mathematics

MDPI

Article

Infusion-Net: Inter- and Intra-Weighted Cross-Fusion Network for Multispectral Object Detection

Jun-Seok Yun, Seon-Hoo Park and Seok Bong Yoo *

Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, Korea
* Correspondence: sbyoo@jnu.ac.kr; Tel.: +82-(62)-5303437

Abstract: Object recognition is conducted using red, green, and blue (RGB) images in object recognition studies. However, RGB images in low-light environments or environments where other objects occlude the target objects cause poor object recognition performance. In contrast, infrared (IR) images provide acceptable object recognition performance in these environments because they detect IR waves rather than visible illumination. In this paper, we propose an inter- and intra-weighted cross-fusion network (Infusion-Net), which improves object recognition performance by combining the strengths of the RGB-IR image pairs. Infusion-Net connects dual object detection models using a high-frequency (HF) assistant (HFA) to combine the advantages of RGB-IR images. To extract HF components, the HFA transforms input images into a discrete cosine transform domain. The extracted HF components are weighted via pretrained inter- and intra-weights for feature-domain cross-fusion. The inter-weighted fused features are transmitted to each other's networks to complement the limitations of each modality. The intra-weighted features are also used to enhance any insufficient HF components of the target objects. Thus, the experimental results present the superiority of the proposed network and present improved performance of the multispectral object recognition task.

Keywords: multispectral object detection; inter- and intra-weighted fusion; high-frequency component; discrete cosine transform

MSC: 68T45

Citation: Yun, J.-S.; Park, S.-H.; Yoo, S.B. Infusion-Net: Inter- and Intra-Weighted Cross-Fusion Network for Multispectral Object Detection. *Mathematics* **2022**, *10*, 3966. https://doi.org/10.3390/math10213966

Academic Editors: Wen-Yu Chung and Sebastian Iwaszenko

Received: 5 October 2022
Accepted: 24 October 2022
Published: 25 October 2022

1. Introduction

Among various developments in computer vision, object detection is an essential funciton. Multimodality object detection can be employed in surveillance cameras, drones, autonomous driving, license plate recognition [1,2], crack detection [3], and other applications, as shown in Figure 1. The current object detection models display remarkable speed and accuracy. Among them, YOLOv7 [4] recently surpassed all known object detection models in standard datasets, such as the Microsoft Common Objects in Context dataset [5], which included areas such as proper illumination and clear boundaries between objects and the background.

However, in the real world, some object detection degradation components are unsuitable for object detection, such as rain, fog, shadow, low light, and low resolution. Hence, under these conditions, object detection models suffer severe performance degradation. Infrared (IR) images with clearer object edges than RGB images can be used as input images to prevent performance degradation. Nevertheless, object detection models still exhibit unacceptable performance due to the lack of information about the IR image (object color, texture, etc.).

In addition, object detection models cannot purify object detection-irrelevant features, such as dark illumination, occluded objects in RGB images, and IR image colors. As unpurified object detection-irrelevant features interrupt the model training stability, the accuracy of object detection models diminishes, as illustrated in Figure 2. For example, Figure 3

presents examples of the advantages of RGB and IR images under certain conditions. On a bright day, the RGB image in Figure 3a provides more information, including color, edges, and textures, than the IR image in Figure 3a. In contrast, at night, the IR image in Figure 3b displays a more precise outline of pedestrians under dark illumination than the RGB image in Figure 3b. In addition, even if a branch occludes the object, the IR image presents a clearer boundary of the item. Hence, an object detection model that can exploit the advantages of each image is necessary to achieve the best recognition performance in various environments.

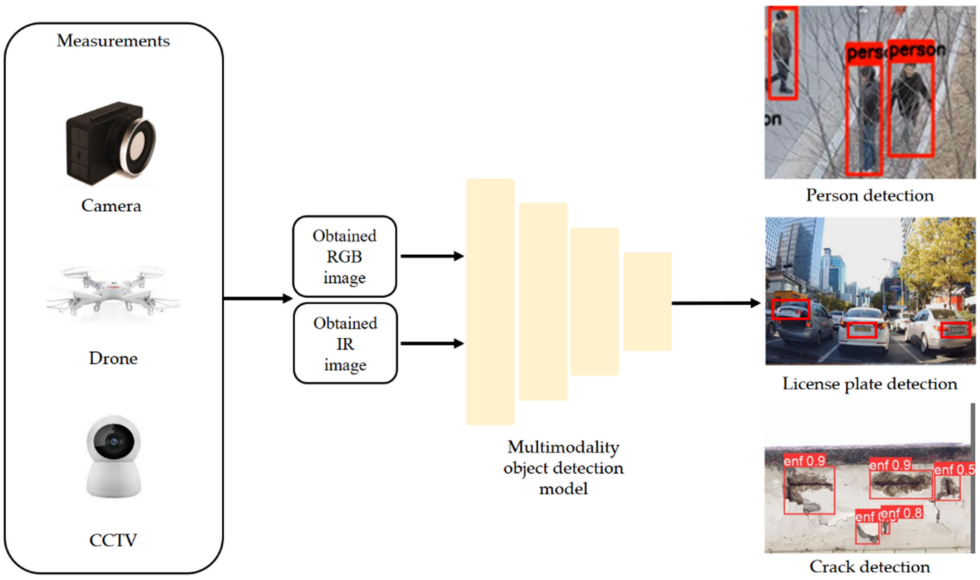


Figure 1. Example applications of the multimodality object detection model.

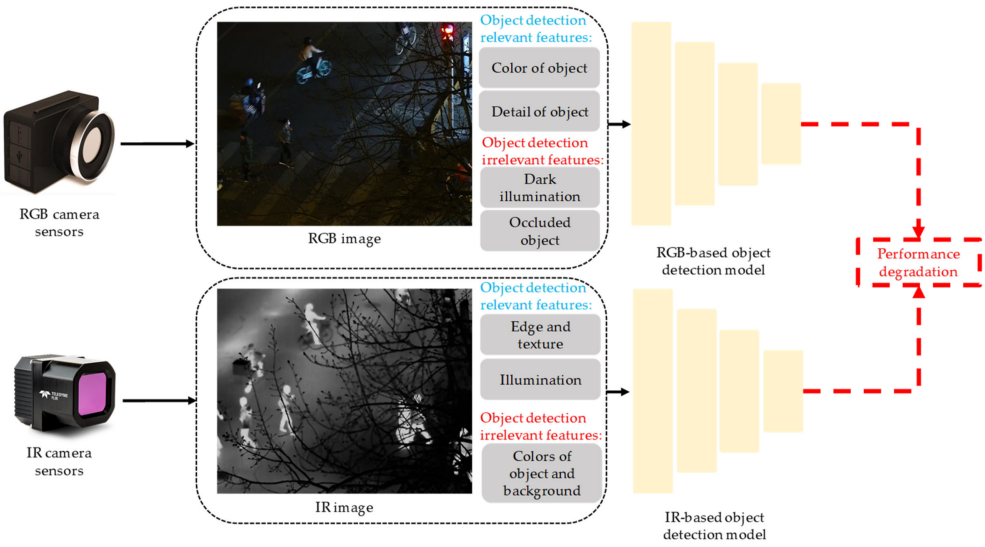
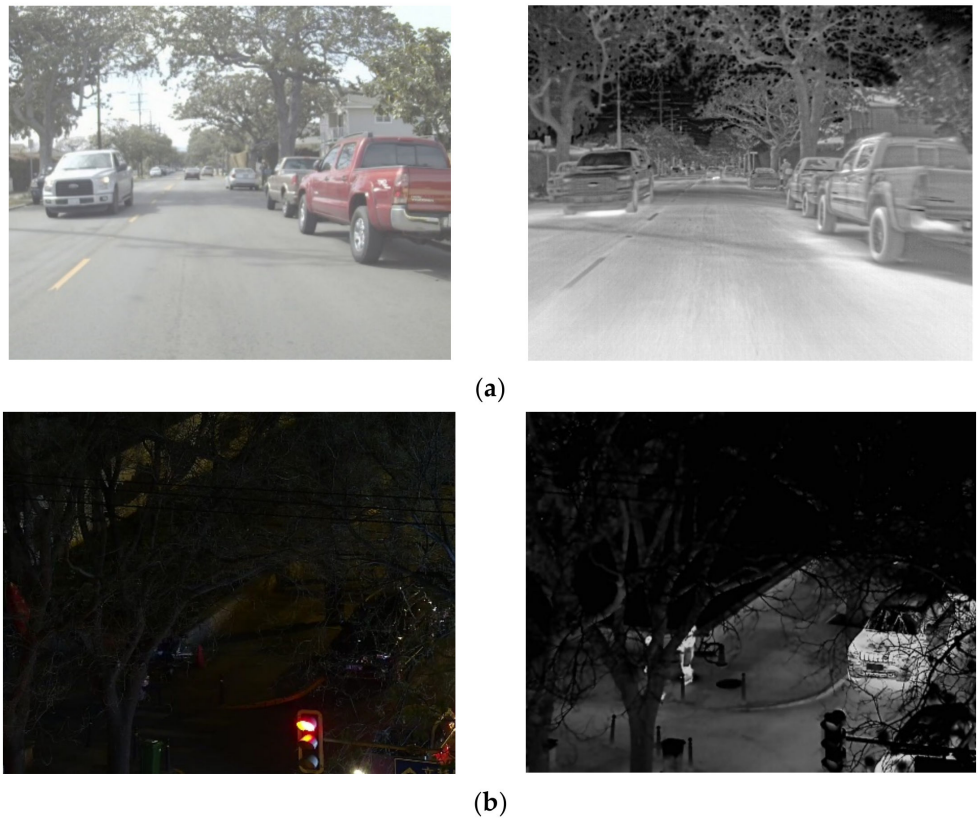


Figure 2. Problems with RGB- or IR-based object detection models.

Due to the development of convolution neural networks (CNNs), dual-stream CNN-based object detectors have been proposed. The dual-stream CNN-based object detectors consist of two streams: an RGB and IR stream. Using the two streams as features, these models improve the performance of object detection in a multispectral environment. In addition, multispectral datasets, such as FLIR [6] and LLVIP [7], that match the resolution of RGB images and IR images, have also led to continuous development in this field.

Recent dual-stream CNN-based object detectors fuse the features of RGB and IR images in the detection model. Various studies have been conducted regarding object recognition in a multispectral environment based on these fusion approaches. However, these approaches

cannot properly exploit the advantages of RGB and IR streams. As these models add or multiply the features of each stream, models cannot adaptively fuse information on features according to each feature level. In addition, object detection-irrelevant features (occluded objects in RGB images, the color of IR images, etc.) may be fused, causing instability in model training and performance degradation.



(**a**)

(**b**)

Figure 3. RGB-IR image pair on (**a**) a bright day and (**b**) at night.

To address this issue, in this paper, we propose an inter- and intra-weighted cross-fusion network (Infusion-Net) using a high-frequency (HF) assistant (HFA). The Infusion-Net exchanges the features of RGB and IR with four HFA blocks, depending on the feature level. In this process, the Infusion-Net adaptively exploits the object detection-relevant features of each stream. The HFA block consists of a HF extraction and an enhancing process based on the discrete cosine transform (DCT). The HF extraction in the HFA block purifies the object detection-irrelevant features. In enhancing the process in the HFA, a residual channel attention block (RCAB) [8] reinforces the purified features. The Infusion-Net adaptably adjusts information utilization and feature enhancement for each stream according to each fusion phase via learnable inter- and intra-weight parameters. Moreover, Infusion-Net surpasses other multistream approaches on multispectral datasets, such as FLIR and LLVIP. In addition, the extensive experiments present the effectiveness of these approaches.

The contributions of this paper are summarized as follows:

- We propose an Infusion-Net that gradually fuses the features of RGB and IR streams according to feature level to exploit the advantages of each stream.
- We propose an HFA block that interchanges, purifies, and reinforces the HF information based on DCT. In the HFA block, the HF information is extracted and reinforced by the proposed extraction method and RCAB.
- We propose the learnable inter- and intra-weight parameters for HF interchange and reinforcement according to the fusion phases. The features of each stream can be adaptably enhanced and fused in each phase using learnable parameters.

2. Related Work

2.1. Object Detection Models

Owing to the development of the CNN, object detectors demonstrate remarkable performance. Several outstanding works have been proposed for object detection, including CenterNet [9], Faster R-CNN [10], and the YOLO series. The YOLO series has shown remarkable accuracy and inference speed, advancing the one-stage object detection design.

In addition, YOLOv4 [11] employs cross stage partial (CSP) darknet, which matches almost all optimal architecture features obtained by the network architecture search technique as a backbone. The scaled YOLOv4 was proposed based on the CSP approach by scaling the features, and it applies to both small and large networks. Moreover, YOLOv5 [12] has four models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Generally, YOLOv5 uses the architecture of CSPDarknet with a spatial pyramid pooling layer as a backbone. Further, YOLOv6 [13] was proposed using a self-distillation strategy performed on the classification and regression tasks. Furthermore, YOLOv6 dynamically adjusts the information from the teacher and labels it to help the student model learn knowledge more efficiently during all training phases.

Next, YOLOv7 [4] is the latest work in the YOLO series. This network further improves the detection speed and accuracy based on the previous work. Specifically, in terms of the overall architecture, this paper proposes extended-ELAN (E-ELAN). Specifically, the ELAN uses expand, shuffle, and merge cardinality to continuously enhance the network learning ability without destroying the original gradient path by considering the following design strategy [14]. In addition, E-ELAN can guide various groups of computational blocks to learn diverse features. Further, YOLOv7 also proposes a compound model scaling method to maintain the model properties from the initial design and the optimal structure. Regarding network optimization strategy, YOLOv7 introduces model reparameterization and dynamic label assignment, analyzes the existing problems, and reduces them.

Specifically, YOLOv7 was proposed with direct access to the cascade of ResNet [15] or DenseNet [16], providing more gradient diversity for various characteristic graphs. However, these structures destroy the network structure because RepConv [17] has an identity connection. Therefore, the YOLOv7 was proposed by removing the identity connection in RepConv and designing the planned reparameterized convolution, realizing the efficient combination of the reparameterized convolution and various networks. In addition, YOLOv7 uses the idea of deep supervision and adds an additional auxiliary head structure in the middle network layer as an auxiliary loss to guide the weight of the shallow network. These mono-modality object detection methods provide real-time inference and achieve object detection performance.

However, these object detection models use only one stream. Hence, these models cannot exploit the advantages of each stream, such as the object color and detail information in RGB images and the precise edges and proper illumination in IR images. The proper feature utilization is required across each stream to acquire better object detection performance.

2.2. Fusion Mechanism of the Multispectral Object Detection

Despite numerous attempts at object detection, problems with improving performance by fusing features from other input modalities still exist. The conventional fusion mechanism-based approaches [18–20] focus on preprocessing the input image to obtain a fusion of RGB and IR information. Moreover, these fusion mechanism-based models can be categorized by the position of the feature fusion, such as early, late, and halfway fusion. Moreover, these approaches focus on fusing the features, such as through element-wise adding, multiplication, max-pooling, average-pooling, element-wise product, and other methods. Among the previous studies, Wagner et al. [21] analyzed two fusion methods, early and late fusion, and their performance on multispectral datasets. Based on the analysis, Wagner et al. proposed the cyclic fuse-and-refine (CFR) method to improve performance by using cyclical fuses and refines for each spectral feature. Liu et al. [22] designed two other fusion methods, halfway and score fusion, using two convolutional

networks. The guided attentive feature fusion (GAFF) [23] was proposed to guide efficient and effective multispectral feature fusion by using attention modules. A gated fuse unit [24] was proposed to learn the combination of feature maps generated between RGB and IR streams to determine an optimal fusion mechanism. Recently, a cross-modality fusion transformer (CFT) [25] was proposed to combine the features of RGB and IR streams with state-of-the-art (SOTA) performance in a multispectral object detection task by using a transformer [26]. In particular, the CFT learns long-range dependencies and integrates global contextual information at the feature level. The network can robustly capture the latent interactions between RGB and IR images. These multimodality object detection methods present higher accuracy than mono-modality methods by exploiting the features of each stream. Moreover, these methods efficiently fuse the features of each stream in the feature domain. Hence, these models contain more powerful feature information than other approaches.

However, these fusion-based models [21–25] utilized the unpurified object detection-irrelevant features (occluded objects in RGB images, the color of IR images, etc.). As result, these object detection-irrelevant features may be fused, leading to instability in model training and performance degradation. Moreover, even though these fusion-based models fuse the features of each stream, they do not fully adaptably fuse the features of each stream according to the feature level. Therefore, the fusion information must be adaptably fused according to the feature level and fusion phase. Thus, we propose the Infusion-Net, which gradually fuses the features of each stream according to the feature level using learnable weights. It provides optimal feature utilization and selectively enhances the features according to the fusion phase. When fusing the features in the Infusion-Net, the object detection-irrelevant features are eliminated through HF extraction. Only object detection-relevant features are interchanged and reinforced by HFA.

3. Proposed Method

This section describes the Infusion-Net for multispectral object detection. Section 3.1 provides an overview of DCT. Section 3.2 describes the HFA based on the DCT. Finally, Section 3.3 presents the overall architecture of Infusion-Net.

3.1. Discrete Cosine Transform

This section describes the HFA block, which interchanges, purifies, and reinforces the HF information based on DCT. Before describing the HFA, we explain why we use the DCT in the HFA blocks and the principle behind the DCT [27].

The image of the spatial domain can be transformed into the frequency of the spectral domain. Inversely, the frequency can be transformed into an image without loss of image quality. The discrete Fourier transform is commonly employed to transform the image or frequency. However, even if the input value consists of integer values, the transformed output includes complex numbers. Calculating complex numbers causes memory overhead and high computational costs in CNN-based models. To address this issue, we apply the DCT, transforming the input values into integer values using the cosine function. We convert the two-dimensional (2D) images to a frequency; thus, we employ the 2D DCT and 2D inverse DCT. The equations for the 2D DCT and 2D inverse DCT (IDCT) are represented as follows:

$$2D\_DCT(u,v) = \alpha(u)\beta(v)\sum_{\mathrm{i}=0}^{N}\sum_{j=0}^{M}\mathbf{f}(i,j)\cos(\frac{\pi(2i+1)u}{2N})\cos\left(\frac{\pi(2j+1)v}{2M}\right), \quad (1)$$

$$2D\_IDCT(i,j) = \sum_{\mathrm{u}=0}^{N}\sum_{j=0}^{M}\alpha(u)\beta(\mathrm{v})\mathbf{F}(u,v)\cos(\frac{\pi(2u+1)i}{2N})\cos\left(\frac{\pi(2v+1)j}{2M}\right), \quad (2)$$

$$\alpha(u) = \begin{cases} \sqrt{\frac{1}{N}}, & u = 0 \\ \sqrt{\frac{2}{N}}, & u \neq 0 \end{cases}, \tag{3}$$

$$\beta(v) = \begin{cases} \sqrt{\frac{1}{M}}, & v = 0 \\ \sqrt{\frac{2}{M}}, & v \neq 0 \end{cases}, \tag{4}$$

where $\mathbf{F}(u, v)$ denotes the value of the transformed frequency, where the pixel value $\mathbf{f}(i, j)$ of the (i, j) position of the image is transformed, and $\mathbf{f}(i, j)$ denotes the value of the transformed image. Equations (1) and (2) represent the 2D DCT and 2D IDCT, respectively. Equations (3) and (4) display the cosine basis function and regularization constant.

In the DCT, the high frequency is concentrated on the bottom-right side of the frequency. In contrast, the low frequency is concentrated on the top-left side. With this principle, we can adaptably extract the desired HF information via the binary mask $\boldsymbol{\mathcal{M}}$. The mask has zero values at the top of the diagonal and a value of one at the bottom of the diagonal. Hence, we can extract the desired HF information through hyperparameters τ that adjust diagonal positions, as presented in Figure 4a. Black pixels indicate a value of zero, and white pixels indicate a value of one. The binary mask function can be formulated as follows:

$$\boldsymbol{\mathcal{M}}(i, j) = \begin{cases} 0, & y < -i + 2\tau w \\ 1, & \text{otherwise} \end{cases}, \tag{5}$$

where w denotes the image width, and i and j denote the horizontal and vertical coordinates of $\boldsymbol{\mathcal{M}}$, respectively. The hyperparameter τ ranges from 0 to 0.5. The generated $\boldsymbol{\mathcal{M}}$ is used to extract HF components $\boldsymbol{Hf}$ by multiplying the binary mask $\boldsymbol{\mathcal{M}}$ and the input, transformed by the 2D DCT. The equation for the HF extraction HE is formulated as follows:

$$\boldsymbol{Hf}_{DCT} = 2D\_DCT(\boldsymbol{I}) \bigotimes \boldsymbol{\mathcal{M}}, \tag{6}$$

$$\boldsymbol{Hf} = 2D\_IDCT(\boldsymbol{Hf}_{DCT}), \tag{7}$$

where $\boldsymbol{I}$ denotes the input and $\bigotimes$ denotes the element-wise product. In addition, $2D\_DCT$ and $2D\_IDCT$ denote the 2D DCT and 2D IDCT functions, respectively. Figure 4b depicts the results of the HF extraction, according to hyperparameter τ. As illustrated in Figure 4b, when hyperparameter τ is too small, object detection-irrelevant information, such as an unwanted color, exists, and when the hyperparameter τ is too large, object detection-relevant information, such as an edge of an object, is removed. Hence, the proper hyperparameter τ is needed to improve object detection performance.

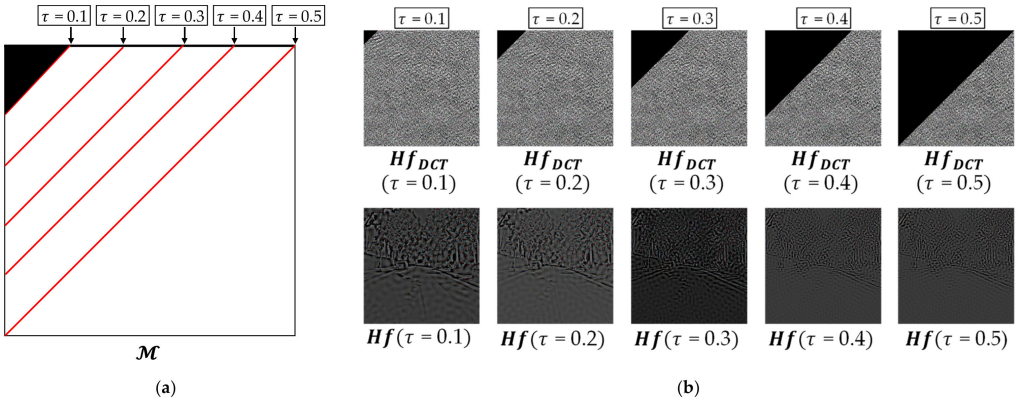


Figure 4. (**a**) Examples of binary mask $\boldsymbol{\mathcal{M}}$ according to hyperparameter τ; (**b**) results of high-frequency extraction according to hyperparameter τ.

In object detection tasks, the edges and texture information for objects are essential to improve object detection performance. This information helps detect the object position

and classify the object class. This paper demonstrates that edges and texture information are included in HF information. Figure 5 presents the DCT-based HF extraction results according to the HF extraction order. The RGB and IR images in Figure 5a,b are transformed into the DCT frequency domain. Afterward, HF information is extracted using a predefined binary mask. The extracted HF information is presented in Figure 5c,d. As illustrated in Figure 5e,f, the HF information contains edge and texture information that significantly affects the detection performance. Figure 5e displays more noticeable edges of objects than the original RGB image, making it difficult to recognize the object position with the naked eye. These results demonstrate that HF extraction can help improve object detection performance.

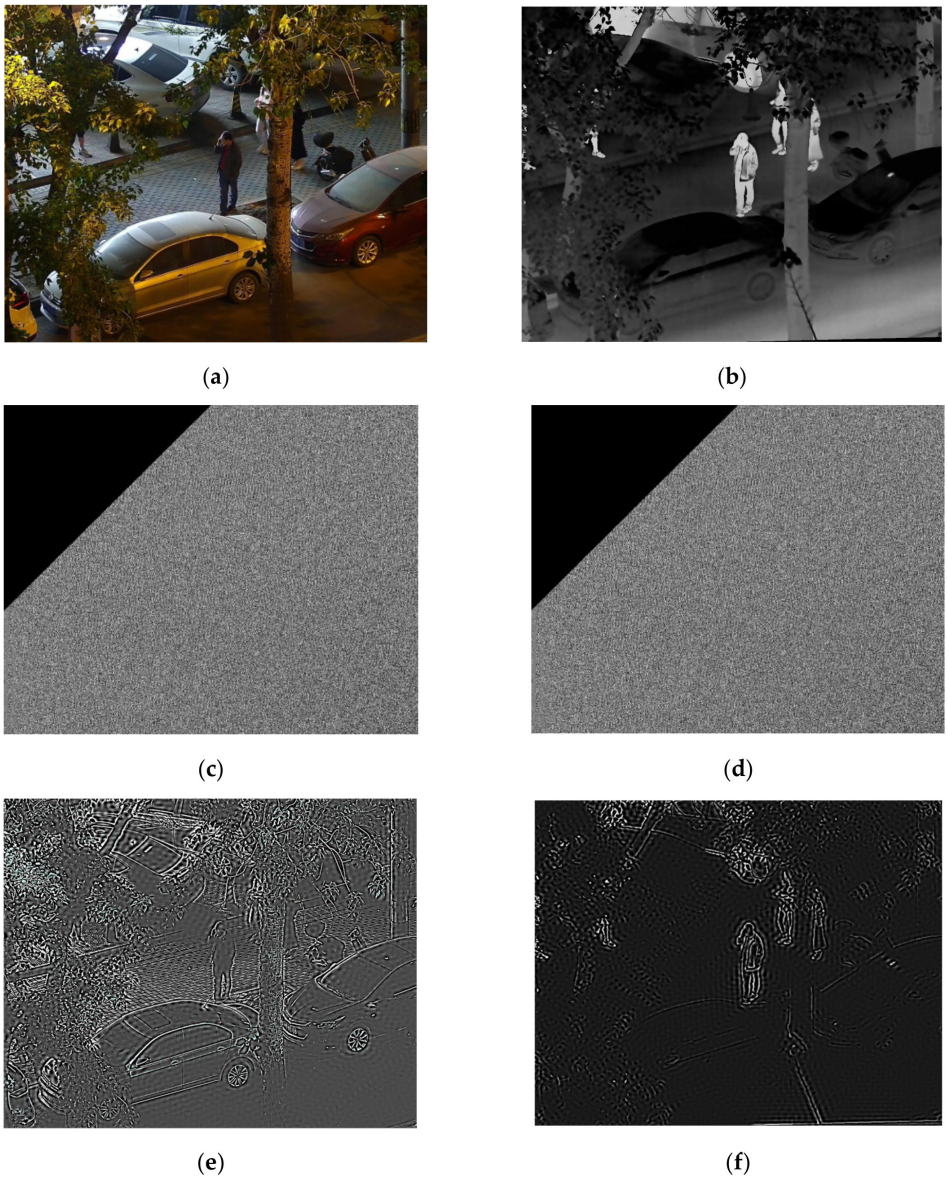


Figure 5. (**a**) RGB image; (**b**) IR image; (**c**) HF extraction in the DCT domain (RGB image); (**d**) HF extraction in the DCT domain (IR image); (**e**) extracted HF result (RGB image). (**f**) extracted HF result (IR image).

3.2. High-Frequency Assistant Based on the Discrete Cosine Transform

Based on the HF observation, we devised the HFA block, which interchanges, purifies, and reinforces HF features, as depicted in Figure 6. We applied the HF extraction to the feature level, not the image level. The HFA block transforms the features of the RGB and IR

streams to the DCT frequency. Then, only the necessary HF features are extracted using a predefined binary mask $\mathcal{M}$. The extracted HF features are transformed by the IDCT. With this process, the object detection-irrelevant features are purified. In addition, we reinforced the HF information using an RCAB [8]. As the RCAB focuses on the correlation around the features, the HF information is more precise and sharper. The reinforced HF information is added to the original feature. The enhanced features of the RGB and IR streams are interchanged with each stream. Hence, the proposed model has powerful representation compared with other approaches. The equation for the HFA is formulated as follows:

$$\left(f_{rgb} + \mathcal{F}(f_{ir}),\ f_{ir} + \mathcal{F}\left(f_{rgb}\right)\right) = HFA\left(f_{rgb}, f_{ir}\right), \tag{8}$$

$$\mathcal{F}(f) = f + \mathcal{R}(HE(f)), \tag{9}$$

where f_{rgb} and f_{ir} denote features of the RGB and IR streams, respectively, $\mathcal{R}$ represents the process of the RCAB, $\mathcal{F}$ denotes the function for reinforcing the HF information in the HFA, and HE is the process of HF extraction, as in Equations (6) and (7).

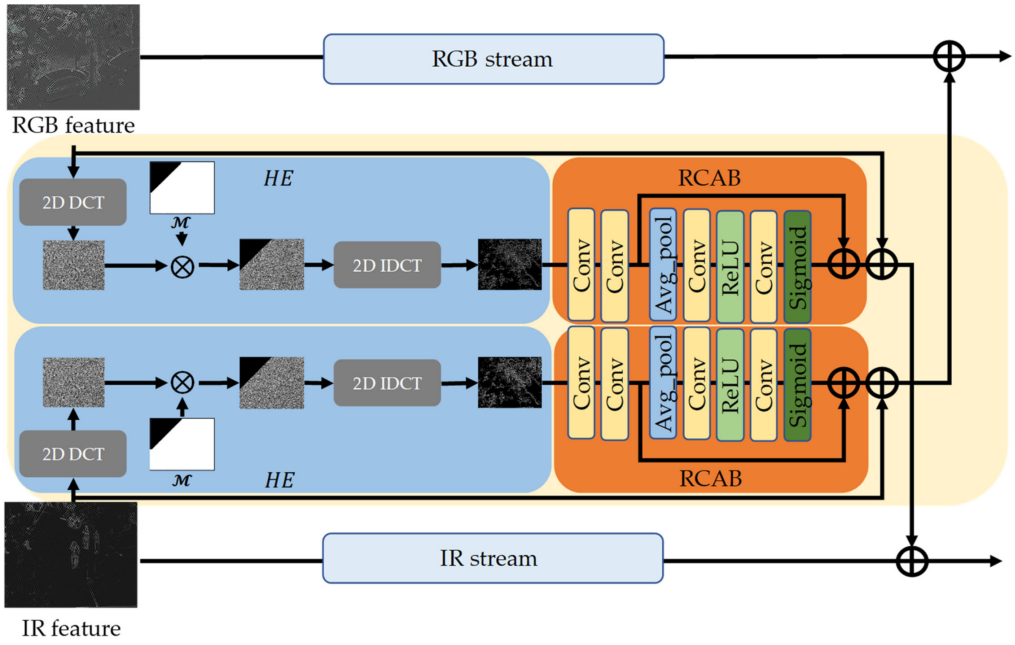


Figure 6. Architecture of the high-frequency assistant (HFA).

3.3. Overall Architecture of Inter- and Intra-Weighted Cross-Fusion Network (Infusion-Net)

This section describes the Infusion-Net for multispectral object detection. As presented in Figure 7, the Infusion-Net primarily consists of RGB and IR streams. Each stream receives an RGB image and an IR image as input. Then, four HFA blocks are employed to interchange the features of each stream, purifying only the necessary information. By dividing the interchange part according to the phase, we adaptably fuse the RGB and IR features by feature level. When fusing the features of each stream, the learnable intra- and inter-weight parameters are employed for the components of each stream. The intra-weight parameter α controls the degree of feature enhancement. As revealed in Figure 7, the extracted HF features of each stream are multiplied by each intra-weight parameter α according to each phase. The features are adaptably enhanced, according to each phase; thus, the Infusion-Net can exploit the optimal feature information depending on the fusion phase. The inter-weight parameter β controls the degree of utilization of the other stream feature, as depicted in Figure 7. Each stream adaptably receives features of another stream, according to feature level; thus, the Infusion-Net comprises the advantages of each stream.

The features of each stream according to the phase in the Infusion-Net can be formulated as follows:

$$f_{rgb}^{i+1} = f_{rgb}^{i} + \left(HE\left(f_{rgb}^{i}\right) \times \alpha_{rgb}^{pi}\right) + \left(\mathcal{F}\left(f_{ir}^{i} \times \beta_{ir}^{pi}\right)\right), \tag{10}$$

$$f_{ir}^{i+1} = f_{ir}^{i} + \left(HE\left(f_{ir}^{i}\right) \times \alpha_{ir}^{pi}\right) + \left(\mathcal{F}\left(f_{rgb}^{i} \times \beta_{rgb}^{pi}\right)\right), \tag{11}$$

where i denotes the phase of the Infusion-Net, f_{rgb}^{i} and f_{ir}^{i} represent the features of the RGB and IR streams when the phase is i, respectively, α_{ir}^{pi} and β_{ir}^{pi} are the intra- and inter-weight parameters of the IR stream, respectively, and α_{rgb}^{pi} and β_{rgb}^{pi} denote the intra- and inter-weight weight parameters of the RGB stream, respectively.

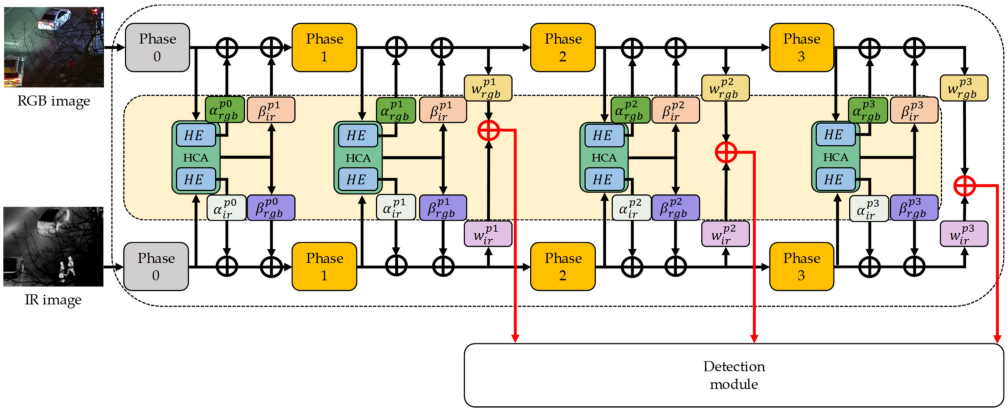


Figure 7. Overall architecture of the proposed Infusion-Net.

The fused features of each phase are concatenated with the detection module. When concatenating the features of each phase, the stream scaler weights are applied to emphasize the features of the stream according to phase selectively. Therefore, the Infusion-Net can filter the object-irrelevant features and emphasize more precise object-relevant features. A concatenated feature C is formulated as follows:

$$f_{m}^{i} = \left(f_{rgb}^{i} \times w_{rgb}^{pi}\right) + \left(f_{ir}^{i} \times w_{ir}^{pi}\right) \tag{12}$$

$$C = Concat\left(f_{m}^{1}, f_{m}^{2}, f_{m}^{3}\right), \tag{13}$$

where w_{rgb}^{pi} and w_{ir}^{pi} denote the stream scaler weights of each stream in phase i, f_{m}^{i} denotes fused features by using stream scaler, and *Concat* denotes the feature concatenation function.

As depicted in Figure 7, the concatenated features are used to predict the position and class of the object using the detection module. As in other approaches, we designed the detection module by applying YOLO v7 [4].

The structures of each phase are described in Figure 8. In Phase 0, the features are extracted by four convolutional layers from each image. The extracted features are extracted to another size of features to exploit the different scale features to improve the performance. In another phase, the max-pooling layer and more convolutional layers are employed to reduce the model parameters and inference time while maintaining the model representability. The extracted features are concatenated in the concatenation layer and aggregated by the convolutional layer.

The step-by-step block diagram of an application of Infusion-Net, as shown in Figure 9. The RGB and IR images are obtained by each measurement, such as from the RGB and IR camera. In general, the resolution of an RGB image is higher than that of an IR image. Hence, the images are aligned to match the position of the objects; then, aligned images are inputted into Infusion-Net. In Infusion-Net, the features of each stream are gradually fused with inter- and intra- weights and HFA. The features of each phase are utilized in the

detection module. The detection module predicts the class and position of objects in the input image.

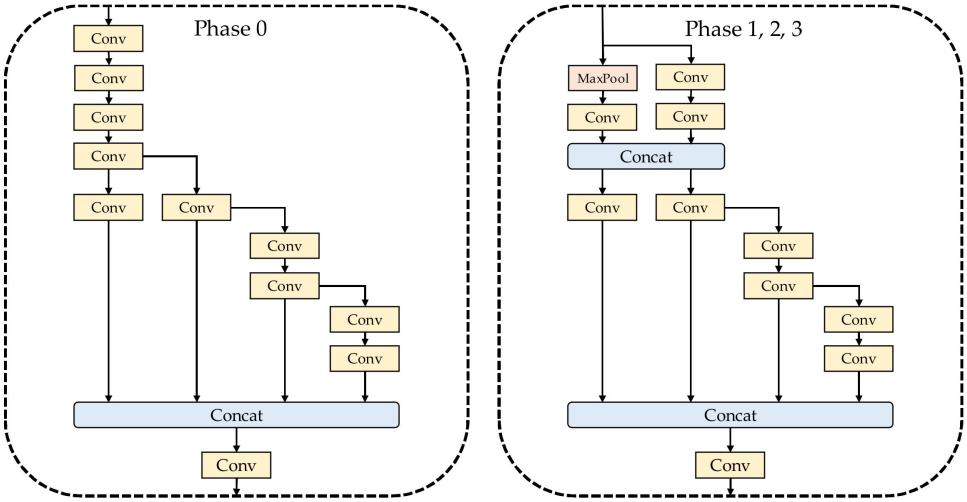


Figure 8. Structure of each phase in the Infusion-Net.

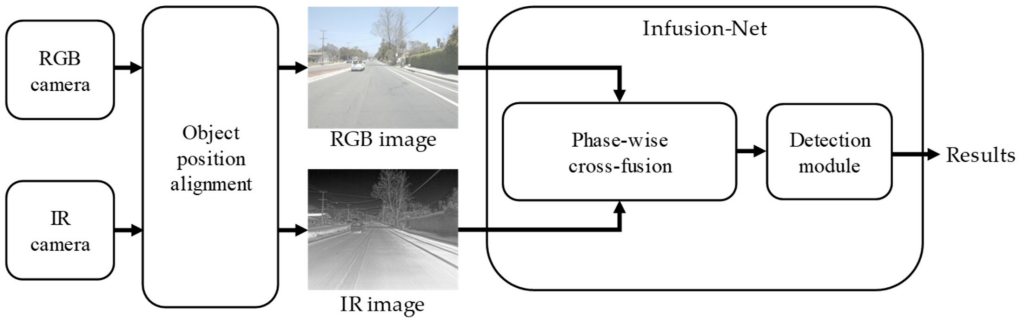


Figure 9. Step-by-step block diagram of an application of Infusion-Net.

4. Experiments and Analysis

4.1. Experimental Setup

4.1.1. Dataset

In the real world, the resolution of an RGB image is commonly higher than that of an IR image. As the positions of objects are mismatched, object detection models suffer performance degradation in multispectral object detection tasks. Hence, we used the FLIR [6] and LLVIP [7] datasets, which have RGB-IR pairs with the object position exactly matched. The FLIR dataset [6] is a multispectral object detection dataset that includes bright day and night scenes. The FLIR dataset consists of 5142 RGB-IR image pairs with 640 $\times$ 512 resolution. The classes of the FLIR dataset comprise people, cars, and motorcycles. In this experiment, 4129 RGB-IR image pairs were used for model training and 1013 for evaluation. The LLVIP [7] is a pedestrian detection dataset in a low-light night environment. The LLVIP dataset consists of 15,487 RGB-IR image pairs, and the object positions are accurately aligned. The resolution of the LLVIP dataset is 1280 $\times$ 1024. We split LLVIPR image pairs into 12,025 RGB-IR pairs for model training and 3,462 RGB-IR image pairs for evaluation.

4.1.2. Metric

We used the mean average precision (mAP) to analyze the detection results, an accuracy metric generally used to evaluate object detection models. We employed the mAP50, mAP75, and mAP50:95. The mAP50 indicates the mAP at the intersection over union (IoU) threshold of 0.5. The mAP75 is the mAP at the IoU threshold of 0.75, and mAP50:95 represents the average mAP at the IoU threshold of 0.5 to 0.95, with intervals of 0.05.

4.1.3. Environment

We implemented the proposed framework in PyTorch 1.8.0 and used Python 3.8.3, CUDA 11.2, and cuDNN 8.2.0. The experiment was performed using an AMD Ryzen 5 5600X 6-Core Processor CPU with 32 GB of memory and an NVIDIA RTX 3090 GPU. The hyperparameters of the framework were as follows: the initial learning rate was 10^{-2}, the momentum was 0.937, and the weight attenuation was 0.0005. The training batch size was set to 10, and the number of epochs was set to 100.

4.2. Comparison of Conventional and Proposed Multispectral Object Detection Methods

We set YOLOv5 [12] and YOLOv7 [4] as the baselines in this experiment. Specifically, YOLOv7 provides SOTA accuracy and speed in the object detection task. In addition, we conducted experiments on the mono-modality object detection models to demonstrate the advantage of multimodality. Furthermore, we set the YOLO series-based multimodality models as baselines; they consist of two back-bone structures, similar to the multimodality-based models. The multimodality-based models, such as CFR [21], GAFF [23], and CFT [25], are used to verify the Infusion-Net. Each object detection model is trained from scratch according to their losses and training methods on FLIR and LLVIP datasets.

Tables 1 and 2 compare the Infusion-Net and other models on the FLIR and LLVIP datasets. The Infusion-Net has 3.5% and 7.8% higher mAP50 values than the YOLOv7 mono-modality model on the FLIR and LLVIP datasets, respectively, demonstrating the superiority of the dual-stream approach. Among the multimodality models, the Infusion-Net performs best for mAP50, mAP75, and mAP50:95 on the FLIR and LLVIP datasets, as listed in Tables 1 and 2. In particular, the Infusion-Net scores 1.4% higher on the FLIR dataset and 1.1% higher on the LLVIP dataset than CFT, which has the second-best score. The performance of the Infusion-Net proves the superiority of this approach.

Table 1. Comparison of the Infusion-Net and other models on the FLIR dataset [6].

Model	Data	Backbone	mAP50 (%)	mAP75 (%)	mAP50:95 (%)
Mono-modality model					
YOLOv5 [12]	RGB	CSPDarknet	67.8	25.9	31.8
YOLOv5 [12]	IR	CSPDarknet	73.9	35.7	39.5
YOLOv7 [4]	RGB	E-ELAN	70.2	32.7	31.6
YOLOv7 [4]	IR	E-ELAN	75.6	32.2	38.2
Multimodality model					
halfway fusion [21]	RGB + IR	VGG16	71.2	-	-
CFR [21]	RGB + IR	VGG16	72.4	-	-
GAFF [23]	RGB + IR	ResNet18	72.9	32.9	37.5
GAFF [23]	RGB + IR	VGG16	72.7	30.9	37.3
YOLOv7 [4]	RGB + IR	E-ELAN	77.5	34.0	39.0
CFT [25]	RGB + IR	CFB	77.7	34.8	40.0
Infusion-Net (Ours)	RGB + IR	Infusion-Net	79.1	35.2	40.3

In this work, we propose an Infusion-Net that gradually fuses the features of RGB and IR streams, depending on each fusion phase, using four HFA blocks to exploit the object detection-relevant features of each stream. The HFA extracts HF information, which has clear boundaries and textures of the object, and the extracted HF information is strengthened by the RCAB. Enhanced features are exchanged between each stream with learnable inter- and intra-weight parameters and stream scaler weights. Using learnable weights, the Infusion-Net enables effective feature utilization and enhancement with the fusion phase, leading to improved object detection accuracy.

Table 2. Comparison of the Infusion-Net and other models on the LLVIP dataset [7].

Model	Data	Backbone	mAP50 (%)	mAP75 (%)	mAP50:95 (%)
Mono-modality model					
YOLOv5 [12]	RGB	CSPDarknet	90.8	51.9	50.0
YOLOv5 [12]	IR	CSPDarknet	94.6	72.2	61.9
YOLOv7 [4]	RGB	E-ELAN	91.9	52.9	51.2
YOLOv7 [4]	IR	E-ELAN	96.0	72.9	63.9
Multimodality model					
YOLOv5 [12]	RGB + IR	CSPDarknet	95.8	71.4	62.3
YOLOv7 [4]	RGB + IR	E-ELAN	96.9	73.1	64.4
CFT [25]	RGB + IR	CFB	97.5	72.9	63.6
Infusion-Net (Ours)	RGB + IR	Infusion-Net	98.6	73.3	64.6

To qualitatively evaluate the Infusion-Net, we compared it with CFT [25], which has the second-best mAP50 on the FLIR and LLVIP datasets. In Figure 10, the Infusion-Net detects a car, whereas CFT provides the missing object. As the Infusion-Net gradually fuses the features of RGB and IR streams according to the feature level, the Infusion-Net is better at detecting small objects than the CFT model, improving the object detection performance. As depicted in Figure 11, the Infusion-Net detects the person, whereas CFT provides the missing person. The visual outcomes present that Infusion-Net exploits HF information using HFA, which shows a precise edge of the object and texture, even if the input image exhibits dark illumination. Therefore, the Infusion-Net is robust in low-light environments.

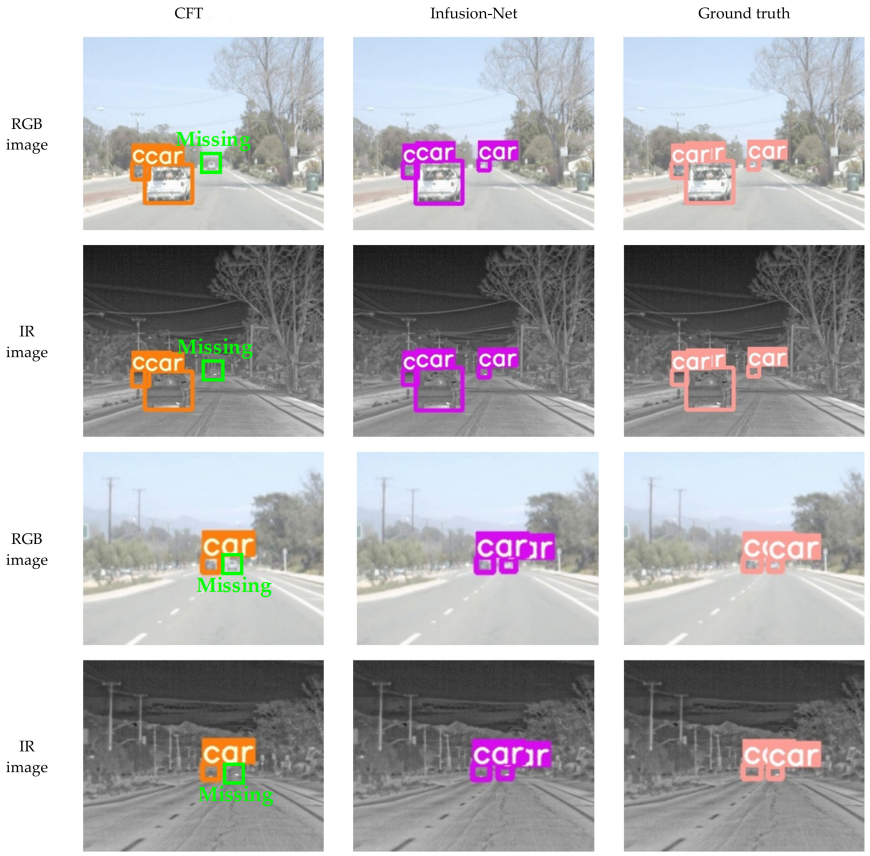


Figure 10. Qualitative comparison of the Infusion-Net results and CFT [25] results on the FLIR dataset. Green boxes denote missing objects.

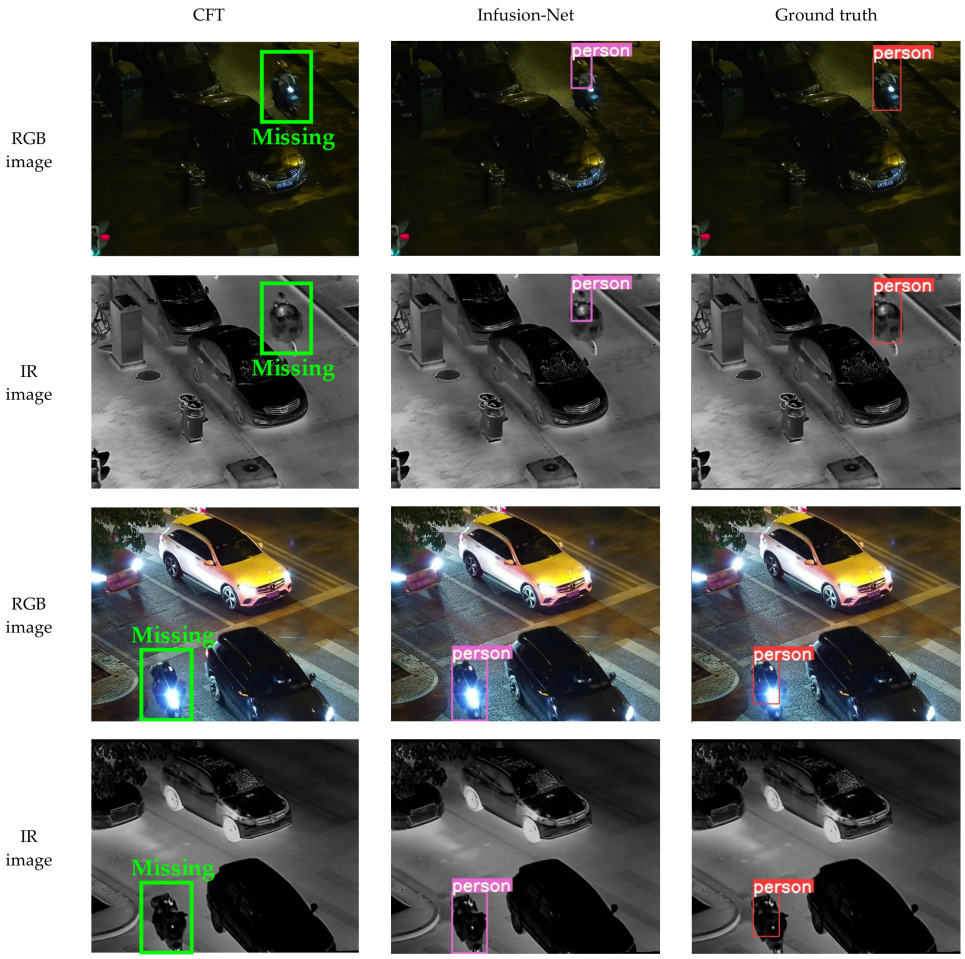


Figure 11. Qualitative comparison of the Infusion-Net results and CFT [25] results on the LLVIP dataset. Green boxes denote missing objects.

4.3. Ablation Study

4.3.1. Learnable Weights

This section investigates the effects of learnable weights, such as inter- and intra-weight parameters and stream scaler weights. Table 3 reveals that the learnable weights applied to the Infusion-Net improve object detection performance. In particular, the Infusion-Net exhibits the best mAP50 when all learnable weights are employed, indicating that the proposed learnable weights provide optimal feature utilization and selectively emphasize the features. Thus, the learnable weight approach is suitable for the multimodality model.

Table 3. Quantitative results for evaluating the effects of learnable weights on the FLIR dataset.

Intra-Weights α	Inter-Weights β	Scaler Weights w	mAP50 (%)
✗	✗	✗	77.9
✗	✗	O	78.6
✗	O	✗	78.5
O	✗	✗	78.0
O	O	O	79.1

4.3.2. Hyperparameters for High-Frequency Extraction

This section specifies how we determined the hyperparameters for HF extraction. As listed in Table 4, the Infusion-Net exhibits the best mAP50 when hyperparameter τ

is 0.2. In contrast, when the hyperparameter τ is 0, the Infusion-Net does not provide the best mAP50 score because the object detection-irrelevant features are not purified. In addition, when the hyperparameter is greater than 0.2, the object detection-relevant features are eliminated. Thus, the Infusion-Net cannot exhibit the optimal mAP50 score, demonstrating that the detection-relevant features, such as the edges and textures, are included in properly extracted HF information. Hence, the proper hyperparameter is needed to improve accuracy. In this paper, we empirically set the hyperparameter τ to 0.2.

Table 4. Quantitative results for evaluating the effects according to hyperparameter τ for the high-frequency extraction on the FLIR dataset.

Hyperparameter τ	**mAP50 (%)**
0	77.1
0.1	78.3
0.2	79.1
0.3	77.4
0.4	76.8
0.5	75.5

4.3.3. Computational Complexity

The proposed Infusion-Net consists of two backbone structures for multimodality; hence, the proposed model requires more inference time than the mono-modality models. Moreover, the resolution of an RGB image is higher than that of an IR image. Hence, Infusion-Net requires the image alignment process between an RGB and IR image, unlike other mono-modality models. Hence, Infusion-Net requires more computational complexity than mono-modality models. Nevertheless, the inference time for the Infusion-Net is 30 ms, whereas the inference time for the multimodality model CFT is 78 ms, when the resolution of the input images is 640×512. Even if the proposed model requires more inference time than the mono-modality models, the Infusion-Net performs an average of 5.6% higher on mAP50 than the mono-modality models on the FLIR and LLVIP datasets.

5. Conclusions

In this paper, to obtain the benefits of each stream, we suggest the Infusion-Net, which gradually fuses the features of RGB and IR streams, depending on the feature level. To this end, we devised an HFA that interchanges, purifies, and reinforces the object detection-relevant information based on the DCT and RCAB. In addition, the learnable inter- and intra-weight parameters and stream scaler weights provide optimal feature utilization across streams and selectively emphasize the features of the stream according to the fusion phase. The experimental results of the proposed model reveal the best performance in the mAP50, mAP70, and mAP50:95, with the fastest inference time, demonstrating the effectiveness of the Infusion-Net. The visual outcomes show more exact results than other SOTA models, demonstrating the superiority of Infusion-Net. In addition, the Infusion-Net can be considered for various object detection applications, such as fault diagnosis [28,29], autonomous driving [30], face attribute recognition [31,32], and smart parking in low-light environments. In particular, our Infusion-Net can be applied to prevent accidents such as car and pedestrian accidents at night because Infusion-Net presents a robust performance in low-light environments. In future work, the Infusion-Net will be extended to multi-input-based computer vision fields, such as RGB-LiDAR, RGB-Depth, and 3D images. In addition, the super-resolution methods can be employed to improve the object detection performance when input images are small.

Author Contributions: Conceptualization, S.B.Y.; methodology, J.-S.Y., S.-H.P. and S.B.Y.; software, J.-S.Y. and S.-H.P.; validation, J.-S.Y.; formal analysis, S.B.Y.; investigation, S.B.Y.; resources, J.-S.Y. and S.-H.P.; data curation, J.-S.Y. and S.-H.P.; writing—original draft preparation, J.-S.Y.; writing—review and editing, S.B.Y.; visualization, S.B.Y.; supervision, S.B.Y.; project administration, S.B.Y.; funding acquisition, S.B.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2020R1G1A1100798, NRF-2020R1A4A1019191) and the Industrial Fundamental Technology Development Program (No. 20018699) funded by the Ministry of Trade, Industry & Energy (MOTIE) of Korea.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lee, S.-J.; Yun, J.-S.; Lee, E.J.; Yoo, S.B. HIFA-LPR: High-Frequency Augmented License Plate Recognition in Low-Quality Legacy Conditions via Gradual End-to-End Learning. *Mathematics* **2022**, *10*, 1569. [CrossRef]
2. Lee, S.; Yun, J.S.; Yoo, S.B. Alternative Collaborative Learning for Character Recognition in Low-Resolution Images. *IEEE Access* **2022**, *10*, 22003–22017. [CrossRef]
3. Hong, Y.; Lee, S.; Yoo, S.B. AugMoCrack: Augmented Morphological Attention Network for weakly supervised crack detection. *Electron. Lett.* **2022**, *58*, 651–653. [CrossRef]
4. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. *arXiv* **2022**, arXiv:2207.02696. Available online: https://arxiv.org/abs/2207.02696 (accessed on 4 October 2022).
5. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Piotr, D.; Zitnick, C.L. Microsoft Coco: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
6. FREE FLIR Thermal Dataset for Algorithm Training. Available online: https://www.flir.in/oem/adas/adas-dataset-form (accessed on 4 October 2022).
7. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A Visible-Infrared Paired Dataset for Low-Light Vision. In Proceedings of the International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 3496–3504.
8. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 286–301.
9. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint Triplets for Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6569–6578.
10. Girshick, R. Fast R-Cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
11. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. Available online: https://arxiv.org/abs/2004.10934 (accessed on 4 October 2022).
12. YOLOv5. Available online: https://github.com/ultralytics/yolov5 (accessed on 2 October 2021).
13. Chuyi, L.; Lulu, L.; Hongliang, J.; Kaiheng, W.; Yifei, G.; Liang, L.; Zaidan, K.; Qingyuan, L.; Meng, C.; Weiqiang, N.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976. Available online: https://arxiv.org/abs/2209.02976 (accessed on 4 October 2022).
14. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.; Dollár, P. Designing network design spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–23 June 2020; pp. 10428–10436.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
16. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
17. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13733–13742.
18. Li, C.; Song, D.; Tong, R.; Tang, M. Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation. In Proceedings of the British Machine Vision Conference 2018, Newcastle, UK, 3–6 September 2018; p. 225.
19. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171. [CrossRef]
20. Zhang, L.; Zhu, X.; Chen, X.; Yang, X.; Lei, Z.; Liu, Z. Weakly Aligned Cross-Modal Learning for Multispectral Pedestrian Detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5126–5136.
21. Wagner, J.; Fischer, V.; Herman, M.; Behnke, S. Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks. In Proceedings of the 24th European Symposium on Artificial Neural Networks, Bruges, Belgium, 27–29 April 2016.

22. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D.N. Multispectral Deep Neural Networks for Pedestrian Detection. In Proceedings of the British Machine Vision Conference 2016, York, UK, 19–22 September 2016.
23. Zhang, H.; Fromont, E.; Lefèvre, S.; Avignon, B. Guided attentive feature fusion for multispectral pedestrian detection. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 72–80.
24. Zheng, Y.; Izzat, I.H.; Ziaee, S. GFD-SSD: Gated fusion double SSD for multispectral pedestrian detection. *arXiv* **2019**, arXiv:1903.06999. Available online: https://arxiv.org/abs/1903.06999 (accessed on 2 October 2021).
25. Qingyun, F.; Dapeng, H.; Zhaokui, W. Cross-modality fusion transformer for multispectral object detection. *arXiv* **2021**, arXiv:2111.00273. Available online: https://arxiv.org/abs/2111.00273 (accessed on 2 October 2021).
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
27. Yun, J.S.; Yoo, S.B. Single image super-resolution with arbitrary magnification based on high-frequency attention network. *Mathematics* **2022**, *10*, 275. [CrossRef]
28. Glowacz, A. Thermographic fault diagnosis of ventilation in BLDC motors. *Sensors* **2021**, *21*, 7245. [CrossRef] [PubMed]
29. Glowacz, A. Fault diagnosis of electric impact drills using thermal imaging. *Measurement* **2021**, *171*, 108815. [CrossRef]
30. Haris, M.; Glowacz, A. Navigating an Automated Driving Vehicle via the Early Fusion of Multi-Modality. *Sensors* **2022**, *22*, 1425. [CrossRef] [PubMed]
31. Lee, I.; Yun, J.S.; Kim, H.H.; Na, Y.; Yoo, S.B. LatentGaze: Cross-Domain Gaze Estimation through Gaze-Aware Analytic Latent Code Manipulation. *arXiv* **2022**, arXiv:2209.10171. Available online: https://arxiv.org/abs/2209.10171 (accessed on 18 October 2022).
32. Yun, J.S.; Na, Y.; Kim, H.H.; Kim, H.I.; Yoo, S.B. HAZE-Net: High-Frequency Attentive Super-Resolved Gaze Estimation in Low-Resolution Face Images. *arXiv* **2022**, arXiv:2209.10167. Available online: https://arxiv.org/abs/2209.10167 (accessed on 18 October 2022).