Supplementary Materials
ABMDRNet: Adaptive-weighted Bi-directional Modality Difference Reduction Network for RGB-T Semantic Segmentation

1. Experiments

1.1. MDRF

1.1.1 One-directional or Bi-directional for Reducing Modality Difference

In this section, we compare the results of reducing modality differences by using one-directional with that of using bi-directional strategies. Here, one-directional strategy denotes our modality difference reduction strategy is only performed on one of the single-modality RGB and thermal features. While, bi-directional strategy denotes our modality difference reduction strategy is performed on both of the single-modality RGB and thermal features simultaneously. As shown in Table 1, 'BS+T'/'BS+RGB' denotes that the one-directional modality difference reduction strategy is performed on thermal images or RGB images. While, 'BS+B' denotes that the bi-directional strategy is employed.

As shown in Table 1, compared with 'BS', our proposed one-directional strategy (*i.e*., 'BS+T' and 'BS+RGB') can improve the performance of RGB-T semantic segmentation to some extent. Furthermore, the bi-directional strategy (*i.e*., 'BS+B') can further boost the performance of RGB-T semantic segmentation. This indicates that the cross-modality complementary information can be effectively exploited by employing modality difference reduction.

1.1.2 Visualization

In Fig. 1, we provide more visual results to further demonstrate the effectiveness of our proposed MDRF subnetwork. It can be observed that, compared with those fused features obtained by the simple fusion strategy without using MDRF (*i.e*., w/o MDRF in Fig. 1), the features obtained by our MDRF are more discriminative. This may owe to the employed bi-directional bridging strategy in the MDRF.

Metrics	BS	BS+T	BS+RGB	BS+B
mAcc	57.30	57.60	59.32	62.37
mIoU	47.99	49.90	51.28	51.98

Table 1. The quantitative comparative results (%) of using one-directional or bi-directional strategy to reduce modality differences.

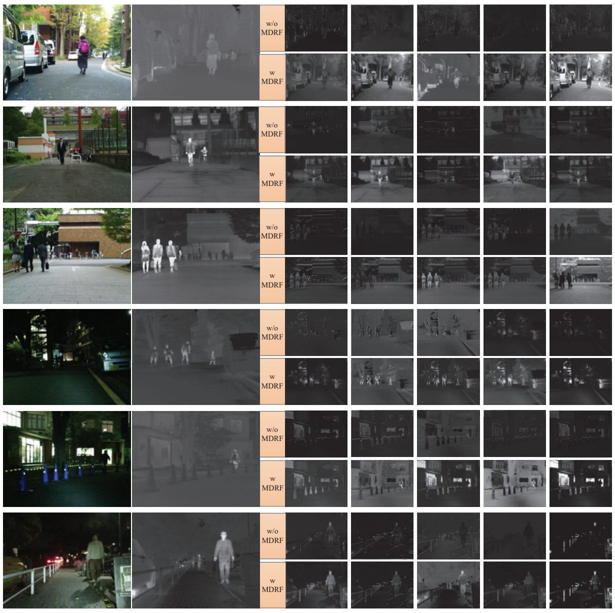

Figure 1. Visual results of the fused cross-modality feature maps with or without using MDRF.

1.2. MSC

In Fig. 2, we provide some fused cross-modality feature maps with or without the MSC module. By comparing with the fused cross-modality feature maps before and after employing MSC, we can observe that the features for both small and large targets (*e.g*., the regions marked by red dotted boxes in Fig. 2) can be better distinguished from those of confusing background areas by employing MSC module. This benefits from the exploitation of effective semantic correlations and long-range relationships between arbitrary two positions in the multi-scale cross-modality feature maps. Thank to the interactions among multi-scale contextual information of cross-modality features together with their long-range dependencies along spatial dimension obtained by our proposed MSC module, the issue of ob-

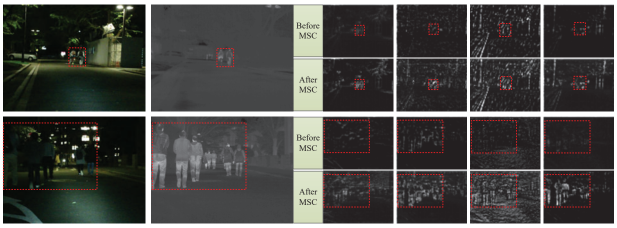

Figure 2. Visual results of cross-modality feature maps before and after employing MSC.

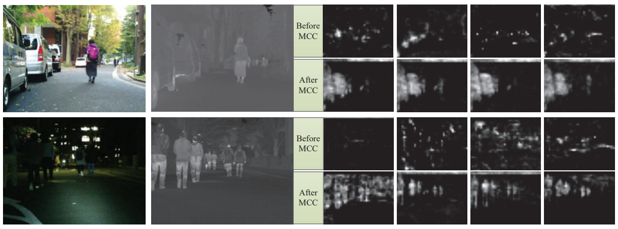

Figure 3. Visual results of cross-modality feature maps before and after employing MCC.

jects diversity in RGB-T semantic segmentation can be addressed to a large extent, thus boosting the RGB-T semantic segmentation performance.

1.3. MCC

Similarly, we also demonstrate the effectiveness of our proposed MCC module by visualizing some fused cross-modality feature maps before and after using MCC module. As shown in Fig. 3, we can observe that the discriminative target features in different channels are all significantly activated by employing the MCC module, which benefits from the exploitation of class-correlations and long-range relationships between arbitrary two channels in the multi-scale cross-modality feature maps. Owing to the proposed MCC module that introduces the multi-scale contextual information and their long-range dependencies along the channel dimension, the RGB-T semantic segmentation performance is greatly boosted.

1.4. The Daytime and Nighttime Results

We also compare our model with other SOTA methods on the daytime and nighttime test sets of MFNet dataset, respectively. As shown in Table 2, our model achieves promising results in both of the two scenarios. Especially, in the nighttime, our model significantly outperforms other models. This indicates that, compared with other models, our model can better exploit the cross-modality complementary information for RGB-T semantic segmentation.

1.5. The Inference Speed

We measure the inference speed of our proposed ABMDRNet and other SOTA models on an NVIDIA GeForce GTX 1080Ti GPU. As shown in Table 3, our model achieves competitive inference speeds, *i.e.*, 40.14 FPS, with other models.

Methods	Daytime		Nighttime	
	mAcc	mIoU	mAcc	mIoU
DUC	53.8	43.6	57.9	50.1
DANet	50.9	37.5	52.4	40.1
HRNet	54.4	46.1	55.1	50.7
LDFNet	55.2	35.9	61.3	40.7
ACNet	60.7	41.6	63.9	47.4
SA-Gate	49.3	37.9	56.9	45.6
MFNet	42.6	36.1	41.4	36.8
RTFNet	57.3	44.4	59.4	52.0
Ours	58.4	46.7	68.3	55.5

Table 2. Quantitative results of different models (%) on the daytime and nighttime test sets of MFNet dataset. The best three results are highlighted in red, green and blue.

Methods	ms	FPS
DUC	15.92	62.83
DANet	15.67	63.82
HRNet	84.25	11.87
LDFNet	13.41	74.58
ACNet	27.12	36.87
SA-Gate	18.39	54.36
MFNet	4.35	229.86
RTFNet-50	11.25	88.87
Ours	24.91	40.14

Table 3. The inference speed of different models on one GTX 1080Ti GPU.