 (2014), 951–963.

[22] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. 2021. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3496–3504.

[23] Claudia Kuenzer and Stefan Dech. 2013. Thermal infrared remote sensing. *Remote Sensing and Digital Image Processing. doi* 10, 1007 (2013), 978–94.

[24] Shuang Li, Bingfeng Han, Zhenjie Yu, Chi Harold Liu, Kai Chen, and Shuigen Wang. 2021. I2v-gan: Unpaired infrared-to-visible video translation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3061–3069.

[25] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. 2022. Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*. Springer, 280–296.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[27] Yongguo Ling, Zhun Zhong, Zhiming Luo, Paolo Rota, Shaozi Li, and Nicu Sebe. 2020. 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*. 889–897.

[28] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. 2022. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838* (2022).

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

[30] Qiao Liu, Xin Li, Zhenyu He, Chenglong Li, Jun Li, Zikun Zhou, Di Yuan, Jing Li, Kai Yang, Nana Fan, et al. 2020. LSOTB-TIR: A large-scale high-diversity thermal infrared object tracking benchmark. In *Proceedings of the 28th ACM international conference on multimedia*. 3847–3856.

[31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. 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, 21–37.

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[33] Fuya Luo, Yijun Cao, and Yongjie Li. 2021. Nighttime thermal infrared image colorization with dynamic label mining. In *Image and Graphics: 11th International Conference, ICIG 2021, Haikou, China, August 6–8, 2021, Proceedings, Part III*. Springer, 388–399.

[34] Fuya Luo, Yunhan Li, Guang Zeng, Peng Peng, Gang Wang, and Yongjie Li. 2022. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. *IEEE Transactions on Intelligent Transportation Systems* 23, 9 (2022), 15808–15823.

[35] Fu-Ya Luo, Yi-Jun Cao, Kai-Fu Yang, and Yong-Jie Li. 2022. Memory-Guided Collaborative Attention for Nighttime Thermal Infrared Image Colorization. *arXiv preprint arXiv:2208.02960* (2022).

[36] Yi Luo, Dechang Pi, Yue Pan, Lingqiang Xie, Wen Yu, and Yufei Liu. 2022. ClawGAN: Claw connection-based generative adversarial networks for facial image translation in thermal to RGB visible light. *Expert Systems with Applications* 191 (2022), 116269.

[37] I Scott MacKenzie. 2012. Human-computer interaction: An empirical research perspective. (2012).

[38] Kangfu Mei, Yiqun Mei, and Vishal M Patel. 2022. Thermal to visible image synthesis under atmospheric turbulence. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2051–2055.

[39] Nithin Gopalakrishnan Nair and Vishal M Patel. 2023. T2V-DDPM: Thermal to Visible Face Translation using Denoising Diffusion Probabilistic Models. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–7.

[40] Jason Nataprawira, Yanlei Gu, Igor Goncharenko, and Shunsuke Kamijo. 2021. Pedestrian detection using multispectral images and a deep neural network. *Sensors* 21, 7 (2021), 2536.

[41] Adam Nyberg, Abdelrahman Eldesokey, David Bergstrom, and David Gustafsson. 2018. Unpaired thermal to visible spectrum transfer using adversarial training. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[42] Bowen Pan and Shangfei Wang. 2018. Facial expression recognition enhanced by thermal images through adversarial learning. In *Proceedings of the 26th ACM international conference on Multimedia*. 1346–1353.

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

[44] Fang Qingyun and Wang Zhaokui. 2022. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition* 130 (2022), 108786.

[45] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[46] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.

[47] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

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

[50] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. 2017. Infrared image colorization based on a triplet dcgan architecture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 18–23.

[51] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. 2018. Learning to colorize infrared images. In *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection-15th International Conference, PAAMS 2017 15*. Springer, 164–172.

[52] Fan Sun, Wujie Zhou, Lv Ye, and Lu Yu. 2022. Hierarchical decoding network based on swin transformer for detecting salient objects in RGB-T images. *IEEE Signal Processing Letters* 29 (2022), 1714–1718.

[53] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. 35–43.

[54] Xiaoheng Tan, Yanxia Chai, Fenglei Chen, and Haijun Liu. 2022. A Fourier-Based Semantic Augmentation for Visible-Thermal Person Re-Identification. *IEEE Signal Processing Letters* 29 (2022), 1684–1688.

[55] Atousa Torabi, Guillaume Massé, and Guillaume-Alexandre Bilodeau. 2012. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding* 116, 2 (2012), 210–221.

[56] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[58] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696* (2022).

[59] Pingyu Wang, Fei Su, Zhicheng Zhao, Yanyun Zhao, Lei Yang, and Yang Li. 2020. Deep hard modality alignment for visible thermal person re-identification. *Pattern Recognition Letters* 133 (2020), 195–201.

[60] Zhongling Wang, Zhenzhong Chen, and Feng Wu. 2018. Thermal to visible facial image translation using generative adversarial networks. *IEEE Signal Processing Letters* 25, 8 (2018), 1161–1165.

[61] Zeyu Wang, Haibin Shen, Changyou Men, Quan Sun, and Kejie Huang. 2023. Thermal Infrared Image Inpainting Via Edge-Aware Guidance. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[62] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[63] Baotai Wu, Yujian Feng, Yunfei Sun, and Yimu Ji. 2023. Feature Aggregation via Attention Mechanism for Visible-Thermal Person Re-Identification. *IEEE Signal Processing Letters* (2023).

[64] Mang Ye, Xiangyuan Lan, and Qingming Leng. 2019. Modality-aware collaborative learning for visible thermal person re-identification. In *Proceedings of the 27th ACM International Conference on Multimedia*. 347–355.

[65] Jun-Seok Yun, Seon-Hoo Park, and Seok Bong Yoo. 2022. Infusion-Net: Inter- and Intra-Weighted Cross-Fusion Network for Multispectral Object Detection. *Mathematics* 10, 21 (2022), 3966.

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

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

[68] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 633–641.

[69] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).