Label-guided Real-time Fusion Network for RGB-T Semantic Segmentation

Zengrong Lin[1]
linzr9@mail2.sysu.edu.cn
Baihong Lin[2]
linbaihong111@126.com
Yulan Guo[1]
guoyulan@sysu.edu.cn

[1] School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China

[2] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China

Abstract

RGB-T semantic segmentation has emerged as a promising solution to handle hard scenes with poor lighting conditions by fusing a pair of RGB and thermal images. Although various deep-learning-based fusion networks have been proposed with excellent performance, most of them are not suitable for real-time applications due to high computational overhead and latency. To realize high-accuracy RGB-T real-time semantic segmentation, this paper proposes a novel *Label-guided Real-time Fusion Network* which fuses multi-level features of RGB and thermal images extracted from double two-pathway lightweight backbones based on the proposed *Label-guided Fusion Module* (LFM). The proposed LFM realizes efficient multi-modal feature fusion by spatial weighted summation, in which a spatial attention map is generated with the guidance of semantic label in the training phase to accurately indicate the contributions of different modalities. Exhaustive experiments on the MFNet and PST900 datasets demonstrate that the proposed method simultaneously achieves higher speed and accuracy compared with other state-of-the-art methods.

1 Introduction

RGB-T semantic segmentation tries to accurately classify each pixel of a RGB image into a specific label by using a thermal image as complementary data. This technique has emerged as a promising solution to address the limitation of RGB-based semantic segmentation mainly caused by poor lighting conditions. So far, various methods have been proposed for RGB-T semantic segmentation, among which deep learning methods have drawn widespread attention with high segmentation accuracy[22]. However, most deep learning methods suffer from low computational efficiency, i.e., they are not suitable for RGB-T real-time semantic segmentation which is often necessary for many applications[21].

*Corresponding author: Baihong Lin. The work is supported in part by Natural Science Foundation of Guangdong Province (2022A1515010493), in part by National Natural Science Foundation of China (No. U20A20185, 61972435, 62372491), and in part by the Guangdong Basic and Applied Basic Research Foundation (2022B1515020103).

To realize high-accuracy RGB-T real-time semantic segmentation, there exist two challenging problems: First, high-accuracy designs usually rely on heavy backbones such as ResNet-50, but these heavy backbones usually bring in high computational cost. To solve this problem, lightweight networks are proposed with a good balance between accuracy and efficiency for segmentation in recent years, *e.g.*, BiSeNetV1 [17], BiSeNetV2 [18]. These lightweight networks have already adopted for RGB-based real-time semantic segmentation, but they are seldom discussed for RGB-T real-time semantic segmentation.

Second, conventional fusion modules based on element-wise addition or concatenation [4, 10] are simple for implementation, but they fail to fully integrate information of paired RGB and thermal images [22] due to their neglect of modality differences caused by different imaging mechanisms. To solve this problem, some researchers [1, 19] introduce attention mechanisms to reduce modality differences laid in features before or after conventional fusion module, while others [7, 21, 22] explore the characteristics of cross-modal features at different levels and devise different fusion modules to merge multi-modal features at different level respectively under multi-label supervision. However, most researches on fusion module designs [1, 7, 19, 21, 22] are conducted based on heavy backbones without consideration of computational complexity. Their performance are not validated based on real-time lightweight backbones which have poorer representation abilities compared with heavy networks. Therefore, it is necessary to discuss fusion module designs based on lightweight networks for RGB-T real-time semantic segmentation.

To address the above problems, this paper proposes a novel *Label-guided Real-time Fusion Network* (LRFNet). Specifically, the proposed LRFNet first introduces double two-pathway lightweight backbones proposed by BiSeNetV1 [17] to extract features of different levels (i.e., detail and context) from RGB and thermal images, and then fuses modality features at different levels based on the proposed core component named *Label-guided Fusion Module* (LFM) respectively to achieve fast and accurate perception. In the proposed LFM, to realize efficient multi-modal feature fusion, we adopt spatial weighted summation based on a spatial attention map which is generated with the guidance of semantic label in the training phase to indicate the contributions of different modalities. Experimental results on MFNet and PST900 datasets show that the proposed LRFNet outperforms other state-of-the-art methods in speed and accuracy.

2 Related work

Real-time Semantic Segmentation. Semantic segmentation can easily achieve high accuracy using methods based on fully convolutional networks [12]. Unfortunately, most semantic segmentation methods cannot achieve real-time performance due to their high computational overhead and latency. To balance the trade-off between accuracy and efficiency, researchers try to devise shallow lightweight networks[2, 18]. However, most lightweight networks do not have sufficient representation ability for accurate segmentation. To overcome this problem, recent studies focus on two mainstreams. The first mainstream is to fuse multi-layer features of a lightweight backbone. For example, MSFNet [12] devises a Multi-features Fusion Module for backbone (ResNet-18) to enlarge the receptive field and recover the spatial information loss. STDC-Seg [2] introduces Feature Fusion Module [17] to fuse features from low-level and high-level layers in a designed lightweight backbone named STDC. The second mainstream is to devise multi-branch architecture. For example, ICNet [20] devises a multi-branch cascade network which can extract features of multi-

resolution images to efficiently achieve high segmentation accuracy. BiSeNetV1 [17] and BiSeNetV2 [18] devises two-stream architectures to extract both detail and context information of an image, and fuse them to enhance segmentation performance. The above network designs have good speed-accuracy trade-offs. However, these methods cannot be directly applied for RGB-T semantic segmentation, since RGB and thermal images cannot be simply regarded as a 4-channel image considering their modality difference [1, 4].

RGB-T semantic Segmentation. Compared with RGB-based semantic segmentation, The main challenging problem of RGB-T semantic segmentation is how to efficiently fuse different modality features of RGB and thermal images. Early studies [4, 10, 13, 14] utilize simple feature fusion methods: Specifically, MFNet [4] employs element-wise addition to fuse different level modality features from two separate symmetric encoders in a shared decoder. PSTNet [10] employs concatenation to fuses RGB and thermal information in a devised dual-stream architecture for real-time segmentation. Similar to MFNet, RTFNet [13] and FuseSeg [14] employ element-wise addition to fuse modality features, but they design network architectures based on heavy backbones like ResNet-152s [5] and DenseNet-161s [6] to further enhance feature representation ability. The above simple fusion methods achieve good segmentation accuracy. However, their performance are gradually stuck in a bottleneck, since simple fusion based on element-wise addition or concatenation does not fully consider modality differences caused by different imaging mechanisms.

To overcome the performance bottleneck, recent studies try to devise more intricate fusion methods. Specifically, some researchers introduce attention mechanisms to refine features before or after conventional fusion module so that modality differences laid in features can be effectively reduced. For example, ABMDRNet [19] devises a Channel Weighted Fusion module based on channel attention. FEANet [1] devises a Feature-Enhanced Attention module before addition-based simple feature fusion. Other researchers devise different fusion modules to merge multi-modal features at different level respectively under multi-label supervision so that modality differences at different level can be fully considered. For example, EGFNet [22] devises a Multimodal Fusion module, and adopts boundary and semantic maps to guide shallow and deep feature fusion respectively based on multitask deep supervision. GMNet [21] devises a Shallow Feature Fusion module and a Deep Feature Fusion module to fully integrate cross-modal information, and uses semantic, binary, and boundary labels to further guide fusion results at different levels. The above mentioned fusion modules show excellent performance. However, they are designed based on heavy backbones and are not validated based on real-time lightweight networks.

3 Proposed Method

In this section, we first introduce our *Label-guided Fusion Module* (LFM) in detail. Then, based on LFM, we introduce our proposed *Label-guided Real-time Fusion Network* (LRFNet). The overall architecture is shown in Fig.1.

3.1 Label-guided Fusion Module

Previous study has pointed out that fusion model can easily achieve sub-optimal performance if always giving equal importance for features from different modality [3]. However, how to efficiently generate a confidence map to indicate the contribution of different modality features for fusion is an intractable problem. Although recent work [3] has provided an

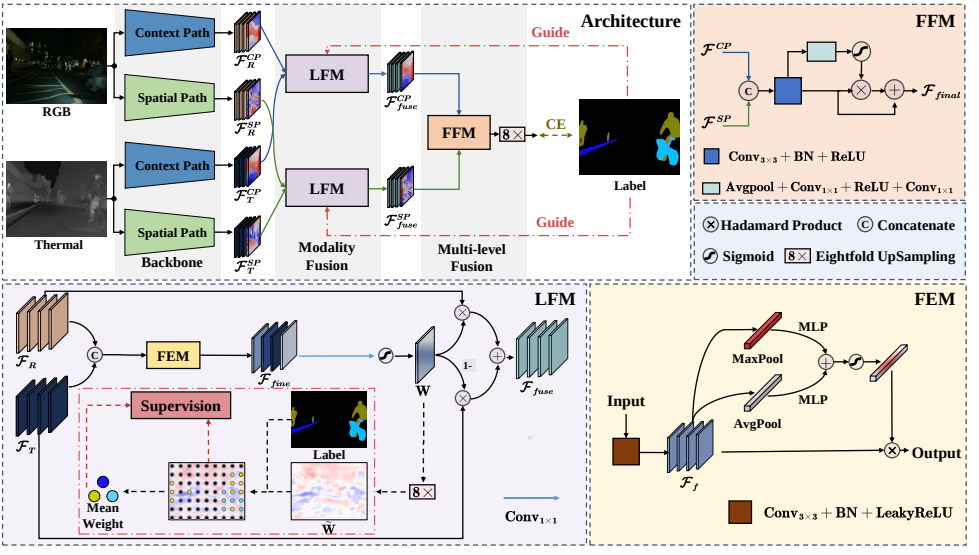

Figure 1: The overall architecture of *Label-guided Real-time Fusion Network*, which adopts *spatial path* and *context path* in BiSeNetV1 [17] as backbones.

effective solution, it is designed based on heavy backbone and has high computational complexity, which is not suitable for lightweight networks with poorer representation ability. To address the above problem, we propose a novel and simple *Label-guided Fusion Module* (LFM) which consists of three parts: *Feature Enhancement Module*, *Spatial Weighted Features Fusion*, and *Semantic Label Supervision*.

Feature Enhancement Module. The *Feature Enhancement Module* (FEM) is designed to refine each coarse feature from lightweight backbones before further processing. Specifically, we first concatenate RGB and thermal features, and use a convolutional layer to achieve a deeper feature following [4]:

$$\mathcal{F}_f = \sigma_L\left(\mathrm{BN}\left(\mathrm{Conv}_{3\times3}\left(\mathrm{Cat}\left(\mathcal{F}_R, \mathcal{F}_T\right)\right)\right)\right) \tag{1}$$

where $\mathrm{Conv}_{3\times3}(*)$ denotes a 3×3 convolutional block, BN denotes a Batch Normalization layer, and σ_L denotes LeakyReLU layer with rate 0.2. Then, we introduce the channel attention module proposed by [15] to further refine the feature with little computational cost, and obtain the fine-grained feature $\mathcal{F}_{fine}$ as the following:

$$\mathcal{F}_{fine} = \sigma_S\left(\mathrm{MLP}\left(\mathrm{GMP}\left(\mathcal{F}_f\right)\right) + \mathrm{MLP}\left(\mathrm{GAP}\left(\mathcal{F}_f\right)\right)\right) \times \mathcal{F}_f \tag{2}$$

where GMP and GAP denote global average and max pooling respectively, two MLPs represent two multi-layer perceptrons with shared parameters, and σ_S denotes Sigmoid function.

Spatial Weighted Features Fusion. Using the fine-grained feature $\mathcal{F}_{fine}$, we generate a spatial attention map $\mathbf{W}$ to indicate the contribution of different modality features at different position as follows:

$$\mathbf{W} = \sigma_S\left(\mathrm{Conv}_{1\times1}\left(\mathcal{F}_{fine}\right)\right) \tag{3}$$

where $\mathrm{Conv}_{1\times1}(*)$ denotes two 1×1 convolutional blocks. Then, based on the spatial attention map $\mathbf{W}$, we obtain the fusion feature $\mathcal{F}_{fuse}$ by fusing the RGB feature $\mathcal{F}_R$ and the

thermal feature $\mathcal{F}_T$ as the following:

$$\mathcal{F}_{fuse} = \mathbf{W} \times \mathcal{F}_R + (1 - \mathbf{W}) \times \mathcal{F}_T \tag{4}$$

However, since the spatial attention map $\mathbf{W}$ usually contains too much noise due to the poor representation ability of lightweight backbones, the above fusion process will easily result in sub-optimal performance. To suppress noise and generate an accurate spatial attention map, we introduce semantic label supervision on the spatial attention map during training process.

Semantic Label Supervision. Noting that modality features within the same category make similar contributions to fusion in a scene, we propose a customized semantic label supervision method which explicitly clusters the intra-category values of a spatial attention map to suppress noise without sacrificing inference speed. Specifically, given a spatial attention map $\mathbf{W} \in \mathbb{R}^{h \times w}$ and the corresponding semantic label map $\mathbf{Y}$ which has k categories, we first upsample the attention map $\mathbf{W}$ to align the corresponding semantic label map, and obtain the mean weight of each category $\mathbf{m} \in \mathbb{R}^{k \times 1}$ as the following:

$$\mathbf{m}_i = \frac{1}{|\mathcal{C}_i|} \sum_{j \in \mathcal{C}_i} \widetilde{\mathbf{W}}_j, \quad \widetilde{\mathbf{W}} = upsample(\mathbf{W}), \quad i = 0, 1..., k-1 \tag{5}$$

where $\mathcal{C}_i$ denotes a set of pixel positions with the i-th category in $\mathbf{Y}$, $|\mathcal{C}_i|$ denotes the number of elements in $\mathcal{C}_i$. Then, we design a regularization term $\mathcal{L}$ for $\widetilde{\mathbf{W}}$ as the following:

$$\mathcal{L}(\widetilde{\mathbf{W}}, \mathbf{Y}) = \frac{1}{k} \sum_{i=0}^{k-1} \frac{1}{|\mathcal{C}_i|} \sum_{j \in \mathcal{C}_i} \left(\widetilde{\mathbf{W}}_j - \mathbf{m}_i \right)^2 \tag{6}$$

By minimizing the regularization term $\mathcal{L}(*,*)$, we suppress noise laid in a spatial attention map and enhance fusion performance.

3.2 Network Architecture

Based on the proposed *Label-guided Fusion Module* (LFM), we design a *Label-guided Real-time Fusion Network* as shown in Fig.1. First, we introduce two Spatial Paths and two Context Paths proposed by BiSeNet [17] to extract both detailed and contextual features $\left\{\mathcal{F}_R^{SP}, \mathcal{F}_R^{CP}, \mathcal{F}_T^{SP}, \mathcal{F}_T^{CP}\right\}$ from RGB and thermal images respectively. The context path is a pretrained ResNet-18 network [5] whereas the spatial path is composed of three convolution layers. Then, we propose a novel fusion module called *Label-guided Fusion Module* (LFM) to integrate features from the RGB and thermal features in both detail and context paths separately, obtaining detailed and contextual fusion features $\left\{\mathcal{F}_{fuse}^{SP}, \mathcal{F}_{fuse}^{CP}\right\}$. Finally, we introduce a *Feature Fusion module* (FFM) proposed by [17] to further fuse detailed and contextual fusion features, obtaining the segmentation fusion feature $\mathcal{F}_{final}$ and upsampling $\mathcal{F}_{final}$ to achieve the segmentation prediction.

Considering the class-imbalance in the datasets [4, 10], we adopt weighted cross entropy loss function [9] to supervise the output of the whole network, and design the loss function as the following:

$$\mathcal{L}_{total} = \frac{1}{N} \sum_{b=1}^{N} \left[\mathrm{WCE}\left(\hat{\mathbf{Y}}_b, \mathbf{Y}_b\right) + \eta \left(\mathcal{L}_{CP}(\widetilde{\mathbf{W}}_{CP}^b, \mathbf{Y}_b) + \mathcal{L}_{SP}(\widetilde{\mathbf{W}}_{SP}^b, \mathbf{Y}_b) \right) \right] \tag{7}$$

where N denotes the number of samples, $\mathrm{WCE}(*,*)$ denotes the average weighted cross-entropy loss between ground-truth label $\mathbf{Y}$ and segmentation prediction $\hat{\mathbf{Y}}$, $\mathcal{L}_{CP}(*,*)$ and $\mathcal{L}_{SP}(*,*)$ denotes two regularization terms based on Eqn.(6) for two spatial attention maps, $\widetilde{\mathbf{W}}^b_{CP}$ and $\widetilde{\mathbf{W}}^b_{SP}$, in spatial path and context path respectively. η is a hyper-parameter to balance the weight between the weighted cross entropy loss and the regularization terms. In our paper, η is set to 0.1.

4 Experiment

In this section, we conduct all experiments based on two public RGB-T semantic segmentation datasets: MFNet [4] and PST900 [10]. We first introduce two datasets and implementation details. Then, we use ablation studies to analyze the proposed LFM and the designed loss function. Finally, we compare the proposed method with other state-of-the-art methods in accuracy and computational cost.

4.1 Dataset

MFNet. The MFNet dataset is a challenging dataset which collects urban street scenes at different time. It contains 1569 pairs of RGB and thermal images with the same spatial size of 480$\times$ 640, which can be split into 820 pairs taken at daytime and 749 pairs taken at nighttime. The dataset has 8 hand-labeled object classes and background. For fairness, we split the dataset into training, testing, and validation sets following the scheme in [4].

PST900. The PST900 dataset comes from the DARPA Subterranean Challenge. It has 894 pairs of aligned RGB and thermal nighttime image with 4 hand-labeled object classes and background. For fairness, we split the dataset into training and testing sets following the scheme in [10], and utilize the approach in [22] to resize each image to 640$\times$ 1280 pixels.

4.2 Implementation Details

Experiment Setup. We utilize Ranger optimizer [16] and Cosine Annealing [8] to gradually reduce the learning rate from 7e-5 to 0 within 300 epochs as in [22]. Note that our training process involves two steps: Firstly, we train a model without semantic label supervision to determine the mean weights $\{\mathbf{m}_i\}$ of Eqn.(5). Then, we retrain a new model using these mean weights under the proposed semantic label supervision without considering the background class, and introduce OHEM [11] to improve the training process. For data augmentation, we apply random flipping and cropping as described in previous literature [4, 22]. The batch size is set to 3 and the weight decay is set to 5e-4. We train and evaluate the performance and the inference speed on a single Tesla V100S.

Evaluation Metrics. We evaluate segmentation performance using mean accuracy (mAcc) and mean intersection over union (mIoU). Additionally, we adopt the number of parameters (params.), floating point operations (FLOPs), and frames per second (FPS) to evaluate the computational cost of all comparison methods comprehensively.

4.3 Ablation Study

In this section, we conduct a series of ablation experiments on the MFNet dataset to validate the effectiveness of the proposed *Label-guided Fusion Module* (LFM), *Feature Enhance-*

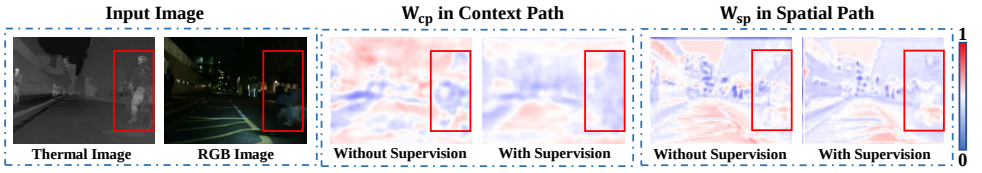

Figure 2: Visualization examples of the spatial attention maps W_{CP} and W_{SP} in Context Path and Spatial Path with our proposed Semantic Label Supervision.

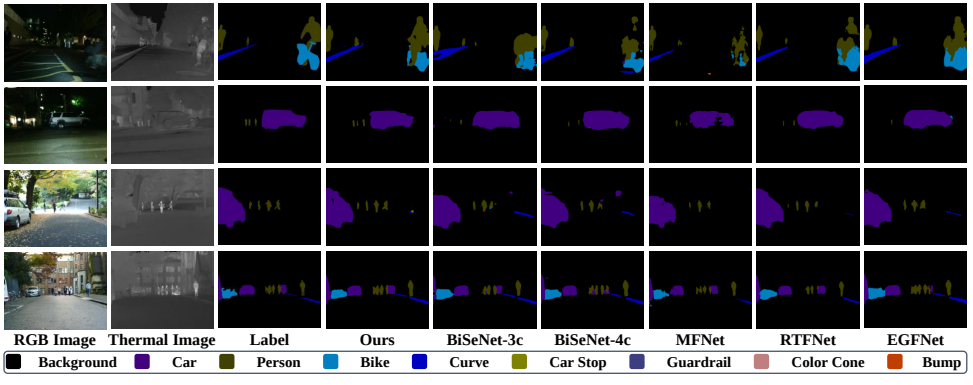

Figure 3: Visualization segmentation examples of our method and five representative state-of-the-art methods on the test set of MFNet.

ment Module (FEM) and loss function. The results are shown in Table.1, in which "w/o" represents removing a certain component.

Effectiveness of LFM. To validate the effectiveness of the proposed LFM, we compare LFM with simple fusion methods in the "Simple Fusion Methods" and "Placement of LFM" parts in Table.1. In the "Simple Fusion Methods" part of Table.1; "4 Channels" denotes the result of BiSeNet with input of a 4 channels RGB-T image; "Concatenation" presents replacing LFM with concatenation in LRFNet; "Addition" presents replacing LFM with addition in LRFNet. From the results of the Simple Fusion Methods part in Table.1, we can find that simple fusion methods based on concatenation or addition outperform the method of "4 Channel" based on BiSeNet, which implies that directly regarding RGB-T images as a 4 channels image will result in sub-optimal performance.

In the "Placement of LFM" part of Table.1, "before FFM" represents fusing features of different modalities on Spatial Path and Context Path respectively using two LFMs before FFM; "after FFM" represents fusing features of different modalities using a singe FFM after FFM. From the results, we can find that adopting two LFMs to fuse the detail and context features respectively before FFM can produce better fusion features compared with the fusion results after FFM. The main reason is that detail/context information of RGB and thermal images usually makes different contributions to fusion whereas a single LFM can not handle the mixture of detail and context information accurately.

Effectiveness of FEM. To validate the effectiveness of the proposed FEM in LFM, we conduct ablation experiments on FEM and show the results in the "Structure of FEM" part of Table.1. From the results, we can find that FEM can improve the accuracy performance. Obviously, FEM can effectively extract the representative information from the coarse modality

	Simple Fusion Methods (w/o LFM)			Placement of LFM	
Model	4 Channels	Concatenation	Addition	before FFM	after FFM
mAcc	64.7	65.7	65.5	68.0	64.5
mIoU	52.4	53.8	53.7	55.1	53.6
	Loss Function			**Structure of FEM**	
Model	CE	WCE	WCE+$\mathcal{L}(*,*)$	with FEM	w/o FEM
mAcc	61.2	66.7	68.0	68.0	67.1
mIoU	53.6	54.3	55.1	55.1	54.2

Table 1: Results from ablation experiments on the MFNet dataset.

Methods	Type	Publication	Backbone	Params. (M)	FLOPs (G)	FPS	mAcc	mIoU
BiSeNet-3c	RGB	ECCV 2018	ResNet18	13.3	17.4	241.7	61.4	48.2
BiSeNet-4c	RGB	ECCV 2018	ResNet18	13.3	17.9	237.3	64.7	52.4
MFNet	RGBT	IROS 2017	No	0.7	8.4	178.1	45.1	39.7
RTFNet-152	RGBT	RAL 2019	ResNet152	254.5	290.3	16.4	63.1	53.2
FuseSeg	RGBT	T-ASE 2021	DenseNet161	100.1	141.0	20.5	70.6	54.5
ABMDRNet	RGBT	CVPR 2021	ResNet50	64.6	194.3	23.1	69.5	54.8
EGFNet	RGBT	AAAI 2022	ResNet101	62.8	201.3	20.5	**72.7**	54.8
Ours	RGBT	-	ResNet18	**25.9**	**32.0**	**111.3**	68.0	**55.1**

Table 2: Comparison results from the MFNet dataset.

features and produce the accurate spatial attention map **W** to indicate the contribution of different modality features at different positions.

Effectiveness of Loss Function. To validate the effectiveness of loss function, we compare different loss functions in the "Loss Function" part of Table.1. From the results, we can find that the weighted cross-entropy loss function (WCE) outperforms the cross-entropy loss function, since WCE can alleviate the influence of class-imbalance in a dataset [9]. More importantly, the regularization term $\mathcal{L}(*,*)$ of semantic label supervision enhances the performance effectively. To further validate the regularization term $\mathcal{L}(*,*)$ of semantic label supervision, we visualize the spatial maps of the Context Path and Spatial Path in Fig.2. Fig.2 shows that the proposed semantic label supervision can effectively suppress the noise in the spatial attention map, leading to intra-category consistent fusion.

4.4 Comparison with State-of-the-Arts

Comparison on MFNet dataset. For the MFNet dataset, we compare our proposed LRFNet with six state-of-the-art semantic segmentation methods, including a real-time RGB-based segmentation method (BiSeNet [17]) and five RGB-T segmentation methods (MFNet [4], RTFNet [13], FuseSeg [14], ABMDRNet [19], EGFNet [22]). The results are shown in Table.2 and Fig.3. From the results, we can find that our method requires approximately 60% fewer parameters and 85% fewer FLOPs than those of ABMDRNet or EGFNet, but it achieves comparable segmentation accuracy under mIoU with a high frame rate of 111.3 FPS for paired 480×640 RGB and thermal images. Although our method does not have high values of mAcc, recent work [3] has pointed out that mIoU is a better indicator than mAcc for the extremely-class-unbalanced MFNet dataset. Overall, our proposed LRFNet reaches a good balance between accuracy and computational cost.

Methods	Type	Params. (M)	FLOPs (G)	FPS	Daytime		Nighttime	
					mAcc	mIoU	mAcc	mIoU
BiSeNet-3c	RGB	13.3	17.4	241.7	57.1	43.7	56.9	45.8
BiSeNet-4c	RGB	13.3	17.9	237.3	56.2	44.9	61.7	51.4
MFNet	RGBT	0.7	8.4	178.1	42.6	36.1	41.4	36.8
RTFNet-152	RGBT	254.5	290.3	16.4	60.0	45.8	60.7	54.8
FuseSeg	RGBT	100.1	141.0	20.5	62.1	47.8	67.3	54.6
ABMDRNet	RGBT	64.6	194.3	23.1	58.4	46.7	68.3	55.5
EGFNet	RGBT	62.8	201.3	20.5	74.4	47.3	68.0	55.0
Ours	RGBT	**25.9**	**32.0**	**111.3**	60.4	**47.9**	65.1	54.3

Table 3: Results from daytime and nighttime images on the MFNet dataset.

Methods	Type	Publication	Backbone	Params. (M)	FLOPs (G)	FPS	mAcc	mIoU
MFNet	RGBT	IROS 2017	No	0.7	21.4	94.6	63.5	50.3
RTFNet-152	RGBT	RAL 2019	ResNet152	254.5	773.3	7.6	65.7	60.5
PSTNet	RGBT	ICRA 2020	No	-	-	-	-	68.4
FuseSeg	RGBT	T-ASE 2021	DenseNet161	-	-	-	-	-
ABMDRNet	RGBT	CVPR 2021	ResNet50	64.6	808.1	7.1	79.1	71.3
EGFNet	RGBT	AAAI 2022	ResNet101	62.8	536.2	9.9	**94.0**	**78.5**
Ours	RGBT	-	ResNet18	**25.9**	**85.2**	**67.3**	86.1	**78.4**

Table 4: Results from PST900 dataset [10]. '-' represents no public data.

We also summarize segmentation evaluations on the daytime and nighttime test sets of the MFNet dataset in Table 2. Notably, our method performs relatively worse during nighttime. The main reason is that the spatial attention map in Spatial Path tends to indicate higher contributions of RGB image feature due to the rich detail information laid in RGB image (see Fig.2). However, RGB image does not contain much useful semantic information during nighttime, thus, our proposed LRFNet performs worse relatively in this case. In the future, we will improve the network design to solve this problem.

Comparison on PST900 dataset. We additionally evaluate our proposed LRFNet on the PST900 dataset. The comparison results are shown in Table 4. It can be found that our proposed method still achieves competitive performance with a high frame rate of 67.3 FPS for paired 640×1280 RGB and thermal images compared with other state-of-the-art methods.

5 Conclusion

This paper proposes a novel *Label-guided Real-time Fusion Network* which fuses detail and context features of RGB and thermal images extracted from double two-pathway lightweight backbones respectively based on the proposed *Label-guided Fusion Module* (LFM) to achieve fast and accurate perception. The proposed LFM conducts weighted feature fusion based on a spatial attention map generated with the guidance of semantic label in the training phase to accurately indicate the contribution of different modalities. Specifically, it achieves 55.1% mIoU with the speed of 111.3FPS on the MFNet dataset, and 78.4% mIoU with the speed of 67.3FPS on the PST900 dataset, proving that the proposed method reaches a better balance between accuracy and computational cost compared with other state-of-the-art methods.

References

[1] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *IEEE In. Conf. on Intell. Robots and Systems (IROS)*, pages 4467–4473. IEEE, 2021.

[2] Mingyuan Fan, Shenqi Lai, Junshi Huang, Xiaoming Wei, Zhenhua Chai, Junfeng Luo, and Xiaolin Wei. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9725, 2021.

[3] Oriel Frigo, Lucien Martin-Gaffe, and Catherine Wacongne. Doodlenet: Double deeplab enhanced feature fusion for thermal-color semantic segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3020–3028, 2022. doi: 10.1109/CVPRW56347.2022.00341.

[4] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IEEE In. Conf. on Intell. Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 770–778, 2016.

[6] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4700–4708, 2017.

[7] Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE Trans. Circuits and Syst. Video Technol.*, 2022.

[8] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

[9] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

[10] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *proc. IEEE int. conf. on robotics and automation (ICRA)*, pages 9441–9447. IEEE, 2020.

[11] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[12] Haiyang Si, Zhiqiang Zhang, and Feng Lu. Real-time semantic segmentation via multiple spatial fusion network.

[13] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3): 2576–2583, 2019.

[14] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion. *IEEE Trans. Automation Science and Engineering*, 18(3):1000–1011, 2020.

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

[16] Less Wright and Nestor Demeure. Ranger21: a synergistic deep learning optimizer. *arXiv preprint arXiv:2106.13731*, 2021.

[17] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[18] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129:3051–3068, 2021.

[19] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2633–2642, 2021.

[20] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 405–420, 2018.

[21] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Trans. Image Processing*, 30:7790–7802, 2021.

[22] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb–thermal scene parsing. In *proc. AAAI Conf. Artif. Intell.*, volume 36, pages 3571–3579, 2022.