UniRGB-IR: A Unified Framework for Visible -Infrared Downstream Tasks via Adapter Tuning

Maoxun Yuan*, Bo Cui*, Tianyi Zhao and Xingxing Wei‡

***Abstract*—Semantic analysis on visible (RGB) and infrared (IR) images has gained attention for its ability to be more accurate and robust under low-illumination and complex weather conditions. Due to the lack of pre-trained foundation models on the large-scale infrared image datasets, existing methods prefer to design task-specific frameworks and directly fine-tune them with pre-trained foundation models on their RGB-IR semantic relevance datasets, which results in poor scalability and limited generalization. In this work, we propose a scalable and efficient framework called UniRGB-IR to unify RGB-IR downstream tasks, in which a novel adapter is developed to efficiently introduce richer RGB-IR features into the pre-trained RGB-based foundation model. Specifically, our framework consists of a vision transformer (ViT) foundation model, a Multi-modal Feature Pool (MFP) module and a Supplementary Feature Injector (SFI) module. The MFP and SFI modules cooperate with each other as an adpater to effectively complement the ViT features with the contextual multi-scale features. During training process, we freeze the entire foundation model to inherit prior knowledge and only optimize the MFP and SFI modules. Furthermore, to verify the effectiveness of our framework, we utilize the ViT-Base as the pre-trained foundation model to perform extensive experiments. Experimental results on various RGB-IR downstream tasks demonstrate that our method can achieve state-of-the-art performance. The source code and results are available at https://github.com/PoTsui99/UniRGB-IR.git**

***Index Terms*—RGB-IR Downstream Tasks, Multi-modal Fusion, Adapter, Transformer.**

I. Introduction

Single visible (RGB) image semantic analysis is a common practice in computer vision and has been widely used in a variety of vision tasks [1]–[3]. With the development of physical devices, more and more general-purpose foundation backbones [4], [5] pre-trained on a large-scale dataset (*e.g.*, ImageNet [6] and COCO [7]) have been designed for various downstream tasks. Since these pre-trained foundation models implicitly encode a large amount of prior knowledge, many researchers [8], [9] always fine-tune them on their semantic relevance datasets to significantly improve the performance on downstream tasks and speed up training convergence.

However, visible cameras have proved to be difficult in providing reliable imaging [11], [12], due to their limited bandwidth across the spectrum. Especially in low-illumination and complex weather conditions, the useful information of RGB images is relatively scarce. Therefore, infrared (IR) modality with stronger low-light adaptability are used as additional information to supplement RGB modality [11], [13]. Afterwards, the joint use of RGB and IR images has been applied in more and more semantic analysis tasks.

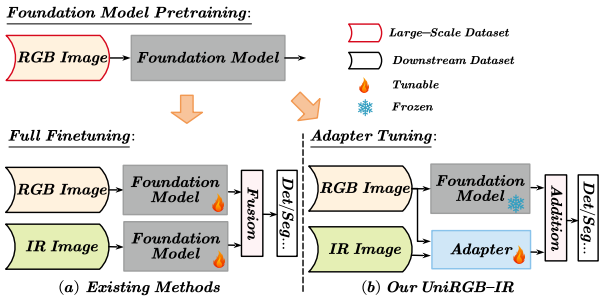

Fig. 1. Existing full fine-tuning methods *vs.* our UniRGB-IR framework. (a) Existing methods use pre-trained RGB-based foundation models and fully fine-tune them on their RGB-IR semantic relevance datasets. (b) We utilize the Adpater [10] to propose a new unified framework called UniRGB-IR, which can efficiently introduce richer RGB-IR features into the pre-trained foundation model for downstream tasks.

Due to the lack of pre-trained foundation models on the large-scale infrared image datasets, most existing methods [14], [15] prevailing trend to use pre-trained RGB baselines and fine-tune them on their RGB-IR semantic relevance datasets as shown in Figure 1 (a). For example, RGB-IR object detectors [14], [16] utilize RGB-based models as strong baselines and fine-tune them to extract RGB and IR features. SwinNet [17] is explored based on the RGB-based transformer to extract hierarchical features of each modality for the detection of salient RGB-IR objects. Similarly, Ha *et al.* [18] propose a new baseline by aggregating infrared modality features into the RGB-based framework and retrain it in the RGB-IR semantic segmentation benchmark. Among the efforts to achieve competitive results, these approaches still suffer from two major issues. (1) **Different downstream tasks require customized model structures**, resulting in poor model scalability and excessive parameter storage. (2) **Full fine-tuning strategy impairs the prior knowledge encoded in the foundation model**, which limits the generalization potential of the fine-tuned model. Based on these issues, a question that naturally arises is whether we can adapt the RGB-based foundation model to efficiently construct a unified framework for RGB-IR downstream tasks.

To achieve this purpose, in this paper, we explore an

Maoxun Yuan and Bo Cui are with the Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (email: yuanmaoxun@buaa.edu.cn).

Tianyi Zhao is with the Institute of Artificial Intelligence, Beihang University, No.37, Xueyuan Road, Haidian District, Beijing, 100191, China.

Xingxing Wei is with Institute of Artificial Intelligence, Hangzhou Innovation Institute, Beihang University, Beijing, 100191, China (email: xxwei@buaa.edu.cn).

‡ indicates the corresponding author.

∗ represents the equal contribution to this work.

efficient and scalable framework for RGB-IR downstream tasks and, for the first time, propose a **Uni**fied model for **RGB-IR** downstream tasks called **UniRGB-IR** as shown in Figure 1 (b). Drawing inspiration from recent advances in adapters [10], [19], which is initially used to yield a extensible model to effectively exploit the potential representation of foundation models in natural language processing (NLP) field, we aim to develop an adapter to dynamically introduce richer RGB-IR features into the pre-trained foundation model through an additional CNN branch. Therefore, without modifying the original foundation model, the powerful pre-trained weights can be directly loaded and frozen to accelerate training convergence and achieve efficient fine-tuning on downstream tasks. Specifically, we design two core modules in our UniRGB-IR: a **M**ulti-modal **F**eature **P**ool module (MFP) and a **S**upplementary **F**eature **I**njector module (SFI). In the MFP module, the multi-receptive field convolution and the feature pyramid are deployed to capture contextual multi-scale features from RGB and IR images. These features are fused at different scales to complement foundation models for different RGB-IR downstream tasks. For the SFI module, a cross-attention mechanism is used to dynamically inject the required features into the pre-trained foundation model, which facilitates our UniRGB-IR framework with a robust feature representation ability. To inherit prior knowledge of foundation model pre-trained on large-scale datasets, we utilize the adapter tuning paradigm instead of the full fine-tuning manner. In specific, we freeze the entire pre-trained weights and only optimize the MFP module and the SFI module. Therefore, our framework can be utilized as a unified framwork to efficiently fine-tune on various RGB-IR downstream tasks.

Overall, our contributions are summarized as follows:

- We explore an efficient and scalable framework called UniRGB-IR to unify RGB-IR downstream tasks. To the best of our knowledge, this is the first attempt to construct a unified framework for RGB-IR downstream tasks.
- We design a Multi-modal Feature Pool module and a Supplementary Feature Injector module. The former extracts contextual multi-scale features from two modality images, and the latter dynamically injects the required features into the pre-trained model. These two modules can be efficiently fine-tuned with adapter tuning paradigm to complement the pre-trained foundation model with richer RGB-IR features for RGB-IR downstream tasks.
- We incorporate the vision transformer foundation model into the UniRGB-IR framework to evaluate the effectiveness of our method on RGB-IR downstream tasks, including RGB-IR object detection, RGB-IR semantic segmentation, and RGB-IR salient object detection. Extensive experimental results demonstrate that our methods can efficiently achieve superior performance on these RGB-IR downstream tasks.

II. Related Work

A. Vision Foundation Models

Recently, thanks to the powerful long-distance modeling capability, vision transformers (ViT) [4] have been widely used as the foundation model in many vision tasks to achieve competitive results. The original ViT is a plain, non-hierarchical structure for image classification. Based on it, ViTDet [20] also constructs a non-hierarchical model by incorporating the feature pyramid. However, the non-hierarchical structure lacks rich feature representation, resulting in unsatisfactory performance. Subsequently, various hierarchical transformers [5], [21]–[23] have been proposed for different downstream vision tasks. Swin Transformer [21] and Multiscale Vision Transformer [22] are designed based on the ViT model to explore multi-scale features to improve the performance of image classification and object detection tasks. Besides, PVT [24] performs global attention on the downsampled key and value maps for dense prediction. In this paper, we also introduce the ViT model as the pre-trained foundation model to construct a unified framework for different RGB-IR downstream tasks.

B. RGB-IR Downstream Tasks

Since infrared cameras have stronger low-light adaptability to capture object thermal radiation, infrared images have been widely used together with RGB images in various RGB-IR downstream tasks in recent years.

RGB-IR Object Detection. Zhang *et al.* [25] explore a two-stream SSD [26] structure to capture the contextual enhanced features for RGB-IR object detection. Besides, AR-CNN [27] is presented based on faster R-CNN [28] to align RGB and IR features. With the emergence of transformers, Yuan *et al.* [29] propose a complementary fusion transformer (CFT) module to achieve advanced detection results. Furthermore, C^2Former [12] is a novel transformer block and can be incorporated into exited pre-trained models to increase intra- and inter-modality feature representations.

RGB-IR Semantic Segmentation. MFNet [18] is proposed to incorporate infrared features into the RGB-based framework to perform RGB-IR semantic segmentation. Based on the transformer structure, Wu *et al.* [30] propose a novel CCFFNet to excavate discriminative and complementary modality features for RGB-IR semantic segmentation. Moreover, CMX [31] is designed as a universal cross-modal fusion framework for RGB-IR semantic segmentation in an interactive fusion manner.

RGB-IR Salient Object Detection. SwinNet [17] is designed based on the Swin Transformer to extract hierarchical information of each modality, which achieves impressive results. CAVER [32] introduces the transformer to rethink the bi-modal salient object detection from a sequence-to-sequence perspective, which increases the model interpretability. Recently, Zhou *et al.* [33] transfer a large amount of knowledge learned in the transformer-based network to lightweight WaveNet through the distillation method.

The above methods attempt to design task-oriented structures to improve performance on corresponding downstream tasks. They either train the designed model from scratch or adopt a full fine-tuning strategy on a pre-trained model. Unlike these methods, we take a further step and deploy an adapter based on the pre-trained foundation model to unify the RGB-IR downstream tasks.

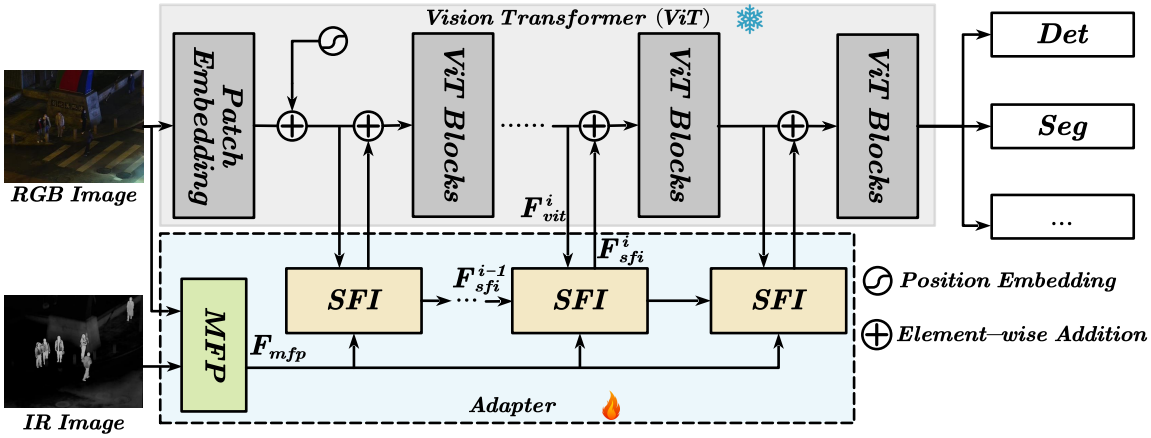

Fig. 2. The overall architecture of our UniRGB-IR. In our framework, a ViT model with different numbers of ViT blocks is deployed as a foundation model, which is divided into N (usually $N = 4$) stages for feature interaction. Besides, a Multi-modal Feature Pool (MFP) is designed to extract contextual multi-scale RGB-IR features and a Supplementary Feature Injector (SFI) is explored to introduce these features into the ViT foundation model. During training, we freeze the entire ViT model weights and only optimize the MFP and SFI modules.

C. Adapter

In the NLP field, Adapter [10] first fixes the original foundation backbone and introduces new modules into the transformer for task-specific fine-tuning, thereby effectively adapting the pre-trained backbone to downstream NLP tasks. Afterwards, Adapter has been widely studied in computer vision. ViT-Adapter [34] and Low-Rank Adapter [35] are designed to introduce a modest number of trainable parameters into the ViT and fine-tuning efficiently for dense prediction tasks. In addition, PC-Adapter [36] explores an attention-based adapter to preserve global shape knowledge for domain adaptation on point cloud data. Recently, Adapter is also used as a parameter-efficient training technology for vision-and-language tasks [37], [38].

In this paper, we aim to explore an adapter capable of converting the IR features into the RGB features to fit the pre-trained foundation model, which remains challenging to design this unified framework. Our UniRGB-IR is the first to propose utilizing adapter to unify RGB-IR downstream tasks.

III. Method

A. Overall Architecture

The overall framework of UniRGB-IR is illustrated in Figure 2, which consists of three parts: vision transformer model, Multi-modal Feature Pool (MFP) module, and Supplementary Feature Injector (SFI) module. In our framework, the ViT model is utilized as the pre-trained foundation model and frozen during the training process. Specifically, for the ViT model, the RGB image is directly fed into the patch embedding process to obtain the D-dimensional feature tokens, which are usually 1/16 of the original image resolution. To complement the richer features required for various RGB-IR downstream tasks, we feed the RGB and IR images into the MFP module to extract contextual multi-scale features from two modalities (*eg.* 1/8, 1/16 and 1/32 of the original image resolution). Afterwards, these richer features are dynamically injected into the features of ViT model through the SFI module, which can adaptively introduce the required RGB-IR features into the ViT model. To fully integrate the extracted features into the ViT model, we add the SFI module at the beginning of each stage. Thus, after N stages of feature injection, the final features from ViT model can be leveraged for various RGB-IR downstream tasks.

B. Multi-modal Feature Pool

To complement the rich feature representations for different RGB-IR downstream tasks, we introduce a multi-modal feature pool (MFP) module, including multiple perception and feature pyramid. The former can extract the contextual features with different convolution kernels, which achieve the long-distance modeling capability of CNNs. Different from existing works [39], [40] that increase the width or depth of the model, we efficiently achieve multi-receptive field perception in the channel dimension. As for feature pyramid, it can obtain multi-scale features to enhance the small object features. Thus, these two operations are connected in series to construct MFP module, as shown in Figure 3.

Specifically, for the input RGB ($H \times W \times 3$) and IR ($H \times W$) images, we first employ a stem block borrowed from ResNet [41] to extract two modality features F_1^{rgb} and $F_1^{ir} \in \mathbb{R}^{H/4 \times W/4 \times C}$. Then, these two features are split into four equal parts by utilizing the channel splitting technique [42]. To extract multi-receptive field perception, each part is subjected to convolution operations with different kernel sizes (3×3, 3×3, 5×5 and 7×7). Then, we fuse each processed feature from two modalities by using a SE attention module (shown in Figure 3). Therefore, we concatenate each fused part to obtain RGB-IR contextual features F_{fus}. The above process can be represented as:

$$F_{fus} = \Gamma_{k=1}^{4}(Fusion(W_k^{rgb} * f_k^{rgb}, W_k^{ir} * f_k^{ir})), \quad (1)$$

where $F_{fus} \in \mathbb{R}^{H/4 \times W/4 \times C}$, f_k^{rgb} and f_k^{ir} are the k-th part of F_1^{rgb} and F_1^{ir} features respectively, W_k is the convolution with k-th kernel size, Γ is the concatenation operation.

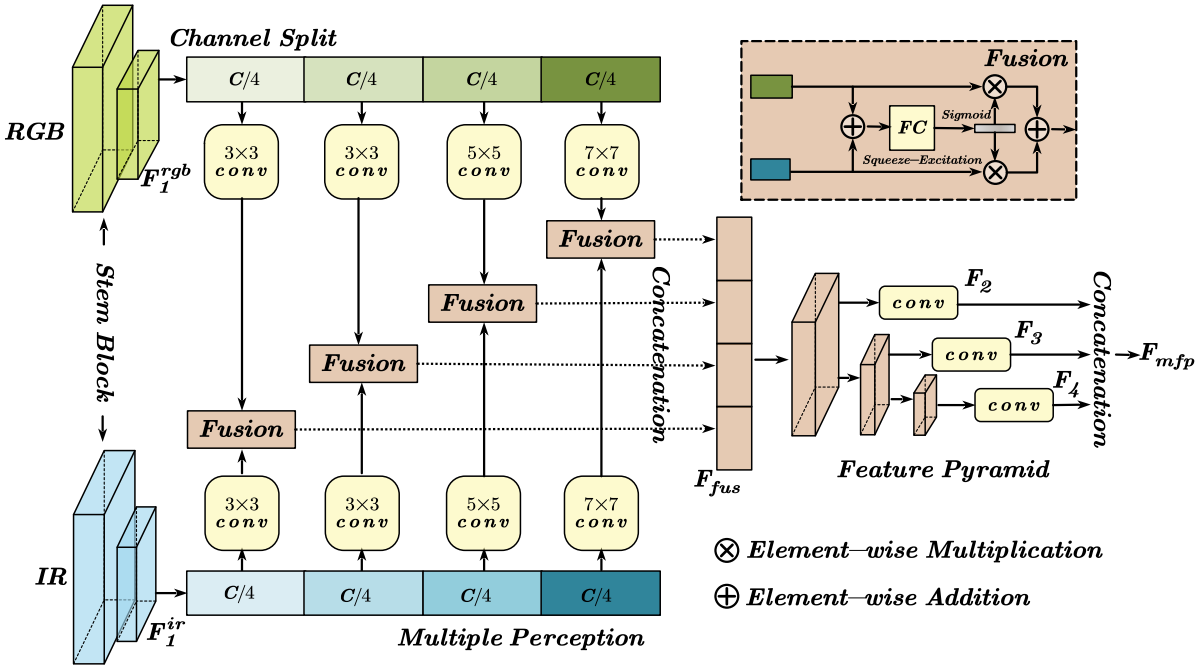

Fig. 3. Structure of the multi-modal feature pool (MFP) module. We explore the multiple perceptions to expand the receptive field of contextual feature extraction and utilize the feature pyramid to obtain the multi-scale features. In this module, Squeeze-Excitation(SE) attention is utilized to fuse two modality features. Thus, the contextual multi-scale features can be provided for the various RGB-IR downstream tasks.

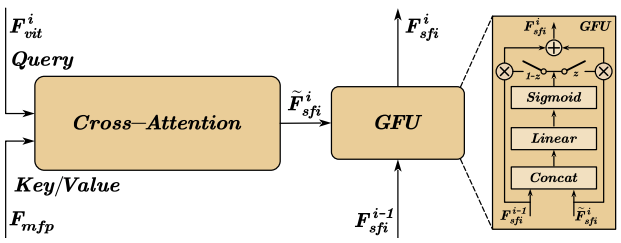

Fig. 4. Structure of supplementary feature injector (SFI) module. A gated fusion unit (GFU) is utilized to dynamically fuse the current and the last injected features.

For the feature pyramid, a stack of three 3×3 convolutions with $stride = 2$ is conducted to downsample the size of the feature maps. Then, features of each scale are fed into a 1×1 convolution to project the feature maps to D dimensions. Therefore, we can obtain a set of multi-scale features $\{F_2, F_3, F_4\}$ with 1/8, 1/16, and 1/32 of the original image resolution, respectively. Finally, we flatten and concatenate these features into feature tokens $F_{mfp} \in \mathbb{R}^{(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2})\times D}$, which will be used as supplementary features for the ViT model.

C. Supplementary Feature Injector

As shown in Figure 4, we propose a supplementary feature injector (SFI) module to adaptively introduce the contextual multi-scale features without altering the ViT structure. Since the sequence lengths of contextual multi-scale features F_{mfp} and the ViT features F^i_{vit} are different, to address this, we employ sparse attention (*eg.* [43] and [44]) to dynamically sample supplementary features from each scale. To be specific, we utilize the ViT features $F^i_{vit} \in \mathbb{R}^{\frac{HW}{16^2}\times D}$ as the query, and the contextual multi-scale features $F_{mfp} \in \mathbb{R}^{(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2})\times D}$ as the key and value:

$$\tilde{F}^i_{sfi} = Attention(LN(F^i_{vit}), LN(F_{mfp})), \tag{2}$$

where $Attention(\cdot)$ is the attention layer and $LN(\cdot)$ is LayerNorm [45] which is intended to reduce modality discrenpancy. Furthermore, we adopt progressive injection to introduce contextual multi-scale features, which can balance the foundation model features and the injected features F^i_{sfi}. Thus, a gated fusion unit is explored to predict the fusion weight z to gate F^{i-1}_{sfi} and $\tilde{F}^i_{sfi}$ for dynamic fusion. Specifically, we concatenate the two features F^{i-1}_{sfi} and $\tilde{F}^i_{sfi}$ and feed it into linear layer to predict the weight z. Then, z and $1-z$ are used to fuse F^{i-1}_{sfi} and $\tilde{F}^i_{sfi}$ features respectively. The final output features F^i_{sfi} of SFI module can be formulated as:

$$F^i_{sfi} = \begin{cases} \tilde{F}^i_{sfi}, & i = 1 \\ GFU(\tilde{F}^i_{sfi}, F^{i-1}_{sfi}), & i = 2 \dots N \end{cases} \tag{3}$$

where $GFU(\cdot,\cdot)$ is the gated fusion unit shown in Figure 4.

D. Adapter Tuning Paradigm

To fully inherit the prior knowledge of the ViT pretrained on large-scale datasets, we explore the adapter tuning paradigm instead of the full fine-tuning manner. For the dataset $D = \{(x_j, gt_j)\}_{j=1}^{M}$ of the downstream task, full fine-tuning process calculates the loss between the prediction and the ground truth, which can be formulated as:

$$\mathcal{L}(D, \theta) = \sum_{j=1}^{M} \text{loss}(F_\theta(x_j), gt_j), \tag{4}$$

TABLE I
COMPARISON RESULTS (MAP, IN%) ON THE FLIR AND LLVIP DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **RED** AND THE SECOND-BEST ARE HIGHLIGHTED IN **BLUE**. "–" INDICATES THAT THE AUTHOR DID NOT PROVIDE THE CORRESPONDING RESULTS.

Methods	Modality	FLIR			LLVIP		
		mAP	**mAP$_{50}$**	**mAP$_{75}$**	**mAP**	**mAP$_{50}$**	**mAP$_{75}$**
SSD [26]	IR	24.6	57.0	18.0	53.5	90.2	57.9
RetinaNet [46]	IR	31.5	66.1	25.3	55.1	94.8	57.6
Faster R-CNN [28]	IR	37.6	75.8	31.6	54.5	94.6	57.6
Cascade R-CNN (ViT-B) [47]	IR	41.9	78.6	37.3	60.4	94.9	67.5
DDQ-DETR [48]	IR	37.1	73.9	32.2	58.6	93.9	64.6
SSD [26]	RGB	18.8	46.3	13.1	39.8	82.6	31.8
RetinaNet [46]	RGB	21.9	51.2	15.2	42.8	88.0	34.4
Faster R-CNN [28]	RGB	27.7	62.2	21.2	45.1	87.0	41.2
Cascade R-CNN (ViT-B) [47]	RGB	33.3	69.3	26.2	51.7	90.3	54.7
DDQ-DETR [48]	RGB	30.9	64.9	24.5	46.7	86.1	45.8
GAFF [49]	RGB+IR	37.4	74.6	31.3	55.8	94.0	60.2
ProbEn [50]	RGB+IR	37.9	75.5	31.8	51.5	93.4	50.2
LGADet [51]	RGB+IR	-	74.5	-	-	-	-
CSAA [52]	RGB+IR	**41.3**	**79.2**	**37.4**	**59.2**	**94.3**	**66.6**
UniRGB-IR (Ours)	RGB+IR	**44.1**	**81.4**	**40.2**	**63.2**	**96.1**	**72.2**

TABLE II
COMPARISON RESULTS (MR^{-2}, IN%) UNDER 'ALL-DATASET' SETTINGS OF DIFFERENT PEDESTRIAN DISTANCES, OCCLUSION LEVELS, AND LIGHT CONDITIONS (DAY AND NIGHT) ON THE KAIST DATASET. THE BEST AND SECOND RESULTS ARE HIGHLIGHTED IN **RED** AND **BLUE** RESPECTIVELY.

Methods	Near	Medium	Far	None	Partial	Heavy	Day	Night	All
ACF [53]	28.74	53.67	88.20	62.94	81.40	88.08	64.31	75.06	67.74
Halfway Fusion [54]	8.13	30.34	75.70	43.13	65.21	74.36	47.58	52.35	49.18
FusionRPN+BF [55]	0.04	30.87	88.86	47.45	56.10	72.20	52.33	51.09	51.70
IAF R-CNN [56]	0.96	25.54	77.84	40.17	48.40	69.76	42.46	47.70	44.23
IATDNN+IASS [57]	0.04	28.55	83.42	45.43	46.25	64.57	49.02	49.37	48.96
CIAN [25]	3.71	19.04	55.82	30.31	41.57	62.48	36.02	32.38	35.53
MSDS-R-CNN [58]	1.29	16.19	63.73	29.86	38.71	63.37	32.06	38.83	34.15
AR-CNN [27]	0.00	16.08	69.00	31.40	38.63	**55.73**	34.36	36.12	34.95
MBNet [59]	0.00	16.07	55.99	27.74	35.43	59.14	32.37	30.95	31.87
TSFADet [16]	0.00	15.99	50.71	25.63	37.29	65.67	31.76	27.44	30.74
CMPD [60]	0.00	**12.99**	51.22	24.04	33.88	59.37	**28.30**	30.56	28.98
CAGTDet [29]	0.00	14.00	49.40	24.48	33.20	59.35	28.79	27.73	28.96
C^{2}Former [12]	0.00	13.71	**48.14**	**23.91**	**32.84**	57.81	28.48	**26.67**	**28.39**
UniRGB-IR (Ours)	**0.00**	**13.44**	**38.21**	**20.26**	**31.67**	**55.03**	**25.93**	**23.95**	**25.21**

where loss represents the loss function and F_θ denotes the entire network parameterized by θ. Afterwards, θ is optimized through the formula:

$$\theta \leftarrow \arg\min_{\theta} \mathcal{L}(D, \theta). \tag{5}$$

However, in our adapter tuning paradigm, the parameter θ consists of two parts, one part is the parameter in the original ViT model θ_V, and the other part is the parameter in our UniRGB-IR θ_A. During training, we freeze the parameter θ_V and only optimize the parameter θ_A. Thus, the loss function and optimization of our a can be represented as:

$$\mathcal{L}(D, \theta_V, \theta_A) = \sum_{j=1}^{M} \text{loss}(F_{\theta_V, \theta_A}(x_j), gt_j), \tag{6}$$

$$\theta_A \leftarrow \arg\min_{\theta_A} \mathcal{L}(D, \theta_V, \theta_A). \tag{7}$$

IV. EXPERIMENTS

To evaluate the effectiveness of our UniRGB-IR, we utilize the ViT-Base model (pre-trained on COCO [7] dataset) as the foundation model and utilize this framework to perform different RGB-IR downstream tasks. During training, we freeze the ViT-Base model and only optimize the MFP and SFI modules. We evaluate and compare our method with various competitive models, including CNN-based and Transformer-based models. Besides, our evaluation spans various tasks, including RGB-IR object detection on FLIR [70], LLVIP [71], and KAIST [53] datasets, RGB-IR semantic segmentation on MFNet [18] and PST9000 [63] datasets and RGB-IR salient object detection (*see supplementary materials*). Furthermore, ablation experiments on the designed modules and qualitative experiments are also conducted to verify that the UniRGB-IR framework can be leveraged as a unified framework to

TABLE III
RESULTS ON THE PST900 DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **RED** AND THE SECOND-BEST ARE HIGHLIGHTED IN **BLUE**. "–" INDICATES THAT THE AUTHOR DID NOT PROVIDE THE CORRESPONDING RESULTS.

Methods	Background		Fire-Extinguisher		Backpack		Hand-Drill		Survivor		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
MFNet [18]	**99.9**	98.7	71.8	67.4	52.8	52.4	46.7	39.3	18.8	18.9	58.0	55.3
RTFNet [61]	**99.9**	99.0	78.6	54.8	62.5	60.8	76.7	61.0	65.2	62.5	76.6	67.6
CCNet [62]	99.6	98.7	88.1	73.8	76.0	73.0	54.1	51.0	49.5	33.5	73.46	66.0
PSTNet [63]	-	98.9	-	70.1	-	69.2	-	53.6	-	50.0	-	68.4
ACNet [64]	**99.8**	99.3	84.9	60.0	85.6	83.2	53.6	51.5	69.1	65.2	78.7	71.8
FDCNet [65]	99.7	99.2	91.8	71.5	77.5	72.2	82.5	70.4	78.4	72.4	86.0	77.1
CGFNet [66]	99.7	99.3	**96.3**	71.7	87.6	82.0	70.5	59.7	**86.3**	**77.4**	88.1	78.0
EGFNet [67]	99.5	99.2	95.2	71.3	**94.2**	83.1	**98.0**	64.7	83.3	74.3	**94.0**	78.5
CCFFNet [30]	**99.9**	**99.4**	87.0	**79.3**	80.2	77.8	88.3	**77.4**	77.9	73.4	86.7	**81.5**
RSFNet [68]	-	99.4	-	75.4	-	84.9	-	72.9	-	70.1	-	80.5
SGFNet [69]	**99.8**	99.4	89.4	**75.6**	90.4	**85.4**	94.0	76.7	82.7	76.7	91.2	**82.8**
UniRGB-IR (Ours)	99.7	**99.5**	**97.0**	72.0	**93.3**	**87.7**	**95.5**	**78.0**	**86.1**	**77.8**	**94.3**	**82.8**

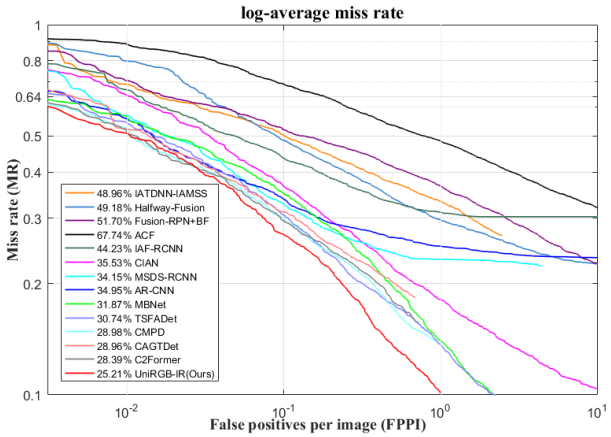

Fig. 5. The Miss rate curves for all detectors are plotted against False Positives Per Image (FPPI) under the IoU threshold of 0.5 on the KAIST dataset.

efficiently introduce IR features into the foundation model to achieve superior performance.

A. RGB-IR Object Detection

Datasets. Our object detection experiments are based on the three paired RGB and IR object detection datasets. FLIR [70] is a paired visible and infrared object detection dataset, including daytime and night scenes, which has 4,129 aligned RGB-IR image pairs for training and 1,013 for testing. For LLVIP [71] dataset, it contains 15,488 aligned RGB-IR image pairs, of which 12,025 images are used for training and 3,463 images for testing. As for KAIST [53] dataset, it is a aligned multispectral pedestrian deteciton dataset, in which 8,963 and 2,252 image pairs are utilized for training and testing.

Metrics. For FLIR and LLVIP datasets, we employ mean Average Precision (mAP) to evaluate the detection performance. As for KAIST dataset, we use log-average miss rate MR over the false positive per image (FPPI) with the range of $[10^{-2}, 10^{0}]$ to evaluate the pedestrian detection performance.

Settings. All the experiments are conducted with NVIDIA GeForce RTX 3090 GPUs. We implement our framework on the MMDetection library and use the Cascade R-CNN [47] as the basic framework to perform RGB-IR object detection. The detector is trained with an initial learning rate of 2×10^{-4} for 48 epochs. The batch size is set to 16, and the AdamW [72] optimizer is employed with a weight decay of 0.1. Horizontal flipping is also used for data augmentation.

Results on FLIR and LLVIP datasets. We compare our method with five common mono-modality methods and four competitive multi-modality methods. As shown in Table I, it can be seen that most of the multi-modality detectors are even worse than mono-modality detectors (*eg.* Faster R-CNN in IR modality). Since the RGB feature interfere with infrared feature under limited illumination conditions, it has a negative impact on the fused features utilized for object detection tasks. However, our UniRGB-IR can effectively solve this problem through the SFI module, enabling the detector to achieve better classification and localization processes.

Results on KAIST dataset. The quantitative results of the different methods on the KAIST dataset are shown in Table II. The experiments are conducted under 'All-dataset' settings [53]. We compare our UniRGB-IR with thirteen multi-modal object detection methods. Our model achieves the best performance on the 'All', 'Day', and 'Night' conditions and the other four of five subsets ('Near', 'Far', 'None', 'Partial' and 'Heavy'), and rank second in the 'Medium' subset. Furthermore, our detector surpasses the previous best competitor C^2Former by 3.18% on the 'All' condition, which indicates UniRGB-IR is robust to the complex scenes. Intuitively, we draw the log-average Miss Rate (MR) over the False Positive Per Image (FPPI) curve of these detectors as shown in Figure 5. The above results verify the superiority of our UniRGB-IR.

B. RGB-IR Semantic Segmentation

Datasets. Our semantic segmentation experiments are performed on the two public RGB-IR semantic segmentation datasets: PST900 [63] and MFNet [18]. The PST900 dataset is divided into 597 pairs for training and 288 pairs for testing,

TABLE IV
RESULTS ON THE MFNET DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **RED** AND THE SECOND-BEST ARE HIGHLIGHTED IN **BLUE**. "–" INDICATES THAT THE AUTHOR DID NOT THE CORRESPONDING RESULTS.

Methods	Unlabeled		Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
MFNet [18]	98.7	96.9	77.2	65.9	67.0	58.9	53.9	42.9	36.2	29.9	12.5	9.9	0.1	0.0	30.3	25.2	30.0	27.7	45.1	39.7
RTFNet [61]	**99.4**	**98.5**	93.0	87.4	79.3	70.3	76.8	62.7	60.7	45.3	38.5	29.8	0.0	0.0	45.5	29.1	74.7	55.7	63.1	53.2
PSTNet [63]	98.8	97.4	89.4	77.0	74.8	67.2	65.8	50.8	42.5	35.2	34.6	25.2	0.0	0.0	32.7	29.9	61.1	55.0	55.5	48.6
FuseSeg [73]	99.0	97.6	93.1	87.9	81.4	71.7	78.5	64.6	68.4	44.8	29.1	22.7	63.7	6.4	55.8	46.9	66.4	47.9	70.6	54.5
ABMDRNet [74]	98.6	97.8	94.3	84.4	90.0	69.6	75.7	60.3	64.0	45.1	44.1	33.1	31.0	5.1	61.7	47.4	66.2	50.0	69.5	54.8
MFFENet [75]	**99.3**	97.8	91.4	87.1	82.6	**74.4**	76.7	61.3	58.7	45.6	44.9	30.6	60.0	5.2	64.4	**57.0**	72.7	40.5	72.3	55.5
EGFNet [67]	98.7	98.0	**95.8**	87.6	89.0	69.8	80.6	58.8	**71.5**	42.8	48.7	33.8	33.6	7.0	65.3	8.3	71.1	47.1	72.7	54.8
MTANet [76]	98.4	98.0	**95.8**	88.1	**90.9**	71.5	80.3	60.7	**75.3**	40.9	**62.8**	38.9	38.7	13.7	63.8	45.9	70.8	47.2	75.2	56.1
MLFNet [77]	-	97.3	-	82.3	-	68.1	-	**67.3**	-	27.3		30.4	-	15.7	-	55.6	-	40.1	-	53.8
FEANet [78]	**99.3**	98.2	93.3	87.8	82.7	71.1	76.7	61.1	65.5	**46.5**	26.6	22.1	**70.8**	6.6	**66.6**	55.4	**77.3**	48.9	73.2	55.3
GMNet [79]	99.2	97.5	94.1	86.5	83.0	73.1	76.9	61.7	59.7	44.0	55.0	**42.3**	**71.2**	14.5	54.7	48.7	73.1	47.4	74.1	57.3
CCFFNet [30]	98.8	**98.3**	94.5	**89.6**	83.6	74.2	73.2	63.1	67.2	**50.5**	38.7	31.9	30.6	4.8	55.2	49.7	72.9	56.3	68.3	57.6
CCAFFMNet [80]	97.4	95.2	**95.0**	88.6	86.1	72.9	**82.1**	67.1	71.2	45.9	32.1	24.8	57.1	17.8	58.0	50.1	76.2	**57.8**	72.8	57.8
CMX [31]	-	**98.3**	-	**89.4**	-	**74.8**	-	64.7	-	47.3	-	30.1	-	8.1	-	52.4	-	**59.4**	-	**58.2**
FDCNet [65]	98.8	98.2	94.1	87.5	**91.4**	72.4	78.1	61.7	70.1	43.8	34.4	27.2	61.5	7.3	64.0	52.0	74.5	56.6	74.1	56.3
ECGFNet [81]	**99.3**	97.5	89.4	83.5	85.2	72.1	72.9	61.6	62.8	40.5	44.8	30.8	45.2	11.1	57.2	49.7	65.1	50.9	69.1	55.3
LASNet [82]	97.6	97.4	94.9	84.2	81.7	67.1	**82.1**	56.9	70.7	41.1	56.8	39.6	59.5	**18.9**	58.1	48.8	**77.2**	40.1	**75.4**	54.9
UniRGB-IR (Ours)	98.3	97.2	94.0	83.7	88.7	64.9	**88.0**	**69.8**	53.3	36.8	**58.5**	**41.0**	69.7	**36.7**	**77.6**	**56.2**	55.2	47.3	**75.7**	**59.3**

TABLE V
QUANTITATIVE COMPARISON OF RGB-IR SALIENT OBJECT DETECTION TASK ON VT821, VT1000 AND VT5000 DATASETS. * REPRESENTS RGB-D SOD TRANSFORMED INTO RGB-T SOD. THE BEST RESULTS ARE HIGHLIGHTED IN **RED** AND THE SECOND-BEST ARE HIGHLIGHTED IN **BLUE**.

Model	VT821				VT1000				VT5000			
	$S\uparrow$	$adpE\uparrow$	$adpF\uparrow$	$MAE\downarrow$	$S\uparrow$	$adpE\uparrow$	$adpF\uparrow$	$MAE\downarrow$	$S\uparrow$	$adpE\uparrow$	$adpF\uparrow$	$MAE\downarrow$
MMCI* [83]	0.763	0.784	0.618	0.087	0.886	0.892	0.803	0.039	0.827	0.859	0.714	0.055
TANet* [84]	0.818	0.852	0.717	0.052	0.902	0.912	0.838	0.030	0.847	0.883	0.754	0.047
S2MA* [85]	0.811	0.813	0.709	0.098	0.918	0.912	0.848	0.029	0.853	0.864	0.743	0.053
JLDCF* [86]	0.839	0.830	0.726	0.076	0.912	0.899	0.829	0.030	0.861	0.860	0.739	0.050
MTMR [87]	0.725	0.815	0.662	0.109	0.706	0.836	0.715	0.119	0.680	0.795	0.595	0.114
M3S-NIR [88]	0.723	0.859	0.734	0.140	0.726	0.827	0.717	0.145	0.652	0.780	0.575	0.168
SGDL [89]	0.765	0.847	0.731	0.085	0.787	0.856	0.764	0.090	0.750	0.824	0.672	0.089
FMSF [90]	0.760	0.796	0.640	0.080	0.873	0.899	0.823	0.037	0.814	0.864	0.734	0.055
MIDD [91]	0.871	0.895	0.803	0.033	0.915	0.933	0.880	0.027	0.868	0.896	0.799	0.043
ADF [92]	0.810	0.842	0.717	0.077	0.910	0.921	0.847	0.034	0.864	0.891	0.778	0.048
LSNet [93]	0.877	0.911	0.827	0.033	0.924	0.936	0.887	0.022	0.876	0.916	0.827	0.036
UniRGB-IR (Ours)	0.881	0.895	0.806	0.039	0.939	0.943	0.894	0.018	0.906	0.935	0.849	0.027

containing five categories (background, fire extinguisher, backpack, hand drill, and survivor). The MFNet dataset provides 1,569 pairs and contains nine classes (Car, Curve, Person, Guardrail, Car stop, Bike, Bump, Color cone, and Background). The dataset is divided into three parts: training, verification and testing according to the ratio of 2:1:1.

Metrics. Two metrics are utilized to evaluate the performance of semantic segmentation, namely mean accuracy (mAcc) and mean intersection over union (mIoU). Both metrics are calculated by averaging the ratios of the intersection and union of all categories.

Settings. Similarly, as the RGB-IR object detection task, we incorporate our method into the SETR [94] basic framework and implement it on the MMSegmentation library. The fine-tuning process spins a total of 10K iterations with an initial learning rate of 0.01. We employ the SGD optimizer and set the batch size to 16.

Results. The quantitative results of the different RGB-IR segmentation methods on the PST900 and MFNet dataset are shown in Table III and Table IV respectively. The comparison results show that our model significantly outperforms other methods. Specifically, our model obtains the best performance in terms of both the mACC and mIoU. In PST900 dataset, our model performs competitive performance in the Backpack, Hand-Drill and Survivor, outperforming the second-best methods by 2.3%, 0.6% and 0.4% IoU, respectively. In predicting Color Cone class on MFNet dataset, our model achieves the first and second best in terms of both evaluation metrics, with about 11% Acc higher than that of the second-best. These results show that our UniRGB-IR can also be used flexibly for

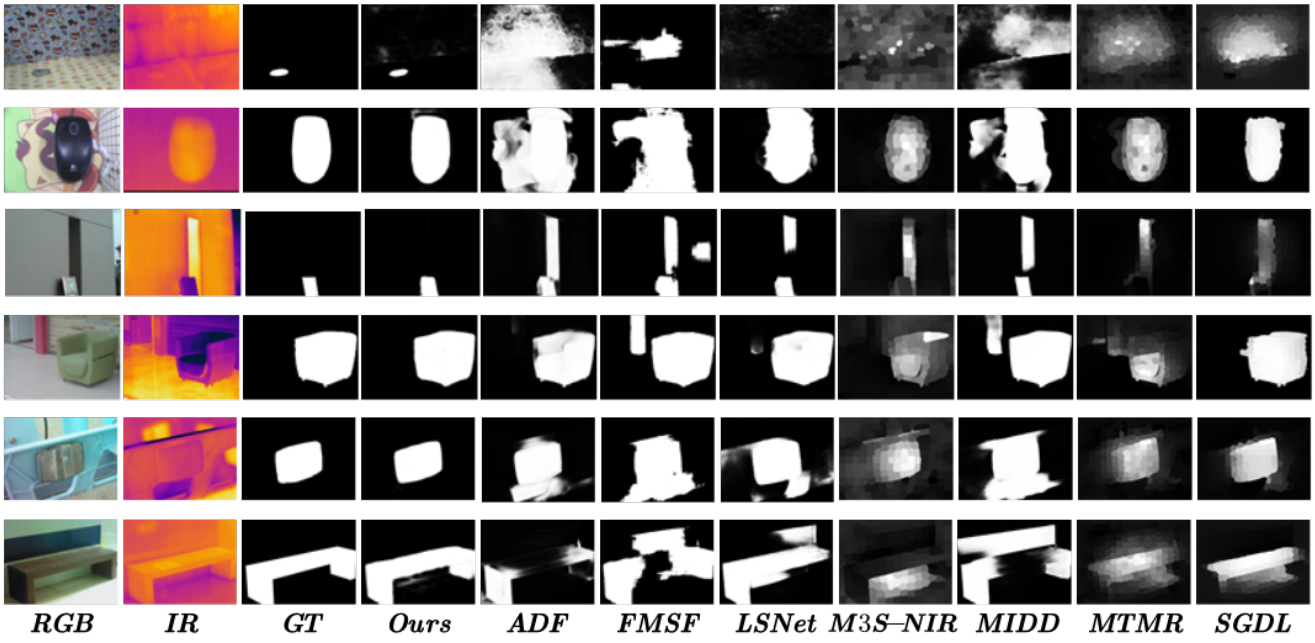

Fig. 6. Saliency maps of proposed UniRGB-IR and seven representative SOTA RGB-IR SOD methods in difference scenes. From top to bottom are visualization examples from the VT821, VT1000 and VT5000 datasets respectively.

TABLE VI
ABLATION STUDIES OF KEY COMPONENTS ON THE FLIR AND PST900 DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

MFP	SFI	FLIR			PST900	
		mAP	mAP_{50}	mAP_{75}	mAcc	mIoU
		39.6	76.3	35.9	90.1	77.5
✓		41.2	77.9	36.5	91.7	78.8
✓	✓	**44.1**	**81.4**	**40.2**	**94.3**	**82.8**

TABLE VII
ABLATION OF ADDING SFI MODULE TO DIFFERENT STAGES.

Different Stages with SFI module				mAP	mAP_{50}	mAP_{75}
Stage 1	Stage 2	Stage 3	Stage 4			
✓				41.7	80.0	37.2
✓	✓			43.3	**81.6**	39.7
✓	✓	✓		**44.1**	81.4	**40.2**
✓	✓	✓	✓	42.7	79.9	38.2

TABLE VIII
ABLATION OF THE DIFFERENT ATTENTION MECHANISMS.

Attention Mechanism	Complexity	mAP	mAP_{50}	mAP_{75}
Global Attention [95]	Quadratic	36.1	72.8	30.6
Pale Attention [43]	Linear	31.6	68.4	24.6
Deformable Attention [44]	Linear	**44.1**	**81.4**	**40.2**

segmentation tasks without any modification, which validates its universality.

C. RGB-IR Salient Object Detection

Datasets. Our salient object detection (SOD) experiments are performed on the three public datasets: VT821 [87], VT1000 [89] and VT5000 [92]. The VT821 dataset includes 821 registered RGB and IR images. The VT1000 dataset contains 1000 registered RGB-IR images with simple scenes and aligned images. The VT5000 dataset is a recent large-scale RGB-IR dataset, including a full-day scene under various limited light conditions. As usual in [92], we utilize 2500 image pairs in the VT5000 dataset as the training dataset, and the remaining image pairs along with the image pairs from the VT821 and VT1000 datasets are used as the test datasets.

Metrics. Four metrics are utilized to evaluate the performance of salient object detection namely F-measure (*adpF* ↑), E-Measure (*adpE* ↑), S-Measure (*S* ↑) and Mean absolute error (*MAE* ↓). ↑ and ↓ denote the higher the better and the lower the better, respectively.

Settings. As same as the RGB-IR semantic segmentation task, we incorporate our method into the SETR basic framework and also implement it on the MMSegmentation library. The fine-tuning process spins a total of 10K iterations with an initial learning rate of 0.01. We use the SGD optimizer and set the batch size to 64. For convenience, all input images are resized to 224 × 224 for testing.

Quantitative Comparison. Table V reports the quantitative comparison results. As can be seen from Table V, our UniRGb-IR outperforms SOTA methods both on VT1000 and VT5000 datasets in all evaluation metrics. Specifically, the S, $adpE$, $adpF$ and MAE matrics of our UniRGB-IR achieve 0.906, 0.935, 0.849 and 0.027 on VT5000, all of which are higher than the previous competitor LSNet [93]. These remarkable results clearly indicate that the saliency maps predicted by UniRGB-IR are close to the corresponding ground-truths, which verifies the effectiveness of our method.

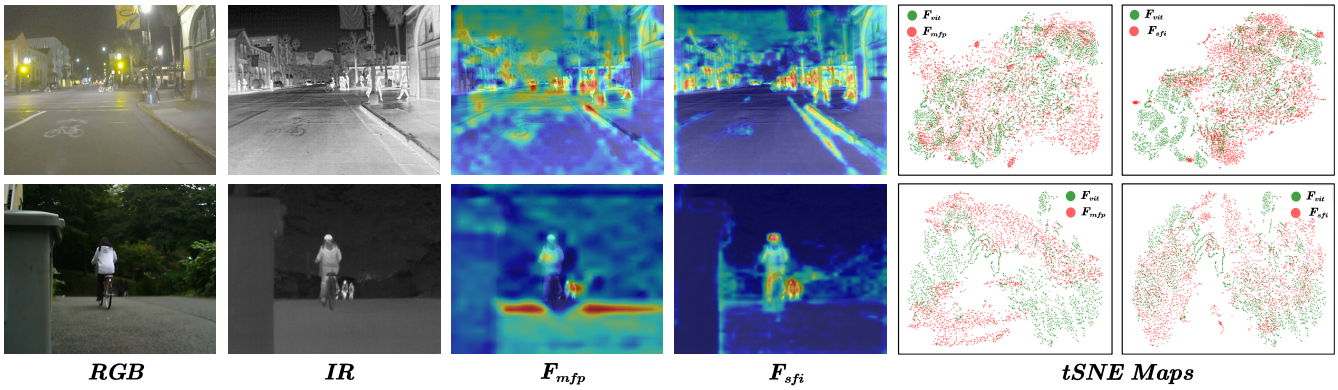

Fig. 7. Visualization of intermediate results. The F_{mfp} and F_{sfi} features from the first stage are visualized in the third and fourth columns. The tSNE visualizations are also shown in the last two columns.

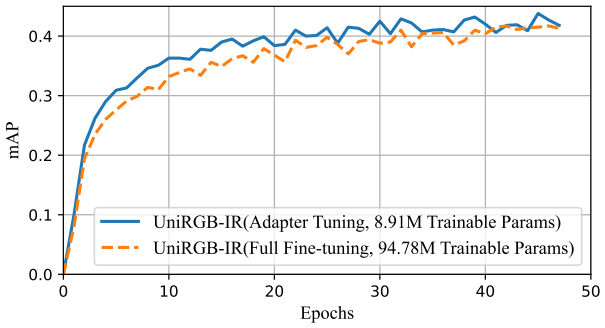

Fig. 8. Training efficiency analysis on the FLIR dataset.

Qualitative Evaluation. Qualitative RGB-IR SOD comparisons are shown in Figure 6. The first and third rows show scenes containing small objects. The second and fifth rows show the salient object with the complex backgrounds. All these scenes are suitably handled by our UniRGB-IR, which verifies that our UniRGB-IR is more effective and robustness while providing high-quality results with appropriate detection capabilities.

D. Ablation Study

Ablation for components. To investigate the contribution of the SFI and MFP modules, we gradually add each module to the baseline model as shown in Table VI. We stack RGB and IR images and feed them into the Cascade R-CNN detector and SETR framework (ViT backbone) for full fine-tuning on the corresponding tasks, which are considered as our baseline models. Then, we freeze the ViT model and introduce the MFP module to extract contextual multi-scale features (element-wise addition), which results in 1.6% mAP and 1.3% mIoU improvements. Finally, we replace the element-wise addition operation with the SFI module and further improve the mAP and mIoU metrics by 2.9% and 4.0% respectively, which achieve the best performance on both datasets.

SFI module at different stages. We add the SFI module to the ViT pre-trained model at the beginning of different stages. From Table VII, we can find that by adding the SFI module at the first stage, the detector achieves 41.7% mAP on FLIR dataset. After adding the SFI module in the second and third stages, the performance further improved by about 2% and 3% mAP, respectively. However, continuing to add it to the final stage will reduce the detection performance while also increasing computational overhead. Therefore, we add the SFI module from the first stage to the third stage of ViT model.

Attention type in SFI module. Since the attention mechanism in our SFI module is replaceable, we adopt three popular attention mechanisms in our UniRGB-IR to discuss their impact on model performance. As shown in Table VIII, the detector achieves the best performance with linear complexity by utilizing deformable attention. Thus, the deformable attention is more suitable for our framework and utilized as the default configuration. It is worth noting that it can be replaced by other attention mechanisms to further achieve superior performance.

E. Visualization Analysis

Intermediate results. To illustrate the effectiveness of the SFI module, we visualize the intermediate results on the FLIR dataset. From F_{mfp} and F_{sfi} in Figure 7, we can find that the foreground objects in the F_{sfi} become salient through the SFI module. Furthermore, we also visualize the tSNE maps of F_{mfp}–F_{vit} and F_{sfi}–F_{vit} respectively. After using the SFI module, the distribution of injected features F_{sfi} is more concentrated thain the distribution of ViT features F_{vit}, indicating that the required richer RGB-IR features can be well supplemented into the ViT model through the SFI module.

Training efficiency. We further plot the mAP curves for each epoch of UniRGB-IR with different training paradigm to demonstrate the efficiency of UniRGB-IR, as shown in Figure 8. During the training process, all hyperparameters of the two models are the same. From Figure 8, we can find that the convergence speed of the adapter tuning paradigm surpasses that of the full fine-tuning strategy. Moreover, by utilizing the adapter tuning paradigm, our UniRGB-IR achieves superior performance with a smaller number of trainable parameters (about 10% of the full fine-tuning model). The above results verify the efficiency of our method.

V. Conclusion

In this paper, we proposed an efficient and scalable framework (named UniRGB-IR) to unify RGB-IR downstream

tasks. Our framework contains a Multi-modal Feature Pool module and a Supplementary Feature Injector module. The former extracts contextual multi-scale features from two modality images, and the latter adaptively injects the features into the transformer model. These two modules can be efficiently optimized to complement the pre-trained foundation model with richer RGB-IR features. To evaluate the effectiveness of our method, we incorporated the ViT-Base model into the framework as the pre-trained foundation model and performed various RGB-IR downstream tasks. Extensive experiments verify that our UniRGB-IR can be effectively leveraged as a unified framework for RGB-IR downstream tasks. We believe that our method can be applied to more multi-modal real-world applications.

REFERENCES

[1] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial intelligence review*, vol. 53, pp. 5455–5516, 2020.

[2] H. Yan, B. Li, H. Zhang, and X. Wei, "An antijamming and lightweight ship detector designed for spaceborne optical images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4468–4481, 2022.

[3] L. Bo, X. Xiaoyang, W. Xingxing, and T. Wenting, "Ship detection and classification from optical remote sensing images: A survey," *Chinese Journal of Aeronautics*, vol. 34, no. 3, pp. 145–163, 2021.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[5] C. Xia, X. Wang, F. Lv, X. Hao, and Y. Shi, "Vit-comer: Vision transformer with convolutional multi-scale feature interaction for dense predictions," *arXiv preprint arXiv:2403.07392*, 2024.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[8] X. Wei, Y. Huang, Y. Sun, and J. Yu, "Unified adversarial patch for visible-infrared cross-modal attacks in the physical world," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[9] Y. Shi, F. Lv, X. Wang, C. Xia, S. Li, S. Yang, T. Xi, and G. Zhang, "Open-transmind: A new baseline and benchmark for 1st foundation model challenge of intelligent transportation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6327–6334.

[10] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[11] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5363–5371.

[12] M. Yuan and X. Wei, "C2former: Calibrated and complementary transformer for rgb-infrared object detection," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2024.

[13] T. Zhao, M. Yuan, and X. Wei, "Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion," *arXiv preprint arXiv:2401.10731*, 2024.

[14] L. Zhang, Z. Liu, X. Zhu, Z. Song, X. Yang, Z. Lei, and H. Qiao, "Weakly aligned feature fusion for multimodal object detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[15] L. Liu, J. Chen, H. Wu, G. Li, C. Li, and L. Lin, "Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4823–4833.

[16] M. Yuan, Y. Wang, and X. Wei, "Translation, scale and rotation: cross-modal alignment meets rgb-infrared vehicle detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 509–525.

[17] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4486–4497, 2021.

[18] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5108–5115.

[19] A. C. Stickland and I. Murray, "Bert and pals: Projected attention layers for efficient adaptation in multi-task learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5986–5995.

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

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[22] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6824–6835.

[23] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3041–3050.

[24] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[25] L. Zhang, Z. Liu, S. Zhang, X. Yang, H. Qiao, K. Huang, and A. Hussain, "Cross-modality interactive attention network for multispectral pedestrian detection," *Information Fusion*, vol. 50, pp. 20–29, 2019.

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

[27] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5127–5137.

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

[29] M. Yuan, X. Shi, N. Wang, Y. Wang, and X. Wei, "Improving rgb-infrared object detection with cascade alignment-guided transformer," *Information Fusion*, vol. 105, p. 102246, 2024.

[30] W. Wu, T. Chu, and Q. Liu, "Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation," *Pattern Recognition*, vol. 131, p. 108881, 2022.

[31] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[32] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Caver: Cross-modal view-mixed transformer for bi-modal salient object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 892–904, 2023.

[33] W. Zhou, F. Sun, Q. Jiang, R. Cong, and J.-N. Hwang, "Wavenet: Wavelet network with knowledge distillation for rgb-t salient object detection," *IEEE Transactions on Image Processing*, 2023.

[34] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," *arXiv preprint arXiv:2205.08534*, 2022.

[35] D. Yin, Y. Yang, Z. Wang, H. Yu, K. Wei, and X. Sun, "1% vs 100%: Parameter-efficient low rank adapter for dense predictions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 116–20 126.

[36] J. Park, H. Seo, and E. Yang, "Pc-adapter: Topology-aware adapter for efficient domain adaption on point clouds with rectified pseudo-label," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 530–11 540.

[37] Y.-L. Sung, J. Cho, and M. Bansal, "Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5227–5237.

[38] U. Upadhyay, S. Karthik, M. Mancini, and Z. Akata, "Probvlm: Probabilistic adapter for frozen vison-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1899–1910.

[39] Z. He, Y. Cao, L. Du, B. Xu, J. Yang, Y. Cao, S. Tang, and Y. Zhuang, "Mrfn: Multi-receptive-field network for fast and accurate single image super-resolution," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1042–1054, 2019.

[40] X. Wu, D. Hong, and J. Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 364–376, 2022.

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

[42] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[43] S. Wu, T. Wu, H. Tan, and G. Guo, "Pale transformer: A general vision transformer backbone with pale-shaped attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2731–2739.

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

[45] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

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

[47] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[48] S. Zhang, X. Wang, J. Wang, J. Pang, C. Lyu, W. Zhang, P. Luo, and K. Chen, "Dense distinct query for end-to-end object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7329–7338.

[49] H. Zhang, E. Fromont, S. Lefèvre, and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 72–80.

[50] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *European Conference on Computer Vision*. Springer, 2022, pp. 139–158.

[51] X. Zuo, Z. Wang, Y. Liu, J. Shen, and H. Wang, "Lgadet: Lightweight anchor-free multispectral pedestrian detection with mixed local and global attention," *Neural Processing Letters*, vol. 55, no. 3, pp. 2935–2952, 2023.

[52] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, "Multimodal object detection by channel switching and spatial attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023, pp. 403–411.

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

[54] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[55] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 49–56.

[56] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster r-cnn for robust multispectral pedestrian detection," *Pattern Recognition*, vol. 85, pp. 161–171, 2019.

[57] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, pp. 148–157, 2019.

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

[59] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 787–803.

[60] Q. Li, C. Zhang, Q. Hu, H. Fu, and P. Zhu, "Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection," *IEEE Transactions on Multimedia*, 2022.

[61] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[62] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[63] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "Pst900: Rgb-thermal calibration, dataset and segmentation network," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 9441–9447.

[64] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *2019 IEEE International conference on image processing (ICIP)*. IEEE, 2019, pp. 1440–1444.

[65] S. Zhao and Q. Zhang, "A feature divide-and-conquer network for rgb-t semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[66] Y. Fu, Q. Chen, and H. Zhao, "Cgfnet: cross-guided fusion network for rgb-thermal semantic segmentation," *The Visual Computer*, vol. 38, no. 9, pp. 3243–3252, 2022.

[67] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for rgb–thermal scene parsing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 3571–3579.

[68] P. Li, J. Chen, B. Lin, and X. Xu, "Residual spatial fusion network for rgb-thermal semantic segmentation," *arXiv preprint arXiv:2306.10364*, 2023.

[69] Y. Wang, G. Li, and Z. Liu, "Sgfnet: semantic-guided fusion network for rgb-thermal semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[70] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," in *2020 IEEE International conference on image processing (ICIP)*. IEEE, 2020, pp. 276–280.

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

[72] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[73] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 3, pp. 1000–1011, 2020.

[74] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2633–2642.

[75] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, "Mffenet: Multiscale feature fusion and enhancement network for rgb–thermal urban road scene parsing," *IEEE Transactions on Multimedia*, vol. 24, pp. 2526–2538, 2021.

[76] W. Zhou, S. Dong, J. Lei, and L. Yu, "Mtanet: Multitask-aware network with hierarchical multimodal fusion for rgb-t urban scene understanding," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 48–58, 2022.

[77] Z. Guo, X. Li, Q. Xu, and Z. Sun, "Robust semantic segmentation based on rgb-thermal in variable lighting scenes," *Measurement*, vol. 186, p. 110176, 2021.

[78] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, and T. L. Lam, "Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4467–4473.

[79] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021.

[80] S. Yi, J. Li, X. Liu, and X. Yuan, "Ccaffmnet: Dual-spectral semantic segmentation network with channel-coordinate attention feature fusion module," *Neurocomputing*, vol. 482, pp. 236–251, 2022.

[81] W. Zhou, Y. Lv, J. Lei, and L. Yu, "Embedded control gate fusion and attention residual learning for rgb–thermal urban scene parsing," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[82] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "Rgb-t semantic segmentation with location, activation, and sharpening," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1223–1235, 2022.

[83] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection," *Pattern Recognition*, vol. 86, pp. 376–385, 2019.

[84] H. Chen and Y. Li, "Three-stream attention-aware network for rgb-d salient object detection," *TIP*, vol. 28, no. 6, pp. 2825–2835, 2019.

[85] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *CVPR*, 2020, pp. 13 756–13 765.

[86] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for rgb-d salient object detection and beyond," *TPAMI*, vol. 44, no. 9, pp. 5541–5559, 2021.

[87] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, "Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *IGTA*. Springer, 2018, pp. 359–369.

[88] Z. Tu, T. Xia, C. Li, Y. Lu, and J. Tang, "M3s-nir: Multi-modal multi-scale noise-insensitive ranking for rgb-t saliency detection," in *MIPR*. IEEE, 2019, pp. 141–146.

[89] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "Rgb-t image saliency detection via collaborative graph learning," *TMM*, vol. 22, no. 1, pp. 160–173, 2019.

[90] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "Rgb-t salient object detection via fusing multi-level cnn features," *TIP*, vol. 29, pp. 3321–3335, 2019.

[91] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive dual-decoder for rgb-thermal salient object detection," *TIP*, vol. 30, pp. 5678–5691, 2021.

[92] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgbt salient object detection: A large-scale dataset and benchmark," *TMM*, 2022.

[93] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images," *TIP*, vol. 32, pp. 1329–1340, 2023.

[94] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *CVPR*, 2021, pp. 6881–6890.

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