Explicit Attention-Enhanced Fusion for RGB-Thermal Perception Tasks

Mingjian Liang*, Junjie Hu*, Chenyu Bao, Hua Feng, Fuqin Deng and Tin Lun Lam†

***Abstract*—Recently, RGB-Thermal based perception has shown significant advances. Thermal information provides useful clues when visual cameras suffer from poor lighting conditions, such as low light and fog. However, how to effectively fuse RGB images and thermal data remains an open challenge. Previous works involve naive fusion strategies such as merging them at the input, concatenating multi-modality features inside models, or applying attention to each data modality. These fusion strategies are straightforward yet insufficient. In this paper, we propose a novel fusion method named Explicit Attention-Enhanced Fusion (EAEF) that fully takes advantage of each type of data. Specifically, we consider the following cases: i) both RGB data and thermal data, ii) only one of the types of data, and iii) none of them generate discriminative features. EAEF uses one branch to enhance feature extraction for i) and iii) and the other branch to remedy insufficient representations for ii). The outputs of two branches are fused to form complementary features. As a result, the proposed fusion method outperforms state-of-the-art by 1.6% in mIoU on semantic segmentation, 3.1% in MAE on salient object detection, 2.3% in mAP on object detection, and 8.1% in MAE on crowd counting. The code is available at https://github.com/FreeformRobotics/EAEFNet.**

***Index Terms*—Multi-modality data fusion, RGB-Thermal fusion, RGB-thermal perception**

I. Introduction

OVER the last decade, we have witnessed significant progress on many perception tasks. Based on data-driven learning, deep neural networks (DNNs) can learn to estimate semantic maps [25], object categories [30], depth maps [15], etc., from only RGB images. This paradigm has continuously boosted perception tasks for robots in which various models, loss functions, and learning strategies have been explored.

However, current methods highly depend on the quality of RGB images. In reality, visual cameras are particularly susceptible to noises [33], poor lighting [14], weather [23], etc. In these cases, DNNs tend to degrade their performance significantly. To handle these issues, researchers sought to employ thermal data to complement RGB images and developed different multi-modality fusion strategies.

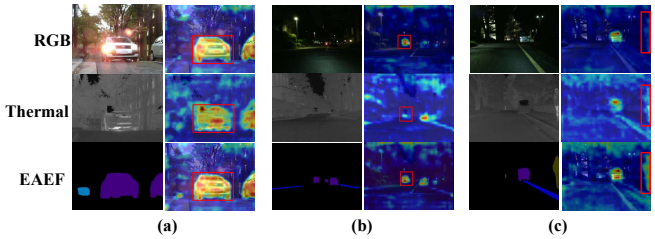

Fig. 1. Visualization of extracted features of RGB, thermal images, and the proposed fusion method, where (a) both RGB and thermal data, (b) only one of them, and (c) none, can yield distinct features. As seen, the proposed fusion method can boost feature extraction for all three cases.

The core of RGB-T based methods is the fusion strategy of RGB data and thermal data. Previous methods [24], [55] directly combine them at the input. Some works [9], [34] use two separate encoders for extracting features from RGB and thermal images, respectively. Then, these features are merged and outputted to a decoder to yield a final prediction. Recently, most studies [52], [5] attempted to utilize the attention mechanism for multi-modality data fusion. These approaches commonly apply channel attention to intermediate features of different data types and obtain the fused features by weighing their importance. However, these fusion strategies are implicit and insufficient. In particular, it is unclear how multi-modality data can (or cannot) complement each other.

Different from existing studies, we explicitly take multi-modality data fusion under three circumstances: i) both RGB and thermal images can extract useful features, as Fig. 1 (a), ii) only one of them can generate meaningful representations, as Fig. 1 (b), and iii) none of them provides useful features, as Fig. 1 (c). In this paper, we propose the Explicit Attention-Enhanced Fusion (EAEF) that performs a more effective fusion. The key inspiration of EAEF is the case-specific design that uses one branch to stick to meaningful representations for i) and enhance feature extraction for iii), and the other branch to force CNNs to pay attention to insufficient representations for ii). One of these branches will generate useful features at least, and their combination will yield complementary features for a final prediction.

To validate the effectiveness of this novel multi-modality fusion method, we design a novel RGB-T framework by integrating EAEF into an encoder-decoder network and evaluate it on various vision tasks with benchmark datasets, including semantic segmentation, object detection, and crowd counting. We confirm through extensive experiments that our method is

Manuscript received: December 15, 2022; Revised: March 17, 2023; Accepted: April 16, 2023. This paper was recommended for publication by Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers' comments.

J. Hu, H. Feng, F. Deng, T.L. Lam are with the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

M. Liang, C. Bao and T.L. Lam are with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen.

* indicates equal contribution.

†Corresponding author: Tin Lun Lam tllam@cuhk.edu.cn

This work was partly supported by the National Natural Science Foundation of China (62073274), Shenzhen Science and Technology Program (JCYJ20220818103000001), and the funding AC01202101103 from the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

Digital Object Identifier (DOI): see top of this page.

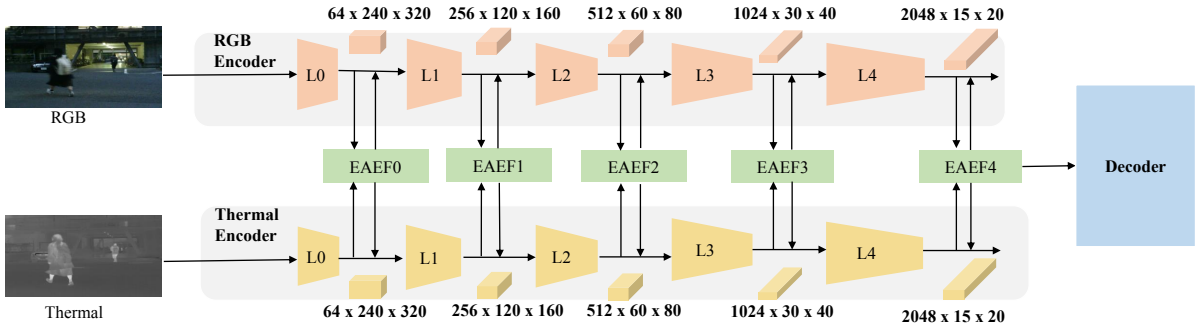

Fig. 2. Diagram of encoder-decoder based network assembled with the proposed EAEF for dense prediction tasks. EAEF is used for fusing features from the RGB branch and the thermal branch at multi-scales.

more effective for RGB-T based vision perception.

In summary, our contributions are:

- A novel multi-modality fusion method that effectively fuses RGB features and thermal features in an explicit manner.
- An effective encoder-decoder based network assembled with the proposed feature fusion strategy for dense prediction tasks.
- State-of-the-art performance on semantic segmentation, object detection, salient detection, and crowd counting with open-source codes.

The remainder of this paper is organized as follows. Section II reviews related studies. Section III presents the framework and the proposed multi-modality data fusion method. Section IV provides quantitative and qualitative experimental results on three tasks. Section V concludes our work.

II. Related Work

Using additional modality data to complement RGB images has shown great improvement in accuracy. Many works attempted to fuse depth and RGB data using attention mechanisms [22], [13], [56] . For example, FRNet [63] and HFNet[60] apply channel attention for RGB-D fusion. BCINet [64] uses spatial attention in their BCIM (Bilateral cross-modal interaction module) to capture cross-modal complementary features. RLLNet [59] leverages both channel and spatial mechanisms in the decoder. Similarly, methods of RGB-T perception have explored the fusion strategy between RGB data and thermal data. In this paper, we mainly review related works from the perspective of feature fusion.

a) RGB-T semantic segmentation: Early works fuse RGB features and thermal features in a straightforward way by applying element-wise addition, such as MFNet [9], RTFNet [34], FuseSeg [35], neglecting the difference of features in their importance on discriminability. Recent works solve this issue by employing channel attention as in ABMDRNet [52], both channel and spatial attention as in FEANet [5] and GMNet [61], self-attention as in MFTNet [57]. Besides, the domain adaptation from day to night using thermal images has also been studied in [43], [18].

b) RGB-T Salient Objection detection: Attention mechanism has also been frequently used in the salient detection task, such as the multi-interaction module using channel attention proposed in MIDD [38]. In most cases, both channel and spatial attention are employed to weigh channels and pixels adaptively, e.g., cross-guidance fusion in SCGFNet [44], cross-modal attention mechanism in [51], Convolutional Block Attention Module (CBAM) in ADFNet [42]. Similarly, based on attentional considerations, LSNet [65] introduced attention selection to determine inter-attention at each pixel position.

c) RGB-T crowd counting: DEFNet [62] fuses the two modalities through simple element-wise addition. IADM [21] and TAFNet [36] introduce a gate mechanism for feature fusion. Not surprisingly, an attempt to incorporate attention mechanism has also been carried out by CSCA [53].

d) RGB-T objection detection: Existing works including TarDAL [24], CDDFuse [55], and U2F[46], directly combine RGB data and thermal data at inputs, without applying attention based fusion strategy.

However, there is no guarantee that existing methods utilizing attention can extract sufficiently effective features since feature generation and attention extraction are performed implicitly inside CNNs. In this paper, we propose a case-specific way to improve the classical channel attention mechanism by explicitly enhancing attention extraction.

III. Methodology

A. Framework Overview

We take the classical encoder-decoder network for dense prediction tasks. The framework consists of an image encoder, a thermal encoder, and a decoder. The proposed Explicit Attention-Enhanced Fusion (EAEF) is applied between the two encoders to fuse features at multi-scales. A diagram of our dense prediction network is given in Fig. 2, where we show a semantic segmentation network built on ResNet [11].

Note that the framework naturally uses different backbones on different tasks. Therefore, the detailed implementation is task-specific. Nevertheless, all models built on our framework have the same technical components, i.e., one RGB encoder, one thermal encoder, EAEF modules, and a decoder.

B. Explicit Attention-Enhanced Fusion

Suppose $F_{rgb} \in \mathbb{R}^{h \times w \times c}$ and $F_t \in \mathbb{R}^{h \times w \times c}$ are the features extracted from RGB encoder and thermal encoder at a certain

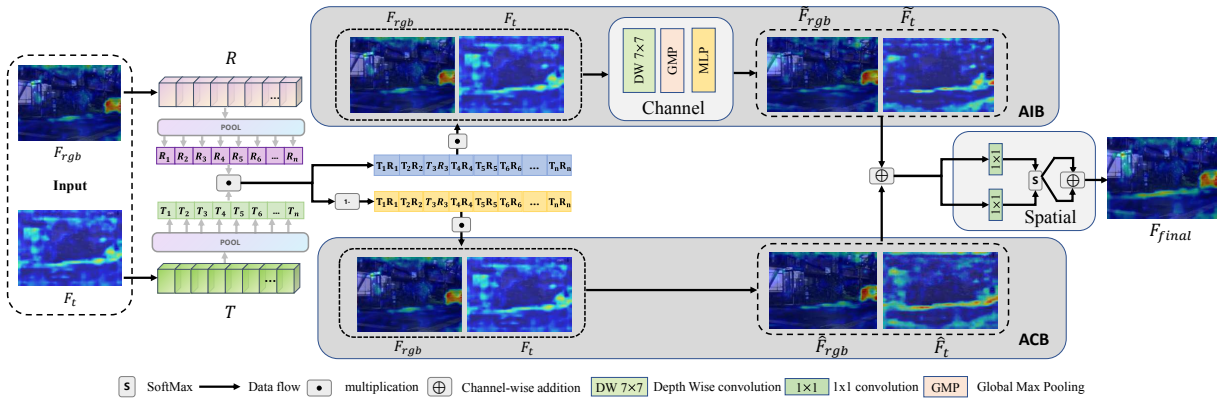

Fig. 3. Overview of the proposed Explicit Attention-Enhanced Fusion (EAEF). EAEF takes RGB features F_{rgb} from the RGB encoder and thermal features F_t from the thermal encoder as inputs, then applies attention interaction and attention complement with two branches. The final features are fused by merging outputs from these two branches. For simplicity, we only show one feature map extracted from the image encoder and the thermal encoder, respectively.

scale, the conventional way to extract channel-wise attention can be represented as

$$\begin{aligned} R &= \Big(f_{MLP}\Big(f_{GAP}(F_{rgb})\Big)\Big) \\ T &= \Big(f_{MLP}\Big(f_{GAP}(F_t)\Big)\Big) \end{aligned} \tag{1}$$

where $R \in \mathbb{R}^c$ and $T \in \mathbb{R}^c$ are extracted weights for RGB features and thermal features, respectively; f_{GAP} and f_{MLP} denote the global average pooling and MLP, respectively. For many previous works, the feature fusion is conducted by $\sigma(R) \circledast F_{rgb} + \sigma(T) \circledast F_t$, where σ is the sigmoid activation that generates channel-wise attention, $\circledast$ denotes channel-wise multiplication. However, this fusion method is effective only if either $\sigma(R)$ or $\sigma(T)$ has been activated sufficiently.

Differing from any existing approaches, we delve into this feature fusion by explicitly considering the interaction of multi-modality features. Specifically, we decompose the feature fusion into an Attention Interaction Branch (AIB) and an Attention Complement Branch (ACB), as shown in Fig. 3. The former handles cases where both RGB and thermal encoder or none of them capture discriminative features, and the latter tackles cases where only one encoder extracts useful features.

AIB takes an element-wise multiplication between R and T to generate correlated attention, then applies channel-wise multiplication to RGB and thermal features. It is represented as:

$$\begin{aligned} F'_{rgb} &= \sigma(c * (R \otimes T)) \circledast F_{rgb} \\ F'_t &= \sigma(c * (R \otimes T)) \circledast F_t \end{aligned} \tag{2}$$

where $\otimes$ denotes element-wise multiplication. c is the number of channels that plays a role of scaling factor for attention enhancement, such that $\sigma(c * (R \otimes T)) \geq \sigma(R)$ and $\sigma(c * (R \otimes T)) \geq \sigma(T)$.

For cases where only one modality data provides sufficiently discriminative features, i.e., $R \geq 0$, $T \leq 0$ or $R \leq 0$, $T \geq 0$, $\sigma(c \circledast (R \otimes T))$ tends to be small. Thus, we use the attention complement branch (ACB) that applies the enhancement by:

$$\begin{aligned} \hat{F}_{rgb} &= (1 - \sigma(c * (R \otimes T))) \circledast F_{rgb} \\ \hat{F}_t &= (1 - \sigma(c * (R \otimes T))) \circledast F_t \end{aligned} \tag{3}$$

such that $1 - \sigma(c * (R \otimes T)) \geq \sigma(R)$ and $1 - \sigma(c * (R \otimes T)) \geq \sigma(T)$.

At least one branch can generate stronger attention than the traditional channel-wise attention mechanism. However, if we directly combine $\hat{F}_{rgb}$ and F'_{rgb} or $\hat{F}_t$ and F'_t, it will lead to an identical mapping, e.g., $\hat{F}_{rgb} + F'_{rgb} = F_{rgb}$. To tackle this issue, we apply non-linear feature interaction within each branch.

Let $F'_{rgb,t}$ be the concatenation of F'_{rgb} and F'_t; then, AIB further performs multi-modality interaction between F'_{rgb} and F'_t by a data interaction module:

$$\tilde{F}_{rgb,t} = F'_{rgb,t} \circledast \sigma\Big(f_{MLP}\Big(f_{GMP}\Big(f_{dw}(F'_{rgb,t})\Big)\Big)\Big) \tag{4}$$

where f_{MLP} and σ denote MLP and sigmoid operations which are the same as Eq.(1), f_{dw} is depth-wise convolution, f_{GMP} is global max pooling operation. $\tilde{F}_{rgb,t}$ is the outputted features by the data interaction module, and it is further split to RGB features $\tilde{F}_{rgb}$ and thermal features $\tilde{F}_{rgb}$, respectively.

We experimentally found that the interaction module contributes less to the ACB regarding the model's performance. Therefore, we do not apply the interaction module in ACB to reduce the model complexity.

Then, the enhanced RGB and thermal features are obtained by aggregating outputs from AIB and ACB:

$$\begin{aligned} \overline{F}_{rgb} &= \tilde{F}_{rgb} + \hat{F}_{rgb} \\ \overline{F}_t &= \tilde{F}_t + \hat{F}_t \end{aligned} \tag{5}$$

Finally, we apply a spatial attention mechanism to merge the enhanced RGB and thermal features with a 1×1 convolutional layer. Formally, the merged features are obtained by:

$$F^*_{rgb,t} = \overline{F}_{rgb,t} \otimes SoftMax\Big(Conv_{1\times1}(\overline{F}_{rgb,t})\Big) \tag{6}$$

where $\overline{F}_{rgb,t}$ denotes the concatenated result of $\overline{F}_{rgb}$ and $\overline{F}_t$. The outputted features of EAEF are obtained by:

$$F_{final} = F^*_{rgb} + F^*_t \tag{7}$$

IV. Experimental Results

A. Semantic Segmentation

1) Datasets: MFNet dataset [9] is the most popular benchmark for RGB-T based semantic segmentation. It records nine semantic categories in urban street scenes, including one unlabeled background category and eight hand-labeled object categories. The dataset contains 1569 pairs of RGB and thermal images with a resolution of 640×480. Following RTFNet [34], we use 784 pairs of images for training, 392 pairs for validation, and the rest 420 pairs for testing.

PST900 dataset [54] is also a popular benchmark for RGB-T based semantic segmentation. It contains five semantic categories and 894 RGB-T image pairs with a resolution of 720×1280. Among them, 597 pairs are split for training, and the rest 297 pairs are used for testing.

2) Implementation Details and Evaluation Metrics: We employ a cascaded decoder structure based on BBSNet [6] with three modifications. First, we replace the asymmetric convolution in GCM with SELayer [12] to reduce the model parameters. Second, we introduce an ASPP [1] structure into Cascaded Decoder to better aggregate low-level features. Finally, we adjust the output dimension of the decoder to meet our nine-categories recognition requirement. We use the stochastic gradient descent (SGD)[28] optimization solver for training. The initial learning rate is set to 0.02, Momentum and weight decay are set to 0.9 and 0.0005, respectively. The batch size is set to 5, and we apply ExponentialLR to gradually decrease the learning rate. The loss function has a DiceLoss [27] term and a SoftCrossEntropy [47] term, each term is weighed with a scalar of 0.5. For MFNet dataset, we train the model with 100 epochs and use the best model on the validation set for evaluation. For PST900 dataset, we train the model with 60 epochs.

Same to previous works [34], we use two measures for quantifying results. The first is Accuracy (Acc) and the second one is Intersection Union (IoU). mAcc and mIoU are the averages over all categories.

3) Results:

a) Results on MFNet: We first conduct quantitative comparisons between the proposed method and other baseline approaches. We compare our method against existing approaches, including MFNet [9], FuseNet [10], RTFNet-152 [34], FusSeg-161 [35], FEANet [5], GMNet [61], MFTNet [57], PSTNet [31], RTFNet-50 [34], ABMDENet [52], EGFNet [7]. Since the model complexity is different for existing methods, we implement our method on two backbones, including a larger ResNet-152 and a smaller ResNet-50, for fair comparisons.

Table I shows the quantitative results. It is clear that our method achieves the best mean accuracy. As seen, when having a similarly smaller model complexity, our method beats PSTNet significantly. Besides, our method built on ResNet152 achieved superior performance for most categories, e.g., the second best performance on "Person", "Bike", "Curve", "'Bump' in IoU. Most importantly, the proposed method gained 0.4% and 1.6% improvements in mAcc and mIoU, respectively, against the current state-of-the-art MFTNet. The quantitative results verify that our method can extract better complementary cross-modality features.

Figure 4 exhibits the qualitative results under different lighting conditions. In general, we find that our method has the following advantages. First, our method demonstrates better results than existing approaches for both night and daytime conditions. It shows slightly better performance for daytime images and more superior results for nighttime images. Second, our method can capture the tiny objects both in RGB and thermal images more effectively, such as the pedestrian in the 3rd column and the bump on the road in the 5th column. These advantages validate the effectiveness of our strategy for multi-modality feature fusion.

Fig. 5 shows attention maps of three examples generated by our method and several baseline methods. It is shown that our method generates better attention than other approaches.

b) Results on PST900: We then conduct experiments on PST900 dataset. We compare our method with Efficient FCN [19], CCNet [17], ACNet [16], SA-Gate [2], RTFNet [34], PSTNet [31], and GMNet [61]. The quantitative results are given in Table II. It can be clearly seen that our method achieves the best results. It outperforms all previous methods, achieving 91.42 in mAcc and 85.56 in mIoU. Besides, it outperforms the state-of-the-art GMNet [61] by 1.81% in mACC and 1.44% in mIoU, respectively.

B. Object Detection

1) Dataset: M^3FD dataset [20]contains a set of auto-driving scenarios. It has 4200 pairs of RGB-T images, including 33603 annotated labels in six classes, including "People", "Car", "Bus", "Motorcycle", "Truck", and "Lamp". Moreover, the dataset was split into "Daytime", "Overcast", "Night", and "challenge" scenarios according to the characteristics of the environments.

2) Implementation Details and Evaluation Metric: We build a network for object detection by integrating EAEF into YoloV5 [3]. We use the stochastic gradient descent (SGD)[28] optimization solver for training. The initial learning rate is set to 0.01, Momentum and weight decay are set to 0.9 and 0.0005, respectively. The batch size is set to 32, and we apply ExponentialLR to gradually decrease the learning rate. The loss function has an IoULoss [48] term and a CrossEntropy [4] term. These two loss terms are weighed with a scalar of 0.3 and 0.7, respectively. For evaluation, we take the mAP@0.5 metric as TarDAL [20].

3) Results: The experimental results are shown in Table III. As seen, the method only using thermal data shows the worst performance. Nevertheless, for "Challenge" scenarios, it attains better performance than using RGB. Both TarDAL and our method obtain better accuracy compared to single modality data based methods. It is also observed that our method outperforms the other approaches by a good margin. We obtain 0.801 mAP, outperforming TarDAL by 2.3%.

TABLE I
QUANTITATIVE COMPARISONS ON THE MFNET DATASET. THE BEST RESULTS ARE SHOWN IN BOLD FONT.

Method	Flops.(G)	Params.(M)	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		
MFNet[9]	0.74	8.4	77.2	65.9	67.0	58.9	53.9	42.9	36.2	29.9	19.1	9.9	0.1	8.5	30.3	25.2	30.0	27.7	45.1	39.7
FuseNet(VGG16)[10]	44.17	284.0	81.0	75.6	75.2	66.3	64.5	51.9	51.0	37.8	17.4	15.0	0.0	0.0	31.1	21.4	51.9	45.0	52.4	45.6
PSTNet(ResNet18)[31]	123.4	105.8	-	76.8	-	52.6	-	55.3	-	29.6	-	25.1	-	**15.1**	-	39.4	-	45.0	-	48.4
RTFNet(ResNet50)[34]	185.2	245.7	91.3	86.3	78.2	67.8	71.5	58.2	59.8	43.7	32.1	24.3	13.4	3.6	40.4	26.0	73.5	57.2	62.2	51.7
ABMDRNet(ResNet50)[52]	-	-	94.3	84.8	**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
GMNet(ResNet50)[61]	153.0	149.8	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
FuseSeg(DenseNet161)[35]	-	-	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	49.7	70.6	54.5
FEANet(ResNet152)[5]	255.2	337.1	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**	55.3	**77.3**	48.9	73.2	55.3
RTFNet(ResNet152)[34]	245.5	337.1	91.3	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
EGFNet(ResNet152)[7]	62.8	201.3	**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	48.3	71.1	47.1	72.7	54.8
MFTNet(ResNet152)[57]	330.6	360.9	95.1	**87.9**	85.2	66.8	**83.9**	64.4	64.3	47.1	50.8	36.1	45.9	8.4	62.8	**55.5**	73.8	**62.2**	74.7	57.3
Ours(ResNet50)	77.9	109.1	93.9	86.8	84.6	71.8	80.4	62.0	66.8	**49.7**	43.5	29.7	58.5	7.1	61.8	50.9	70.9	46.7	73.2	55.9
Ours(ResNet152)	147.3	200.4	95.4	87.6	85.2	72.6	79.9	63.8	70.6	48.6	47.9	35.0	62.8	14.2	62.7	52.4	71.9	58.3	**75.1**	**58.9**

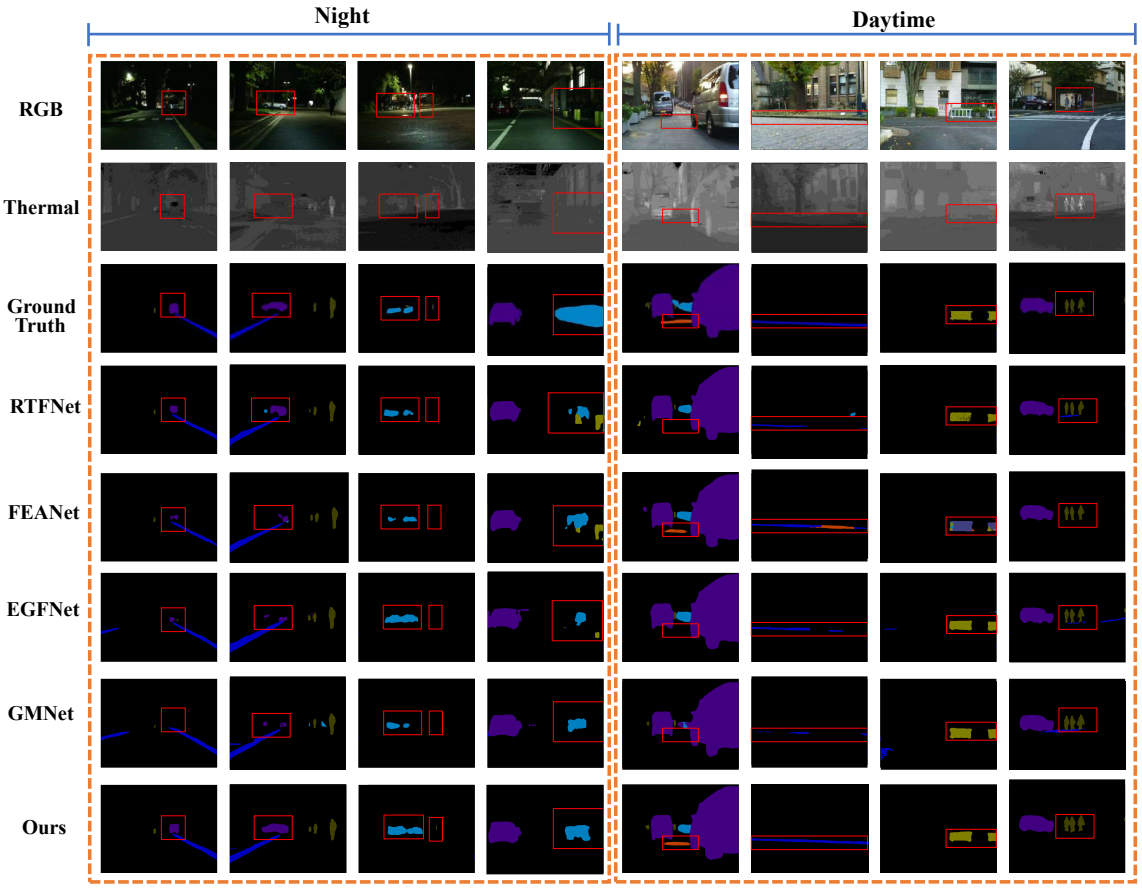

Fig. 4. Qualitative comparisons on the MFNet dataset. We can see that our method can provide better results in various lighting conditions and environments. The comparison results demonstrate our superiority.

C. Salient Objection Detection

1) Dataset: We evaluate our method on VT821 [37], VT1000 [41], and VT5000 [39], consisting of 821, 1000, and 5000 registered pairs of RGB-thermal images, respectively. Specifically, we utilize 2500 pairs from VT5000 for training and the remaining images, including those from the VT821 and VT1000 datasets, for testing.

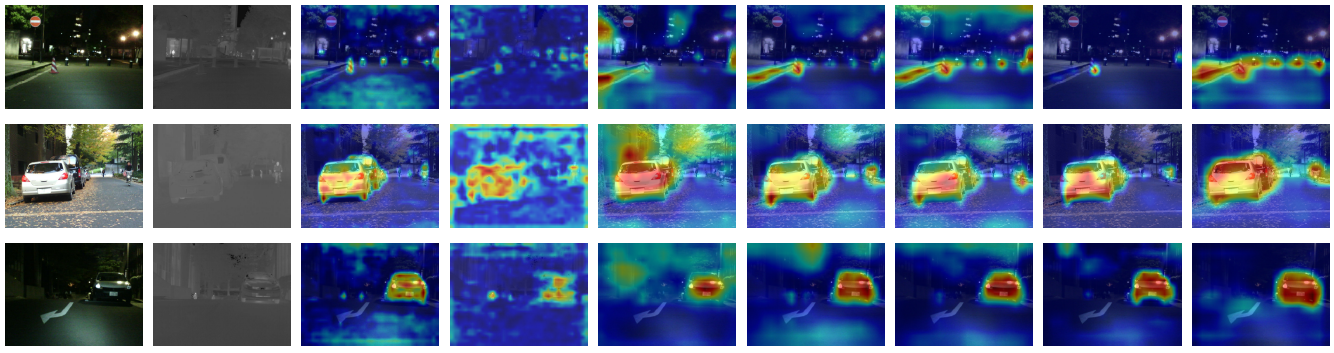

Fig. 5. Visualization of attention maps of different methods on the MFNet dataset.

TABLE II
RESULTS ON THE PST900 DATASET.

Method	mACC	mIoU
Efficient FCN[19]	66.75	57.27
CCNet[17]	73.43	66.00
ACNet[16]	78.67	71.81
SA-Gate[2]	84.71	79.05
RTFNet[34]	65.69	60.46
PSTNet[31]	/	68.36
GMNet[61]	89.61	84.12
Ours	**91.42**	**85.56**

TABLE III
QUANTITATIVE RESULTS ON THE M3FD OBJECT DETECTION DATASET.

Method	Day	Overcast	Night	Challenge	mAP@0.5
RGB	0.759	0.729	0.863	0.815	0.772
Thermal	0.717	0.727	0.852	**0.991**	0.753
Yolov5[30]	0.748	0.732	0.873	0.867	0.763
U2F[46]	0.738	0.731	0.868	0.976	0.775
TarDAL[20]	0.745	0.741	0.893	0.983	0.778
Ours	**0.783**	**0.786**	**0.895**	0.979	**0.801**

2) Implementation Details and Evaluation Metric: We resize the image resolution to 224 × 224 and perform some data augmentations as LSNET [65], including the random flips, random rotations, and clipping. The framework is the same as that used in semantic segmentation tasks built on ResNet50. We evaluate the performance using four metrics (S, adpE, adpF, MAE). The training hyperparameters, such as epoch, batch size, optimizer, and initial learning rate, are set to 20, 8, Adam optimization, and 0.001, respectively.

3) Results: We provide a quantitative comparison against previous approaches, including MTMR [37], M3S-NIR [40], SGDL [41], FMCF [51], ADF [42], MIDD [38], ECFFNet [58], and LSNet [65], CMDBIF-Net [45]. As seen in Table IV, our method achieves the best performance on all metrics, e.g., outperforming the state-of-the-art CMDBIF-Net by 3.1% in MAE on VT5000.

D. Crowd counting

1) Dataset: RGBT-CC dataset [21] has 2,030 RGB-T pairs captured in public scenarios. The images have a resolution of 640 × 480. A total of 138,389 pedestrians are marked with point annotations, and approximately 68 people are marked per image. The training, validation, and test set have 1545, 300, and 1200 RGB-T pairs, respectively.

2) Implementation Details and Evaluation Metric: Following IADM [21], [53], we use the BL [26] network built on the VGG16 [32] as the backbone. We send the feature maps generated by the last EAEF module into an MLP decoder comprising two 1×1 convolutions to get the final prediction. The training strategy is also consistent with IADM [21]. We use the Adam optimizer and set the learning rate to 0.00001. We evaluate the model every 10 epochs out of 300 epochs. The best model on the validation set will be used for evaluation. We measure with the root mean square error (RMSE) and the grid average mean absolute error (GAME) [8].

3) Results: The results are given in Table V where * denotes using pretrained VGG16 on the ImageNet. As shown, our method outperforms all baseline methods. Our method without and with pretraining outperforms TAFNet by 2.7% and DEFNet by 8% in RMSE, respectively.

E. Ablation Study

We analyze the effectiveness of each component of our EAEF through additional experiments on the MFNet dataset. We establish a baseline by removing the AIB and ACB from the EAEF. The results are shown in Table VI. We can observe that both AIB and ACB improved the performance of the baseline, and their combination, i.e., EAEF, gained the best performance.

V. CONCLUSIONS

In this paper, we studied the better fusion strategy of RGB images and thermal data for perception tasks. We explicitly specify cases where i) both RGB and thermal data, ii) only one type of data, and iii) none of them can provide sufficiently useful features. We proposed the explicit attention-enhanced fusion (EAEF) that enhances feature extraction and provides compensation for insufficient representations. We evaluated our method on different perception tasks, including semantic segmentation, object detection, salient detection, and crowd counting. As a result, we achieved state-of-the-art performance

TABLE IV
RESULTS OF SALIENT OBJECT DETECTION ON THE VT821, VT1000, AND VT5000 DATASETS.

Method	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$\downarrow$
MTMR[37]	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[40]	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[41]	0.765	0.847	0.731	0.085	0.787	0.856	0.764	0.090	0.750	0.824	0.672	0.089
FMCF[51]	0.760	0.796	0.640	0.080	0.873	0.899	0.823	0.037	0.814	0.864	0.734	0.055
ADF[39]	0.810	0.842	0.717	0.077	0.910	0.921	0.847	0.034	0.864	0.891	0.778	0.048
MIDD[38]	0.871	0.895	0.803	0.045	0.915	0.933	0.880	0.027	0.868	0.896	0.799	0.043
ECFFNet[58]	0.877	0.835	**0.911**	0.034	0.924	0.919	**0.959**	0.021	0.876	0.850	**0.922**	0.037
LSNet[65]	0.877	0.911	0.827	0.033	0.924	0.936	0.887	0.022	0.876	0.916	0.827	0.036
CMDBIF-Net[45]	0.882	0.923	0.855	0.032	**0.927**	0.952	0.914	0.019	0.886	0.933	0.868	0.032
Ours	**0.885**	**0.927**	0.846	**0.031**	0.926	**0.964**	0.905	**0.017**	**0.885**	**0.934**	0.853	**0.031**

TABLE V
RESULTS ON THE RGBT-CC DATASET. * DENOTES PERTAINING ON THE IMAGENET.

Method	GAME(0)$\downarrow$	GAME(1)$\downarrow$	GAME(2)$\downarrow$	GAME(3)$\downarrow$	RMSE$\downarrow$
UcNet[49]	33.96	42.42	53.06	65.07	56.31
HDFNet[29]	22.36	27.79	33.68	42.48	33.93
BBSNet[6]	19.56	25.07	31.25	39.24	32.48
MVMS[50]	19.97	25.1	31.02	38.91	33.97
IADM[21]	15.61	19.95	24.69	32.89	28.18
CSCA[53]	14.32	18.91	23.81	32.47	26.01
TAFNet[36]	12.38	16.98	21.86	30.19	22.45
Ours	12.58	17.24	22.33	30.88	21.85
DEFNet*[62]	11.90	16.08	20.19	27.27	21.09
Ours*	**11.19**	**14.99**	**19.20**	**27.13**	**19.39**

TABLE VI
RESULTS OF ABLATION STUDY ON THE MFNET DATASET.

Baseline	AIB	ACB	mAcc	mIoU
√			71.7	56.5
√	√		72.5	57.1
√		√	74.3	57.7
√	√	√	**75.1**	**58.9**

on all tasks, providing the robot community with a better fusion approach for RGB-thermal based perception tasks.

REFERENCES

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[2] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *European Conference on Computer Vision*, 2020, pp. 561–577.

[3] M. Contributors, "MMYOLO: OpenMMLab YOLO series toolbox and benchmark," https://github.com/open-mmlab/mmyolo, 2022.

[4] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[5] 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)*, 2021, pp. 4467–4473.

[6] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network," in *European conference on computer vision*, 2020, pp. 275–292.

[7] R. Fan, Z. Wang, and Q. Zhu, "Egfnet: Efficient guided feature fusion network for skin cancer lesion segmentation," in *2022 the 6th International Conference on Innovation in Artificial Intelligence (ICIAI)*, 2022, pp. 95–99.

[8] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, and D. Onoro-Rubio, "Extremely overlapping vehicle counting," in *Iberian Conference on Pattern Recognition and Image Analysis*, 2015, pp. 423–431.

[9] 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)*, 2017, pp. 5108–5115.

[10] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian conference on computer vision*, 2017, pp. 213–228.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[13] J. Hu, C. Bao, M. Ozay, C. Fan, Q. Gao, H. Liu, and T. L. Lam, "Deep depth completion from extremely sparse data: A survey," *arXiv preprint arXiv:2205.05335*, 2022.

[14] J. Hu, X. Guo, J. Chen, G. Liang, F. Deng, and T. L. Lam, "A two-stage unsupervised approach for low light image enhancement," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8363–8370, 2021.

[15] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 1043–1051.

[16] 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)*, 2019, pp. 1440–1444.

[17] 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 Conference on Computer Vision and Pattern Recognition*, 2019, pp. 603–612.

[18] Y.-H. Kim, U. Shin, J. Park, and I. S. Kweon, "Ms-uda: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6497–6504, 2021.

[19] J. Liu, J. He, J. Zhang, J. S. Ren, and H. Li, "Efficientfcn: Holistically-guided decoding for semantic segmentation," in *European Conference on Computer Vision*, 2020, pp. 1–17.

[20] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 5802–5811.

[21] 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 Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 4823–4833.

[22] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 756–13 765.

[23] X. Liu, M. Suganuma, Z. Sun, and T. Okatani, "Dual residual networks leveraging the potential of paired operations for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7007–7016.

[24] Liu, Jinyuan and Fan, Xin and Huang, Zhanbo and Wu, Guanyao and Liu, Risheng and Zhong, Wei and Luo, Zhongxuan, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5802–5811.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

[26] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 6141–6150.

[27] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[28] G. Montavon, G. Orr, and K.-R. Müller, *Neural networks: tricks of the trade*, 2012, vol. 7700.

[29] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for rgb-d salient object detection," in *European Conference on Computer Vision*, 2020, pp. 235–252.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[31] 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)*, 2020, pp. 9441–9447.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[33] M. Suganuma, X. Liu, and T. Okatani, "Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9039–9048.

[34] 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.

[35] 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.

[36] H. Tang, Y. Wang, and L.-P. Chau, "Tafnet: A three-stream adaptive fusion network for rgb-t crowd counting," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2022, pp. 3299–3303.

[37] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "Rgbt salient object detection: Benchmark and a novel cooperative ranking approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4421–4433, 2020.

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

[39] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgbt salient object detection: A large-scale dataset and benchmark," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[40] 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 *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 141–146.

[41] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "Rgb-t image saliency detection via collaborative graph learning," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 160–173, 2020.

[42] Tu, Zhengzheng and Ma, Yan and Li, Zhun and Li, Chenglong and Xu, Jieming and Liu, Yongtao, "Rgbt salient object detection: A large-scale dataset and benchmark," *IEEE Transactions on Multimedia*, pp. 1–1, 2022.

[43] J. Vertens, J. Zürn, and W. Burgard, "Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8461–8468.

[44] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "Cgfnet: Cross-guided fusion network for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2949–2961, 2022.

[45] Z. Xie, F. Shao, G. Chen, H. Chen, Q. Jiang, X. Meng, and Y.-S. Ho, "Cross-modality double bidirectional interaction and fusion network for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[46] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2022.

[47] K. Yi and J. Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7017–7025.

[48] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 516–520.

[49] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8582–8591.

[50] Q. Zhang and A. B. Chan, "Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8297–8306.

[51] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "Rgb-t salient object detection via fusing multi-level cnn features," *IEEE Transactions on Image Processing*, vol. 29, pp. 3321–3335, 2020.

[52] 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 Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 2633–2642.

[53] Y. Zhang, S. Choi, and S. Hong, "Spatio-channel attention blocks for cross-modal crowd counting," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2022, pp. 90–107.

[54] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[55] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," *arXiv preprint arXiv:2211.14461*, 2022.

[56] H. Zhou, L. Qi, Z. Wan, H. Huang, and X. Yang, "Rgb-d co-attention network for semantic segmentation," in *Proceedings of the Asian conference on computer vision*, 2020.

[57] H. Zhou, C. Tian, Z. Zhang, Q. Huo, Y. Xie, and Z. Li, "Multi-spectral fusion transformer network for rgb-thermal urban scene semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, 2022.

[58] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1224–1235, 2022.

[59] W. Zhou, C. Liu, J. Lei, and L. Yu, "Rllnet: a lightweight remaking learning network for saliency redetection on rgb-d images," *Science China Information Sciences*, vol. 65, no. 6, p. 160107, 2022.

[60] W. Zhou, C. Liu, J. Lei, L. Yu, and T. Luo, "Hfnet: Hierarchical feedback network with multilevel atrous spatial pyramid pooling for rgb-d saliency detection," *Neurocomputing*, vol. 490, pp. 347–357, 2022.

[61] 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.

[62] W. Zhou, Y. Pan, J. Lei, L. Ye, and L. Yu, "Defnet: Dual-branch enhanced feature fusion network for rgb-t crowd counting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24 540–24 549, 2022.

[63] W. Zhou, E. Yang, J. Lei, and L. Yu, "Frnet: Feature reconstruction network for rgb-d indoor scene parsing," vol. 16, no. 4, 2022, pp. 677–687.

[64] W. Zhou, Y. Yue, M. Fang, X. Qian, R. Yang, and L. Yu, "Bcinet: Bilateral cross-modal interaction network for indoor scene understanding in rgb-d images," *Information Fusion*, vol. 94, pp. 32–42, 2023.

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