ABMDRNet

Zhang, Qiang; Zhao, Shenlu; Luo, Yongjiang; Zhang, Dingwen; Huang, Nianchang; Han, Jungong

Published in:
IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Publication date:
2021

Citation for published version (APA):
Zhang, Q., Zhao, S., Luo, Y., Zhang, D., Huang, N., & Han, J. (Accepted/In press). ABMDRNet: Adaptive-weighted Bi-directional Modality Difference Reduction Network for RGB-T Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

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

Qiang Zhang[1] Shenlu Zhao[1] Yongjiang Luo[2] Dingwen Zhang[1] Nianchang Huang[1*] Jungong Han[3*]
[1]School of Mechano-Electronic Engineering, Xidian University, China
[2] School of Electronic Engineering, Xidian University, China
[3]Computer Science Department, Aberystwyth University, U.K.
nchuang@stu.xidian.edu.cn, jungonghan77@gmail.com

Abstract

Semantic segmentation models gain robustness against poor lighting conditions by virtue of complementary information from visible (RGB) and thermal images. Despite its importance, most existing RGB-T semantic segmentation models perform primitive fusion strategies, such as concatenation, element-wise summation and weighted summation, to fuse features from different modalities. These strategies, unfortunately, overlook the modality differences due to different imaging mechanisms, so that they suffer from the reduced discriminability of the fused features. To address such an issue, we propose, for the first time, the strategy of ***bridging-then-fusing****, where the innovation lies in a novel Adaptive-weighted Bi-directional Modality Difference Reduction Network (ABMDRNet). Concretely, a Modality Difference Reduction and Fusion (MDRF) subnetwork is designed, which first employs a bi-directional image-to-image translation based method to reduce the modality differences between RGB features and thermal features, and then adaptively selects those discriminative multi-modality features for RGB-T semantic segmentation in a channel-wise weighted fusion way. Furthermore, considering the importance of contextual information in semantic segmentation, a Multi-Scale Spatial Context (MSC) module and a Multi-Scale Channel Context (MCC) module are proposed to exploit the interactions among multi-scale contextual information of cross-modality features together with their long-range dependencies along spatial and channel dimensions, respectively. Comprehensive experiments on MFNet dataset demonstrate that our method achieves new state-of-the-art results.*

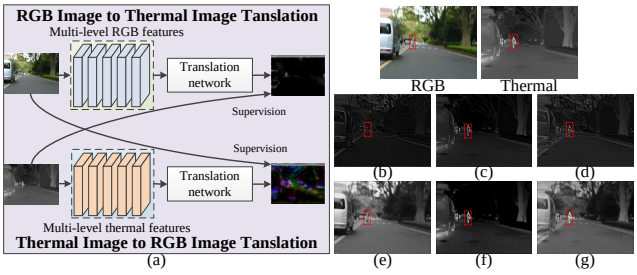


Figure 1. Illustration of modality difference reduction. (a) Bi-directional modality difference reduction. (b)-(d) Original RGB features, thermal features and their fused features, respectively. (e)-(g) RGB features, thermal features and their fused features after reducing modality differences, respectively.

1. Introduction

Semantic segmentation aims to assign category labels to each pixel in a natural image, which plays an important role in many computer vision task, such as autonomous driving [6, 29], pedestrian detection [1], pathological analysis [19, 24] and so on.

So far, CNN-based RGB semantic segmentation methods [13, 14, 19, 25] have achieved prominent results in many large-scale datasets [5, 15]. However, their performance may significantly degrade under poor lighting conditions. To boost semantic segmentation performance, recent researches pay more attention to RGB-T semantic segmentation [8, 20, 23], where thermal images may complement rich contour information and semantic information to RGB images under poor lighting conditions.

Existing models for multi-modality pixel-level prediction tasks, including RGB-T semantic segmentation and RGB-T salient object detection, usually adopt simple strategies, such as element-wise summation [23], concatenation [8] and weighted summation [7, 30], to capture the complementary information from paired RGB and thermal images. However, they usually ignore the modality differences between RGB images and thermal images, which are caused

*Equally corresponding authors.

by different imaging mechanisms. Such negligence may lead to inadequate cross-modality complementary information exploitation. As shown in Fig. 1, the people region, marked by red dotted box in Fig. 1(b), has low intensity values, while the same region in Fig. 1(c) has higher intensity values. If simple fusion operations are employed, the discriminative target information in the thermal image will be noticeably suppressed in the fused features, as shown in Fig. 1(d).

To solve this problem, we propose a novel multi-modality feature fusion subnetwork, *i.e.*, Modality Difference Reduction and Fusion (MDRF), to better exploit the multi-modality complementary information from RGB images and thermal images via a novel strategy of *bridging-then-fusing*. In the bridging stage, as shown in Fig. 1(a), a bi-directional image-to-image translation [12, 31] based method is employed to reduce the differences between RGB and thermal features. The basic idea is that when transferring images from one modality to another, some non-discriminative single-modality information, caused by different imaging mechanisms (*e.g.*, Fig. 1(b) and Fig. 1(c)), will be translated into discriminative ones (*e.g.*, Fig. 1(e) and Fig. 1(f)) by virtue of the complementary supervision information from the images of another modality. As a result, the modality differences between the extracted single-modality RGB and thermal features will be reduced for better fusion (*e.g.*, Fig. 1(d) and Fig. 1(g)). Then, in the fusing stage, a novel fusion module, *i.e.*, Channel Weighted Fusion (CWF) module, is presented to capture the cross-modality information between the corresponding channels of single-modality RGB and thermal features, whose modality differences have been reduced in the first step. As shown in Fig. 1(d) and Fig. 1(g), higher discriminative fused features may be obtained by using the single-modality features that have reduced modality differences than those original ones.

Furthermore, the diversity of objects, *e.g.*, categories, sizes and shapes, in a given image is also problematic for semantic segmentation. Multi-scale contextual information and their long-range dependencies have been proved to be effective to address such an issue in RGB semantic segmentation. However, in multi-modality semantic segmentation, especially for RGB-T semantic segmentation [8, 20, 23], multi-scale contextual information of cross-modality features and their long-range dependencies are not in place yet. In RGB-T semantic segmentation, only MFNet [8] added several mini-inception blocks in the encoder to obtain some contextual information. But this is far limited for semantic segmentation.

Inspired by [3, 6, 28], we propose two novel modules, *i.e.*, a Multi-Scale Spatial Context (MSC) module and a Multi-Scale Channel Context (MCC) module, to exploit the multi-scale contextual information of cross-modality features and their long-range dependencies along spatial and channel dimensions, respectively. First, multi-scale features are obtained by performing the Atrous Spatial Pyramid Pooling (ASPP) module [3] on the original fused cross-modality features. Then the long-range dependencies for these multi-scale features along the spatial and channel dimensions are established by jointly using the original fused cross-modality features and their corresponding multi-scale features in MSC and MCC, respectively. With MSC and MCC cooperative, the multi-scale contextual information of cross-modality features and their long-range dependencies will be fully exploited for RGB-T semantic segmentation.

The main contributions of this paper are summarized as follows:

(1) An end-to-end ABMDRNet is presented to facilitate RGB-T semantic segmentation by simultaneously considering multi-modality difference reduction and multi-scale contextual information of cross-modality data. Comprehensive experimental results show that our model achieves new state-of-the-art performance on MFNet dataset.

(2) An MDRF subnetwork is presented to effectively capture the cross-modality information from the RGB and thermal images via a strategy of *bridging-then-fusing*, which first employs a bi-directional image-to-image translation based method to bridge the modality gaps between multi-modality data and then adaptively selects those discriminative multi-modality features for RGB-T semantic segmentation.

(3) An MSC module and an MCC module are presented to fully exploit the multi-scale contextual information of cross-modality features and their long-range dependencies along the spatial and channel dimensions, respectively.

2. Related work

2.1. RGB-based semantic segmentation

Early RGB-based semantic segmentation methods mainly rely on low-level hand-crafted features combined with flat classifiers, such as Random Forests [21] and multiclass fuzzy Support Vector Machine [18]. Recently, deep learning based semantic segmentation models [2,6,14,17,19,25] have become the mainstream and achieved significant improvements. These models are usually based on Fully Convolutional Network (FCN) [14] for its simple but reasonable architecture for pixel-wise prediction. As well, to address the diversity of objects, these FCN-based models mainly exploit some pyramid structures, such as Pyramid Pooling Module (PPM) [32] and Atrous Spatial Pyramid Pooling (ASPP) [3], to capture the discriminative multi-scale contextual information from input images. Although these multi-scale contextual information extraction modules have achieved great successes in semantic segmentation, their receptive fields are still limited, thus failing to exploit the global contextual information. Recently, many mod-

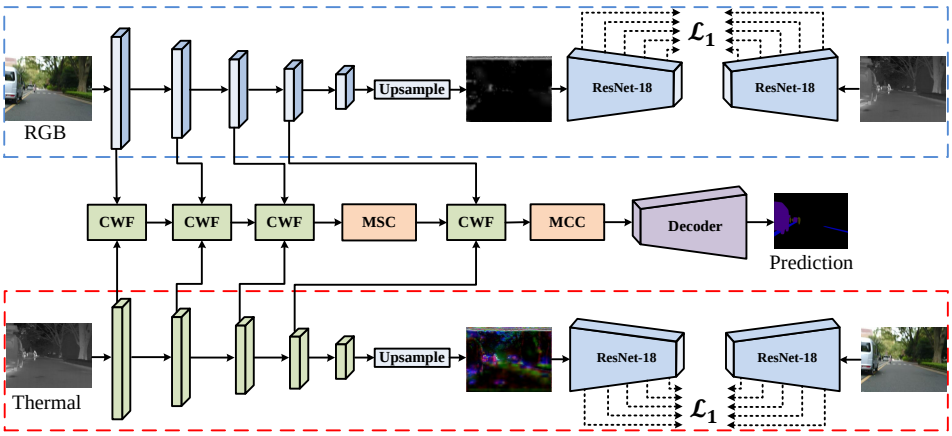


Figure 2. Overall framework of our proposed model. The blue and red dotted boxes represent the bi-directional modality difference reduction stage in MDRF subnetwork.

els [16, 28] try to exploit the long-range dependencies to address such issue and have achieved promising results. For example, non-local operation [28] was proposed to compute the response at a position as a weighted sum of the features for capturing long-range dependencies.

2.2. Multi-modality semantic segmentation

Recently, with the rapid development of imaging techniques, many studies [4, 8, 11, 20, 23, 27] employ multi-modality data (*e.g.*, RGB-T images and RGB-D images) to address some issues arising from the traditional RGB semantic segmentation. These multi-modality semantic segmentation models are usually divided into two categories, *i.e.*, feature-level fusion based and image-level fusion based ones. Specifically, feature-level fusion based models first extract single-modality features from each input modality data and then fuse them to capture complementary information for semantic segmentation. For example, [10] proposed an Attention Complementary Module (ACM) to capture more high-quality single-modality RGB features and depth features from different channels for boosting the RGB-D semantic segmentation. [8] adopted an Encoder-Decoder architecture, which first extracted RGB features and thermal features respectively and then fused them by the tailored short-cut blocks. [23] first fused the multi-level RGB features and thermal features by element-wise summation and then employed an upception block to improve the decoding results. Different from them, image-level fusion based models directly take the combination of multi-modality images as inputs. For example, [20] proposed a sequential dual-stream CNN architecture, which concatenated an RGB image, the matched thermal image and the coarse mask predicted by RGB features as a five-channel input to predict the result.

In contrast to RGB-D semantic segmentation, RGB-T semantic segmentation attracts less attention. Most existing RGB-T semantic segmentation models [8, 20, 23] employ simple fusion strategies, such as element-wise summation [23] and concatenation [8, 20], to capture the cross-modality features, while ignoring the modality differences caused by different imaging mechanisms. Alternatively, in this paper, a novel strategy of bridging-then-fusing is presented to capture the cross-modality features, where the modality differences between multi-modality data are first reduced and then the discriminative multi-modality features are adaptively selected for RGB-T semantic segmentation.

3. Method

As shown in Fig. 2, the proposed RGB-T semantic segmentation framework, *i.e.*, ABMDRNet, consists of three components, including MDRF subnetwork, MSC module and MCC module. The details of them will be discussed in the following contents.

3.1. MDRF

Although paired RGB images and thermal images can provide much complementary information to each other, the modality differences, caused by different imaging mechanisms, may hinder the integration and exploitation of multi-modality complementary information from RGB images and thermal images. Unfortunately, this has been ignored by most existing models. To address this issue, we design a novel multi-modality feature fusion subnetwork, *i.e.*, MDRF subnetwork, via a strategy of bridging-then-fusing, which first reduces the modality differences bi-directionally and then exploits the multi-modality complementary information. Specifically, the MDRF subnetwork consists of two stages. The first stage is bi-directional modality difference reduction, which aims to obtain discriminative single-modality features with fewer modality differences from RGB images and thermal images, respectively. The second stage is discriminative single-modality features fusion,

which aims to effectively exploit the complementary information from multi-modality features.

3.1.1 Bi-directional modality difference reduction

Inspired by those image-to-image translation methods [12, 31], we employ a bi-directional bridging strategy to reduce the modality differences caused by different imaging mechanisms. The strategy starts with a bi-directional difference reduction, including the reduction from RGB to thermal and that from thermal to RGB. Specifically, the multi-level single-modality features extracted from one modality are employed to generate a matched pseudo image of another modality. Meanwhile, its corresponding real image of another modality is also available in RGB-T semantic segmentation. Considering that, we reduce the differences from one modality to another modality by enforcing the features from the pseudo image and those from the real image of the same modality to be similar as possible.

As shown in Fig. 2, we exactly employ the same modality difference reduction structure for RGB images and thermal images (*i.e.*, region marked by blue dotted box and red dotted box). Therefore, in the following contents, we will take the procedure of transferring RGB images into thermal images as an example to show the details of our structure for modality difference reduction.

First, a ResNet-50 [9] is employed to extract the single-modality features from an RGB image. The average pooling and the fully connected layers of the ResNet-50 are removed to maintain more spatial information. Therefore, five levels of single-modality RGB features $\{\mathbf{F}_n^{RGB}|n{=}1,2,3,4,5\}$ are obtained, which have the resolutions of 1/2, 1/4, 1/8, 1/16 and 1/32 of the original image sizes, respectively. Then, the last four levels of single-modality RGB features are fed into an RGB-to-T translation network to generate the corresponding pseudo thermal image. The translation network first performs four 1×1 convolutional layers on the four levels of single-modality RGB features $\{\mathbf{F}_n^{RGB}|n{=}2,3,4,5\}$ to generate one-channel feature maps. Then, all of the generated feature maps are upsampled and fused to generate a pseudo thermal image ($\mathbf{I}^{pse-T}$). After that, to make the generated pseudo thermal image similar to its corresponding real thermal image and further reduce the modality differences, two auxiliary ResNet-18s [9] are employed to extract five levels of auxiliary features (*i.e.*, $\{\mathbf{F}_n^{pse-T}|n{=}1,2,3,4,5\}$ and $\{\mathbf{F}_n^{real-T}|n{=}1,2,3,4,5\}$) from the pseudo thermal image and its real thermal image, respectively. The average pooling and the fully connected layers of the two ResNet-18s are also removed to maintain the spatial information. By enforcing the two sets of features to be similar as possible, the extracted single-modality features $\{\mathbf{F}_n^{RGB}|n{=}1,2,3,4\}$ from RGB modality may share some similar properties with those features $\{\mathbf{F}_n^{T}|n{=}1,2,3,4\}$ from thermal modality, thus reducing the modality differences between the two modality data.

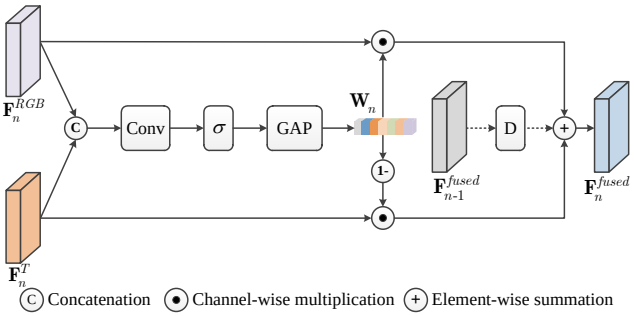


Figure 3. Structure of the proposed CWF module. The weight vector $\mathbf{W}_n$ is able to weigh the importance of feature maps from RGB modality.

For a thermal image, we can also obtain its pseudo RGB image $\mathbf{I}^{pse-RGB}$ together with its corresponding pseudo and real RGB image features $\{\mathbf{F}_n^{pse-RGB}|n{=}1,2,3,4,5\}$ and $\{\mathbf{F}_n^{real-RGB}|n{=}1,2,3,4,5\}$. Furthermore, the extracted single-modality features $\{\mathbf{F}_n^{T}|n{=}1,2,3,4\}$ from the thermal image may also be enforced to share some similar properties with those from the RGB image by using the same way, which will further reduce the modality differences between the two modality data.

In this stage, the following loss is employed for supervision, *i.e.*,

$$\mathcal{L}_{MD}{=}\sum_{n=1}^{5}\mathcal{L}_1\left(\mathbf{F}_n^{pse-T},\mathbf{F}_n^{real-T}\right)+\sum_{n=1}^{5}\mathcal{L}_1\left(\mathbf{F}_n^{pse-RGB},\mathbf{F}_n^{real-RGB}\right), \tag{1}$$

where $\mathcal{L}_1(*)$ denotes L1 loss. Because of the bi-directional modality difference reduction, there will be smaller differences between the RGB features $\{\mathbf{F}_n^{RGB}|n{=}1,2,3,4\}$ and the thermal features $\{\mathbf{F}_n^{T}|n{=}1,2,3,4\}$ extracted from the ResNet-50s. This will improve the discriminability of the fused cross-modality features, as discussed in the earlier section.

3.1.2 Channel-Wise Weighted Features Fusion

Having the single-modality features, the next step is to capture their complementary information by using some fusion strategies for RGB-T semantic segmentation. The most intuitive ways are element-wise summation or concatenation, which cannot exploit the multi-modality complementary information effectively. For that, some complex strategies [7, 30] obtain the fused features in a weighted summation way. However, most existing fusion strategies adopt the same weights for all of the channels. These weights may work well for some channels of features, while some undesirable fusion results may be obtained for some channels of features. In fact, different channels of features may

correspond to different classes for semantic segmentation. Compared with features from different spatial positions, the features from different channels may have higher class-discriminability for semantic segmentation.

Considering that, we propose a novel CWF module in the fusing stage of MDRF to effectively exploit the cross-modality complementary information by re-weighting the importance of single-modality features in a channel-dependent way, rather than in a spatial position-dependent way. Specifically, given the n-th level of single-modality features (*i.e.*, $\{\mathbf{F}_n^{RGB}|n{=}1,2,3,4\}$ and $\{\mathbf{F}_n^{T}|n{=}1,2,3,4\}$) from the first stage, as shown in Fig. 3, the proposed CWF module exploits their multi-modality complementary information by using the following steps.

First, the $\mathbf{F}_n^{RGB}$ and $\mathbf{F}_n^{T}$ are concatenated and then fed into two convolutional layers to obtain the relative importance of the paired features from different modalities but in the same channels. The corresponding importance weight vector $\mathbf{W}_n$ is obtained by

$$\mathbf{W}_n = \mathrm{GAP}(\sigma(\mathrm{Conv}(\mathrm{Cat}(\mathbf{F}_n^{RGB}, \mathbf{F}_n^{T}); \beta))). \tag{2}$$

Here, $\mathrm{Conv}(*;\beta)$ denotes a convolutional block with a 1×1 convolutional layer and a 3×3 convolutional layer and β denotes its parameters. $\mathrm{GAP}(*)$ denotes the global average pooling operation. $\sigma(*)$ denotes the Sigmoid activation function, respectively. Higher values in $\mathbf{W}_n$ indicate that corresponding channels of features in RGB modality are more likely to be important than those corresponding channels of features from thermal images, and vice versa. As a result, the relative importance for each channel of features from different modalities are obtained. By using these channel importance weight vectors $\{\mathbf{W}_n|n{=}1,2,3,4\}$, the fused features $\{\mathbf{F}_n^{fused}|n{=}1,2,3,4\}$ are obtained by

$$\mathbf{F}_n^{fused} = \begin{cases} \mathbf{W}_n \odot \mathbf{F}_n^{RGB} + (\mathbf{1} - \mathbf{W}_n) \odot \mathbf{F}_n^{T}, & n = 1 \\ \mathrm{D}(\mathbf{F}_{n-1}^{fused}) + \mathbf{W}_n \odot \mathbf{F}_n^{RGB} + (1 - \mathbf{W}_n) \odot \mathbf{F}_n^{T}, & n = 2,3,4 \end{cases} \tag{3}$$

where $\odot$ denotes the channel-wise multiplication and $\mathbf{1}$ denotes a vector of 1's with the same size of $\mathbf{W}_n$. $\mathbf{F}_{n-1}^{fused}$ denotes the fused features from the previous level and $\mathrm{D}(*)$ is a convolutional block with stride of 2 for downsampling.

With several CWF modules, corresponding channels of single-modality features from different modalities are re-weighted and fused. Compared with those fusion strategies sharing the same weights for different channels, our proposed CWF module may better select those channels of features with high discriminability from multi-modality data for semantic segmentation.

3.2. MSC Module and MCC Module

As discussed in Section 1, multi-scale contextual information and long-range dependencies have been proved to be effective for alleviating the issue of objects diversity in RGB semantic segmentation, but they are still not well exploited in RGB-T semantic segmentation. For that, we propose an MSC module and an MCC module to simultaneously exploit the multi-scale contextual information of cross-modality features and their long-range dependencies along the spatial and channel dimensions, respectively. More specifically, the MSC module is performed on the 3-rd level of the fused features $\mathbf{F}_3^{fused}$ and the MCC module is performed on the 4-th level of the fused features $\mathbf{F}_4^{fused}$, considering the number of parameters in the entire model.

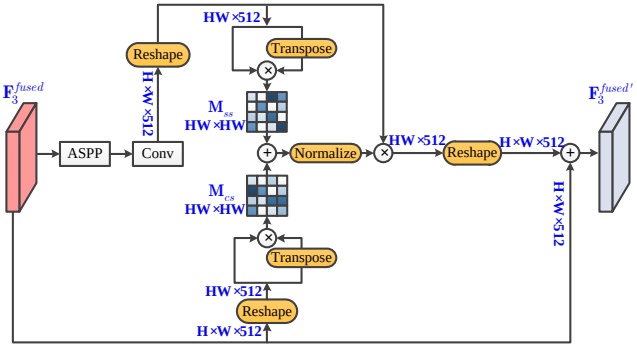


Figure 4. Structure of the proposed MSC module. The blue annotations denote the sizes of the input and output features of MSC.

3.2.1 MSC Module

The structure of MSC is shown in Fig. 4. Given the 3-rd level of the fused features $\mathbf{F}_3^{fused} \in R^{H\times W\times 512}$, MSC exploits their multi-scale contextual information of cross-modality features and their long-range dependencies along the spatial dimension by using the following steps.

First, an ASPP module [3] is employed to extract multi-scale contextual information from the input fused features. For that, the ASPP module employs four parallel convolutional branches to obtain four scales of features. In each branch, a 1×1 standard convolutional layer and a 3×3 atrous convolutional layer with different dilation rates (*i.e.*, 1, 6, 12 and 18, respectively) are employed. Then, the four scales of features are concatenated and fed into a 1×1 convolutional layer to reduce their channels, thus obtaining the final multi-scale features $\mathbf{F}_3^{ms} \in R^{H\times W\times 512}$. Subsequently, inspired by [6] and [16], a self-spatial correlation matrix $\mathbf{M}_{ss} \in R^{HW\times HW}$ is computed from the fused multi-scale features by

$$\mathbf{M}_{ss} = \mathrm{Reshape}\,(\mathbf{F}_3^{ms}) \times (\mathrm{Reshape}\,(\mathbf{F}_3^{ms}))^{\mathrm{T}}, \tag{4}$$

where $(*)^{\mathrm{T}}$ denotes matrix transpose and $\mathrm{Reshape}(*)$ transfers the size of the input matrix from $R^{H\times W\times C}$ to $R^{HW\times C}$. This self-spatial correlation matrix $\mathbf{M}_{ss}$ captures the paired-wise similarities of two arbitrary positions in the multi-scale features and can be employed to extract

the long-range spatial dependencies among the multi-scale contextual features.

Meanwhile, considering that the long-range dependencies among the multi-scale contextual features should consist with those of original input features, a cross-spatial correlation matrix $\mathbf{M}_{cs} \in R^{HW \times HW}$ is also computed from the original input features to complement the self-spatial correlation matrix $\mathbf{M}_{ss}$ for better capturing the long-range dependencies along the spatial dimension.

$$\mathbf{M}_{cs} = \text{Reshape}\left(\mathbf{F}_3^{fused}\right) \times \left(\text{Reshape}\left(\mathbf{F}_3^{fused}\right)\right)^{\text{T}}. \quad (5)$$

With the self-spatial and cross-spatial correlation matrices $\mathbf{M}_{ss}$ and $\mathbf{M}_{cs}$, the final spatial correlation matrix $\mathbf{M}_s \in R^{HW \times HW}$ is obtained by

$$\mathbf{M}_s = \text{Normalization}\left(\mathbf{M}_{ss} + \mathbf{M}_{cs}\right), \quad (6)$$

where $\text{Normalization}(*)$ denotes Min-Max Normalization.

After that, the multi-scale contextual information of the fused features and their corresponding long-range dependencies along the spatial dimension are obtained by

$$\mathbf{F}_3^{fused'} = \text{Reshape}'\left(\left(\mathbf{M}_s \times \text{Reshape}\left(\mathbf{F}_3^{ms}\right)\right)\right) + \mathbf{F}_3^{fused}, \quad (7)$$

where $\text{Reshape}'(*)$ denotes the inverse process of $\text{Reshape}(*)$.

3.2.2 MCC Module

Given the 4-th level of the fused features $\mathbf{F}_4^{fused} \in R^{M \times N \times 1024}$, MCC follows similar steps with MSC to exploit the multi-scale contextual information of cross-modality features and their long-range dependencies along the channel dimension. The differences between MSC and MCC are in the ways of computing correlation matrix. In MCC, a self-channel correlation matrix $\mathbf{M}_{sc} \in R^{1024 \times 1024}$ and a cross-channel correlation matrix $\mathbf{M}_{cc} \in R^{1024 \times 1024}$ are computed. Specifically, after obtaining the multi-scale features $\mathbf{F}_4^{ms} \in R^{M \times N \times 1024}$ from their input features $\mathbf{F}_4^{fused}$, $\mathbf{M}_{sc}$ and $\mathbf{M}_{cc}$ are computed by

$$\mathbf{M}_{sc} = \left(\text{Reshape}\left(\mathbf{F}_4^{ms}\right)\right)^T \times \text{Reshape}\left(\mathbf{F}_4^{ms}\right), \quad (8)$$

$$\mathbf{M}_{cc} = \left(\text{Reshape}\left(\mathbf{F}_4^{fused}\right)\right)^{\text{T}} \times \text{Reshape}\left(\mathbf{F}_4^{fused}\right). \quad (9)$$

Similar to that in MSC, the cross-channel correlation matrix $\mathbf{M}_{cc}$ in MCC is also used to complement the self-channel correlation matrix $\mathbf{M}_{sc}$ for better capturing the long-range dependencies along the channel dimension. Different from that in MSC, the channel correlation matrixes capture the paired-wise similarities of two arbitrary channels in the multi-scale features and can be employed to extract the long-range channel dependencies among the multi-scale contextual features. The final channel correlation matrix $\mathbf{M}_c \in R^{1024 \times 1024}$ is computed by

$$\mathbf{M}_c = \text{Normalization}\left(\mathbf{M}_{sc} + \mathbf{M}_{cc}\right). \quad (10)$$

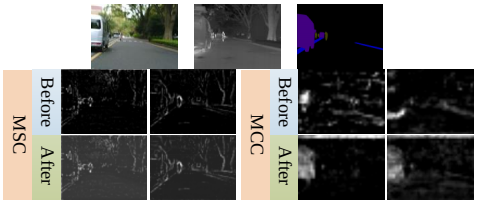


Figure 5. Visual results of multi-scale cross-modality feature maps before and after employing MSC and MCC modules.

After that, the multi-scale contextual information of the fused features and their corresponding long-range dependencies along the channel dimension are obtained by

$$\mathbf{F}_4^{fused'} = \text{Reshape}'\left(\left(\text{Reshape}\left(\mathbf{F}_4^{ms}\right) \times \mathbf{M}_c\right)\right) + \mathbf{F}_4^{fused}. \quad (11)$$

By using MSC and MCC, the multi-scale contextual information of the cross-modality features and their long-range dependencies along the spatial and channel dimensions are captured simultaneously. As shown in Fig. 5, the discriminability of the fused cross-modality features will be greatly boosted by introducing those contextual information.

3.3. Loss Function

The total loss function $\mathcal{L}_{total}$ for training our model consists of the semantic segmentation loss $\mathcal{L}_s$ and multi-modality differences loss $\mathcal{L}_{MD}$, *i.e.*,

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_s\left(\mathbf{S}, \mathbf{G}\right) + \lambda_2 \mathcal{L}_{MD}, \quad (12)$$

where λ_1 and λ_2 denote two hyper-parameters for balancing the two losses. They are empirically set to 1 and 5 in our experiments, respectively. $\mathbf{G}$ denotes the ground truth and $\mathbf{S}$ denotes the final prediction. Considering the imbalance of pixels of each class presented in the MFNet dataset, inspired by [17], we employ the weighted cross-entropy loss as the semantic segmentation loss $\mathcal{L}_s$, which is defined by

$$\mathcal{L}_s = -\sum_{i=1}^{m}\sum_{j=1}^{n} w(x_{ij}) \times p\left(x_{ij}\right) \times \log\left(q\left(x_{ij}\right)\right), \quad (13)$$

where m and n represent the width and height of an image. (i, j) represents the coordinate of a pixel. $w(x_{ij})$ represents the weight coefficient of the pixel class. $p(x_{ij})$ represents the ground truth label of the pixel and $q(x_{ij})$ represents the predicted result on the pixel.

4. Experiments

4.1. The Dataset and Evaluation Metrics

Our model is verified in MFNet dataset [8], which is the only public dataset of natural images for RGB-T semantic segmentation. This dataset contains 1569 annotated RGB and thermal natural image pairs, in which 820 image pairs

Methods	mAcc	mIoU
BS	57.30	47.99
+BMDR	62.37	51.98
+BMDR+IFCNN [49]	62.68	52.62
+BMDR+CW [50]	63.61	52.52
+BMDR+CWF (MDRF)	64.01	52.91
+MDRF+MSC-S	65.57	53.31
+MDRF+MSC-C	64.41	53.27
+MDRF+MSC	66.93	53.39
+MDRF+MSC+MCC-S	68.22	53.67
+MDRF+MSC+MCC-C	66.51	53.60
+MDRF+MSC+MCC	69.52	54.80

Table 1. Quantitative results (%) of ablation study. 'BS' denotes the baseline and 'BMDR' denotes the bi-directional modality difference reduction stage in MDRF. Meanwhile, MSC-S (MSC-C) and MCC-S (MCC-C) indicate only capturing the self (cross) long-range dependencies among cross-modality features along the spatial (channel) dimension, respectively.

are taken at daytime and 749 image pairs are taken at nighttime. There are 9 semantic classes, including the unlabeled background class. All of the images in this dataset have the same resolution of 480×640. For fair comparisons, we follow the same training, testing and verification setting as in [8]. We adopt the widely-used evaluation metrics (*i.e.*, mean Accuracy (mAcc) and mean Intersection over Union (mIoU)) to evaluate the performance of different models.

4.2. Implementation Details

The proposed network is implemented by PyTorch on an NVIDIA GTX 1080 Ti GPU. The stochastic gradient descent (SGD) method with a momentum of 0.9 and a weight decay of 0.0005 is adopted to train our proposed network. The initial learning rate is set to 0.01, which is decreased by adopting the exponential decay scheme with base 0.95 during training. Moreover, the training data are augmented by using random flipping, cropping and noise injecting techniques. We train the network about 300 epochs to its convergence.

4.2.1 Ablation Study

In this section, we validate the effectiveness of each component in our proposed model. The proposed MDRF subnetwork, MSC module and MCC module are first removed from our model as the baseline (denoted by 'BS'). Here, 'BMDR' denotes the bi-directional modality difference reduction stage in MDRF. Meanwhile, MSC-S (MSC-C) and MCC-S (MCC-C) indicate only capturing the self (cross) long-range dependencies among cross-modality features along the spatial (channel) dimension, respectively.

The quantitative experimental results are shown in Table 1. 'BS+BMDR' indicates that reducing the modality differences between multi-modality features benefits the exploitation of cross-modality complementary information, thus boosting the RGB-T semantic segmentation. Furthermore, it can also be observed that, compared with other fusion modules (*e.g.*, 'BS+BMDR+IFCNN', 'BS+BMDR+CW'), our proposed CWF module can more effectively select those discriminative features for semantic segmentation. The results of ('BS+MDRF+MSC-S' and 'BS+MDRF+MSC-C') and ('BS+MDRF+MSC+MCC-S' and 'BS+MDRF+MSC+MCC-C') indicate that the introduction of long-range dependencies along the spatial or channel dimension may provide more effective multi-scale contextual information for semantic segmentation. Meanwhile, the results of 'BS+MDRF+MSC+MCC' indicate that digging complementarity between the self and cross-spatial correlation matrices or the self and cross-channel correlation matrices can further facilitate the exploitation of long-range dependencies for RGB-T semantic segmentation.

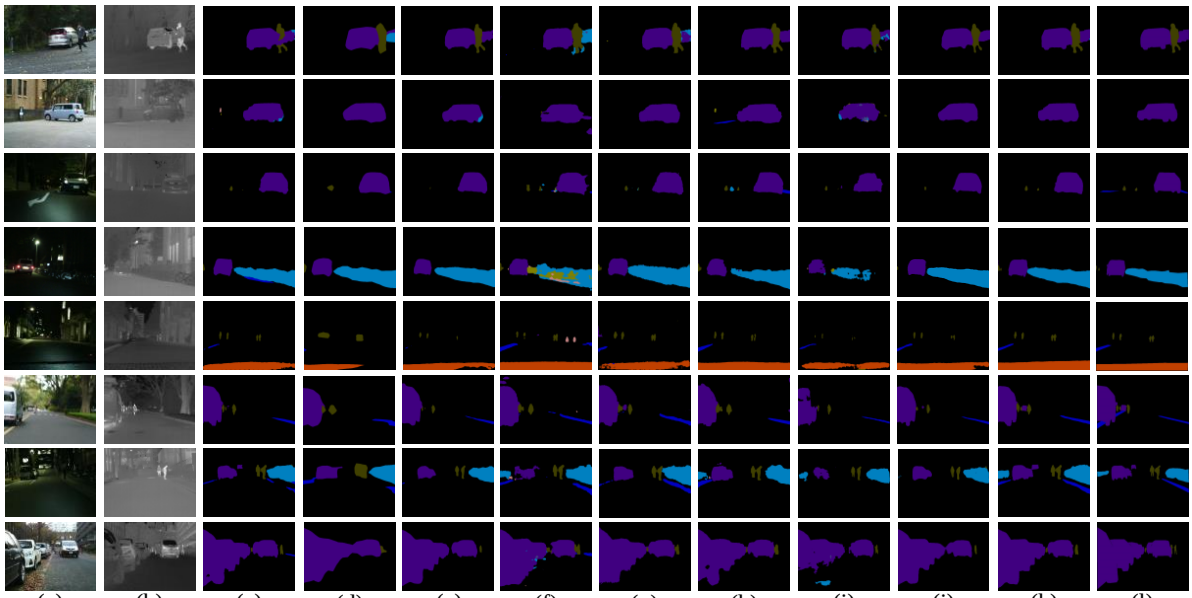
4.3. Comparison with State-of-the-art Methods

We compare our model with 9 state-of-the-art (SOTA) methods, including 3 deep learning based RGB semantic segmentation methods (DUC [26], DANet [6] and HRNet [22]), 3 RGB-T semantic segmentation approaches (MFNet [8], RTFNet [23] and PSTNet [20]) and 3 RGB-D semantic segmentation models (LDFNet [11], ACNet [10] and SA-Gate(ResNet-50) [4]). The procedure of converting the RGB semantic segmentation model into an extended RGB-T model is described as follows. First, we repeat the one-channel thermal image three times as a three-channel image. Then, their proposed networks are taken as the backbones of the RGB and thermal branches, respectively. Finally, the last output features before the prediction layers from the two branches are added and then fed into the prediction layers to obtain the final semantic segmentation maps. For the RGB-D models, we directly replace the input one-channel or HHA-encoded three-channel depth images with the one-channel thermal images or three-channel thermal images obtained by the same way in extending RGB models.

The quantitative results are shown in Table 2, which demonstrates that our method outperforms other SOTA methods by a large margin on the MFNet dataset. This indicates that our method can better exploit the complementary information from RGB-T images for semantic segmentation. Fig. 6 provides the visual comparisons of different models. As shown in the first two rows, under some simple scenes, most models can accurately segment targets. However, as shown in the 3rd-5th rows, our proposed method achieves significant superiorities over other SOTA models under poor lighting conditions. This owes to the strategy of bridging-then-fusing in MDRF. In addition, as shown in the 6th-8th rows, our method still outperforms other SOTA

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		
DUC[26]	98.8	97.7	92.4	82.5	84.1	69.4	71.3	58.9	58.4	40.1	25.5	20.9	17.3	3.4	60.0	42.1	52.2	40.9	61.2	50.7
DANet[6]	97.4	96.3	91.3	71.3	82.7	48.1	79.2	51.8	48.0	30.2	25.5	18.2	5.2	0.7	47.6	30.3	19.9	18.8	55.2	41.3
HRNet[22]	99.4	98.0	90.8	86.9	75.1	67.3	70.2	59.2	39.1	35.3	28.0	23.1	12.1	1.7	50.4	46.6	55.8	47.3	57.9	51.7
LDFNet[11]	96.2	95.3	87.0	67.9	83.9	58.2	82.7	37.2	67.4	30.4	32.9	20.1	8.2	0.8	67.4	27.1	55.6	46.0	64.6	42.5
ACNet [10]	97.6	96.7	93.7	79.4	86.8	64.7	77.8	52.7	57.2	32.9	51.5	28.4	7.0	0.8	57.5	16.9	49.8	44.4	64.3	46.3
SA-Gate[4]	98.2	96.8	86.0	73.8	80.8	59.2	69.4	51.3	56.7	38.4	24.7	19.3	0.0	0.0	56.9	24.5	52.1	48.8	58.3	45.8
MFNet[8]	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[23]	99.6	98.2	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
PSTNet[20]	—	97.0	—	76.8	—	52.6	—	55.3	—	29.6	—	25.1	—	15.1	—	39.4	—	45.0	—	48.4
Ours	98.6	97.8	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

Table 2. Quantitative results of different models (%) on the test set of [8]. The value 0.0 represents that there are no true positives. '—' denotes that the corresponding results are missed in [20].



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Figure 6. Visual comparisons of different methods. (a) RGB images; (b) Thermal images; (c) DUC [26]; (d) DANet [6]; (e) HRNet [22]; (f) LDFNet [11]; (g) ACNet [10]; (h) SA-Gate [4]; (i) MFNet [8]; (j) RTFNet [23]; (k) Ours; (l) GT.

models under complex scenes with multiple objects. This may benefit from the exploitation of multi-scale contextual information of cross-modality features and their long-range dependencies along the spatial and channel dimensions by using our proposed MSC and MCC modules.

5. Conclusion

In this paper, a novel ABMDRNet has been presented for RGB-T semantic segmentation, where the modality difference reduction and multi-scale contextual information are simultaneously considered. By virtue of the strategy of bridging-then-fusing, the proposed MDRF subnetwork can obtain higher discriminative cross-modality features than those traditional fusion modules do. This greatly improves the semantic segmentation performance of our proposed model. Owing to the proposed MSC and MCC modules, the multi-scale contextual information of the cross-modality features and their long-range dependencies along the spatial and channel dimensions are well exploited. Thanks to that, the issue of objects diversity in semantic segmentation can be addressed to a large extent. With the collaboration of these subnetworks and modules, our proposed RGB-T semantic segmentation model achieves new SOTA results on MFNet dataset.

Acknowledgements This work is supported by the National Natural Science Foundation of China under Grant No.61773301.

References

[1] Yanpeng Cao, Dayan Guan, Yulun Wu, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. *ISPRS*, 150:70–79, 2019. 1

[2] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *VCIP*, pages 1–4. IEEE, 2017. 2

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *PAMI*, 40(4):834–848, 2017. 2, 5

[4] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. *arXiv preprint arXiv:2007.09183*, 2020. 3, 7, 8

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 1

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019. 1, 2, 5, 7, 8

[7] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019. 1, 4

[8] 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 *IROS*, pages 5108–5115, 2017. 1, 2, 3, 6, 7, 8

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[10] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *ICIP*, pages 1440–1444, 2019. 3, 7, 8

[11] Shang-Wei Hung, Shao-Yuan Lo, and Hsueh-Ming Hang. Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation. In *ICIP*, pages 2374–2378, 2019. 3, 7, 8

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 2, 4

[13] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *CVPRW*, pages 11–19, 2017. 1

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1, 2

[15] Everingham Mark, Ali Eslami S, M, Van Gool Luc, Williams Christopher K, I, Winn John, and Zisserman Andrew. The pascal visual object classes challenge: A retrospective. volume 111, pages 98–136, 2015. 1

[16] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *CVPR*, pages 12416–12425, 2019. 3, 5

[17] 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. 2, 6

[18] Chang-Yong Ri and Min Yao. Semantic image segmentation based on spatial context relations. In *ISISE*, pages 104–108, 2012. 2

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 1, 2

[20] 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 *ICRA*, pages 9441–9447, 2020. 1, 2, 3, 7, 8

[21] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, pages 1–8, 2008. 2

[22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 7, 8

[23] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *RAL*, 4(3):2576–2583, 2019. 1, 2, 3, 7, 8

[24] Hiroki Tokunaga, Yuki Teramoto, Akihiko Yoshizawa, and Ryoma Bise. Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology. In *CVPR*, pages 12597–12606, 2019. 1

[25] Badrinarayanan Vijay, Kendall Alex, and Cipolla Roberto. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *PAMI*, 39(12):2481–2495, 2017. 1, 2

[26] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *WACV*, pages 1451–1460. IEEE, 2018. 7, 8

[27] Weiyue Wang and Ulrich Neumann. Depth-aware cnn for rgb-d segmentation. In *ECCV*, pages 135–150, 2018. 3

[28] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2, 3

[29] Pingping Zhang, Wei Liu, Hongyu Wang, Yinjie Lei, and Huchuan Lu. Deep gated attention networks for large-scale street-level scene segmentation. *PR*, 88:702–714, 2019. 1

[30] Qiang Zhang, Tonglin Xiao, Nianchang Huang, Dingwen Zhang, and Jungong Han. Revisiting feature fusion for rgb-t salient object detection. *TCSVT*, 2020. 1, 4

[31] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, pages 649–666, 2016. 2, 4

[32] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 2