# Mask-guided modality difference reduction network for RGB-T semantic segmentation

Neurocomputing 523 (2023) 9–17

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Mask-guided modality difference reduction network for RGB-T semantic segmentation

Wenli Liang [a], Yuanjian Yang [a], Fangyu Li [b], Xi Long [c], Caifeng Shan [a,*]

[a] College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao 266590, China
[b] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[c] Department of Electrical Engineering, Eindhoven University of Technology, 5612 AP Eindhoven, The Netherlands

ARTICLE INFO

*Article history:*
Received 2 September 2022
Revised 16 November 2022
Accepted 11 December 2022
Available online 15 December 2022

*Keywords:*
RGB-T semantic segmentation
Modality difference reduction
Multi-task learning

ABSTRACT

By exploiting the complementary information of RGB modality and thermal modality, RGB-thermal (RGB-T) semantic segmentation is robust to adverse lighting conditions. When fusing features from RGB images and thermal images, the existing methods design different feature fusion strategies, but most of these methods overlook the modality differences caused by different imaging mechanisms. This may result in insufficient usage of complementary information. To address this issue, we propose a novel Mask-guided Modality Difference Reduction Network (MMDRNet), where the mask is utilized in the image reconstruction to ensure that the modality discrepancy within foreground regions is minimized. Doing so enables the generation of more discriminative representations for foreground pixels, thus facilitating the segmentation task. On top of this, we present a Dynamic Task Balance (DTB) method to balance the modality difference reduction task and semantic segmentation task dynamically. The experimental results on the MFNet dataset and the PST900 dataset demonstrate the superiority of the proposed mask-guided modality difference reduction strategy and the effectiveness of the DTB method.



## 1. Introduction

Semantic segmentation aims to densely attach each pixel with a category label, and is widely used in many computer vision tasks, such as autonomous driving [1–3], medical diagnostics [4,5], geographic information system [6,7], and so on. Boosted by the extraordinary capability of convolutional neural networks (CNNs) in deriving meaningful image features, semantic segmentation based on deep CNNs has become prevalent in recent years [8–10].

Most deep learning based semantic segmentation methods are designed for RGB single-modality data, achieving prominent performance on many challenging large-scale datasets [11,12]. However, the RGB imaging sensors are highly sensitive to light and therefore susceptible to adverse lighting conditions, like darkness or overexposure. Instead, thermal imaging sensors can detect radiation of the wavelength up to 14 $\mu$m by imaging thermal radiation emitted by substances with the temperature above absolute zero [13]. Compared with RGB sensors, thermal sensors lose color information and detailed information but are robust to challenging lighting conditions. Thus, thermal images can complement RGB images with rich and clear contour information and semantic cues in challenging lighting conditions. With the increasing popularity of thermal sensors, RGB-T semantic segmentation has been investigated recently to improve the performance of scene segmentation [13–17].

Existing methods [13–15] for RGB-T semantic segmentation routinely employ two independent feature extractors to extract single-modality features from RGB modality and thermal modality, respectively, and then fuse these extracted features using a feature fusion strategy. For example, FuseSeg [15] utilizes a two-stage fusion strategy to fuse single-modality features. The thermal feature maps are hierarchically added with the RGB feature maps at the encoder, and then the fused feature maps are concatenated with the corresponding decoder feature maps. However, RGB images and thermal images have different imaging mechanisms. Therefore, integrating single-modality features without considering the modality difference caused by different imaging mechanisms inevitably leads to insufficient utilization of cross-modality complementary information.

Alternatively, Zhang et al. [16] recently proposed an Adaptive-weighted Bi-directional Modality Difference Reduction Network (ABMDRNet) by the bridging-then-fusing strategy, which first alleviate the modality differences and then fuse the multi-modality

* Corresponding author.
*E-mail address:* caifeng.shan@gmail.com (C. Shan).

https://doi.org/10.1016/j.neucom.2022.12.036
0925-2312/

features for RGB-T semantic segmentation. Specifically, the input RGB (thermal) image is first fed into a feature extractor to compute hierarchical single-modality features, which are then utilized to reconstruct a thermal (RGB) image by an image-to-image translation network. By minimizing the difference between the reconstructed thermal (RGB) image with the corresponding ground-truth thermal (RGB) image, the modality differences between RGB and thermal features could be reduced effectively. In other words, the core of the bridging stage is to supervise the RGB (thermal) hierarchical features being similar to the thermal (RGB) counterparts as much as possible, so that the modality difference caused by the imaging mechanism could be effectively reduced. The core of the fusing stage is to adaptively choose the discriminative cross-modality complementary information from the single-modality features after the bridging stage for fusion and exploit its contextual information for RGB-T semantic segmentation.

Despite preliminary experiments that have proved its effectiveness, ABMDRNet still suffers from two problems. Firstly, when using the image-to-image translation method to reduce the modality difference, ABMDRNet performs the processing on the entire image. In actual scenes, the pixels in the foreground region account for only a small proportion of the entire image, so performing the same processing on the entire image may result in the insufficient reduction of the modality difference in the foreground region. Secondly, the bridging-then-fusing strategy involves two tasks, the modality difference reduction task and the semantic segmentation task, but these two tasks fail to be balanced well in the ABMDRNet, because the fixed loss weights are adopted. However, the learning difficulty of these two tasks is different, and different tasks may be in different convergence stages, so the fixed weighting may lead to inadequate learning of the model.

To address these issues, we propose a Mask-guided Modality Difference Reduction Network (MMDRNet), with the aim to improve and extend the ABMDRNet for better RGB-T semantic segmentation. To tackle the aforementioned first problem, at the image reconstruction stage, we first separate the foreground and background from the entire image based on the given image mask and then process them individually. By using a feature extractor with better learning ability for the foreground region, we intend to reduce the modality discrepancy within the foreground region specifically. Doing so enables feature extractors to extract discriminative multi-modality information about the to-be-segmented foreground objects. To address the second problem, inspired by the work in [18], we propose a Dynamic Task Balance (DTB) method, which dynamically adjusts the weight of each task over time by considering the magnitude of the loss of each task. Doing so ensures that these two tasks with different learning difficulties can be well balanced and that both tasks can be balanced at the same convergence stage. The Dynamic Weight Average (DWA) method proposed in [18] adapts the task weighting over time by considering the rate of change of the loss for each task. However, DWA overlooks the magnitude of the loss of different tasks and requires manual tuning to balance the magnitude of the loss of different tasks before training. In contrast, our method can dynamically balance the magnitude of the loss of different tasks during training.

The main contributions of this paper are summarized as follows:

- We improve the image-to-image translation branch of ABMDRNet [16] by involving the mask that indicates the foreground region. The mask guidance enables our method to minimize the reconstruction errors of the foreground region specifically. It would potentially limit the modality difference reduction to the foreground region only rather than the entire image, thus making it more tractable.
- We introduce a multi-task loss optimization method for RGB-T semantic segmentation. Our proposed DTB method dynamically adjusts the weights of the image reconstruction task and the semantic segmentation task in the training by considering the magnitude of the loss of each task in each iteration. Doing so ensures that the learning difficulty and convergence phase of both tasks are well balanced, thus making the two-task learning more sufficient.

The rest of this article is organized as follows. In Section 2, we review the related work. In Section 3, we describe our method in detail. In Section 4, we present our experiments on two datasets. Conclusions and future work are presented in Section 5.

## 2. Related work

### 2.1. RGB semantic segmentation

In this section, we review the representative algorithms of deep learning-based RGB semantic segmentation. Shelhamer et al. [19] replaced the last fully connected layer of Convolutional Neural Networks (CNNs) [20] with a convolutional layer, and first proposed a Fully Convolutional Network (FCN) for pixel-level classification. Inspired by VGG16 and FCN, DeconvNet [21] is the first Encoder-Decoder architecture for semantic segmentation, which contains two symmetrical parts of the convolutional network and deconvolutional network. SegNet [22] also adopts a symmetric structure consisting of an encoder and a decoder, but it is more lightweight in the sense that there are no fully connected layers. U-Net [23] was originally proposed for biomedical image segmentation, but has been generalized and utilized in other domains. It combines the high-resolution information from the contracting path with the upsampled output to keep the spatial information, which is proven effective for semantic segmentation. DeepLab V1 [24] utilizes atrous convolution to expand the receptive field without increasing the number of parameters and adopts Conditional Random Fields (CRF) to refine the boundary. DeepLab V2 [25] adopts Atrous Spatial Pyramid Pooling (ASPP) with multiple sampling rates to make the target still segmentable when the target is represented in different sizes in the image. DeepLab V3 [26] applies atrous convolution to the cascade module and improves the ASPP module. Pyramid Scene Parsing Network (PSPNet) [27] was designed to consider more context information in semantic segmentation, which can avoid the mis-segmentation to a certain extent. Dense Upsampling Convolution (DUC) [28] introduces the hybrid dilated convolution for feature extraction. Hierarchical Decoupled Convolution Network (HDCNet) [29] was designed to alleviate grid problems caused by standard dilated convolution operations, which employs different dilation rates in different layers of the encoder.

To sum up, RGB semantic segmentation has been widely studied and has achieved good results on large-scale datasets in many application scenarios. However, these methods do not perform well under adverse lighting conditions, thus limiting their application.

### 2.2. RGB-T semantic segmentation

In recent years, with the popularity of thermal cameras, researchers have proposed to complement RGB imaging with thermal imaging for better semantic segmentation, e.g., road scene segmentation. Existing RGB-T semantic segmentation algorithms usually seek innovative fusion strategies to better utilize the complementary information between different modalities. In the early days, researchers mostly adopted simple fusion strategies, such as

concatenation [14] and element-wise summation [13]. Ha et al. proposed the Multi-spectral Fusion Network (MFNet) [14]. The mini-inception block with dilated convolution is used in RGB encoder and thermal encoder, respectively, to enlarge the size of the receptive field, and the shortcut is utilized to concatenate the feature maps from two encoders for improving up-sampling. Sun et al. proposed RGB-thermal Fusion Network (RTFNet) [13], which consists of two encoders and one decoder. In the encoder, the thermal feature maps are gradually integrated into the RGB encoder through element-wise summation. In the decoder, the Upception blocks with the residual structures are utilized to extract features and restore the resolution. Shivakumar et al. [30] designed PSTNet with dual-stream CNN architecture, the concatenation strategy is utilized to fuse the features from two modalities. Recently, researchers tend to explore more sophisticated fusion strategies. For example, FEANet [31] proposed a Feature-Enhanced Attention Module (FEAM) to enhance multi-level features and fuse RGB and thermal information in a complementary way. Sun et al. [15] proposed the FuseSeg by fusing RGB and thermal features with a two-stage fusion strategy. The feature maps from the thermal encoder are added into the RGB encoder gradually, and then the fused feature maps are concatenated with the corresponding decoder feature maps. MLFNet [32] was designed by employing multi-level skip connections and an auxiliary decoding module to capture contextual information comprehensively.

Nevertheless, the above approaches pay less attention to the modality difference caused by different imaging mechanisms, which may lead to insufficient information exploitation of the RGB image and thermal image. Considering this, ABMDRNet [16] proposed a bridging-then-fusing strategy. In the bridging stage, an image-to-image translation structure was introduced to reconstruct the entire image of the cross-modality to reduce the modality difference between RGB images and thermal images. In the fusing stage, a channel-wise weighted fusion module was proposed to capture the cross-modality information between the corresponding channels of single-modality RGB and thermal features. However, since the foreground with a small proportion of pixels in the entire image and the background with a large proportion of pixels are processed in the same way, the image-to-image translation network cannot optimally reduce the modality difference. Besides, ABMDRNet neglects the balance of the learning of different tasks in the network training, which may cause insufficient training, thus weakening the learning ability of the method.

### 2.3. Multi-task learning optimization

Multi-task learning (MTL) is to learn multiple related tasks together, and is common in the field of deep learning [18,33–35]. Different tasks may have problems such as different learning rates, different loss magnitudes, and mutual inhibition between tasks during the training. Existing MTL methods investigate multi-task architectures, optimization strategies and methods for learning task relationships [33]. A classic optimization strategy is Weighting by Uncertainty [34], where the loss weight is designed as a learnable parameter to model the uncertainty of each task. GradNorm [35] also treates the loss weights as learnable parameters, but the weights are optimized separately from the network parameters to realize ideal magnitudes of loss gradients for each task. However, in this method, each iteration requires additional computation of gradients, which affects the training efficiency when there are many parameters selected for balancing weights. DWA [18] was introduced as a computationally efficient alternative to GradNorm, and it only requires the numerical task loss, therefore it would not increase the complexity of the network. Because of its simpler implementation, DWA stands out among the multi-task learning optimization methods. However, it overlooks the different magnitude of the loss and manual tuning is needed in the DWA method.

## 3. Method

### 3.1. Architecture

As our backbone, ABMDRNet consists of three parts: Modality Difference Reduction and Fusion (MDRF) subnetwork, Multi-scale Spatial Context (MSC) module and Multi-scale Channel Context (MCC) module. The MDRF subnetwork employs a bridging-then-fusing strategy to reduce the modality difference first and then fuse the discriminative single-modality features. A bi-directional image-to-image translation based method is utilized to reduce the modality difference. A Channel Weighted Fusion (CWF) module is proposed in the fusing stage to adaptively select the discriminative multi-modality features for RGB-T semantic segmentation. MSC and MCC modules are designed to fully utilize the multi-scale contextual information of cross-modality features and their long-range dependencies in spatial and channel dimensions, respectively. More details can be found in [16].

Fig. 1 shows our proposed MMDRNet for RGB-T semantic segmentation. Specifically, the proposed network enhances the ABMDRNet from two aspects: 1) a better modality difference reduction strategy, and 2) the optimization of its multi-task learning.

As mentioned in Section 1, when reducing the modality difference based on the image-to-image translation, ABMDRNet processes the entire (RGB or thermal) image. In actual scenes, the to-be-segmented foreground region sometimes accounts for a small proportion of the entire image. For example, as shown in Fig. 2, in the MFNet dataset [14], the pixels in the foreground area only accounts for 7.862% of the entire dataset. Here, we define the foreground region as the to-be-segmented regions (labeled as various classes) in the image except the background region. We think that performing the same processing on the entire image will result in the insufficient reduction of the modality difference in the foreground region. To remedy this issue, we introduce the image mask in the modality difference reduction stage. By focusing on the foreground region in the image-to-image translation, our method makes the reconstructed foreground region as similar as possible to that of the real image. In this manner, the modality discrepancy within foreground regions is minimized, thus boosting the feature extractors to mining sufficient discriminative modality complementarity information for RGB-T semantic segmentation.

To better balance the modality difference reduction task and the semantic segmentation task in the network training, we propose a DTB method which dynamically adjusts the weight of each task over time by considering the magnitude of the loss of each task. By assigning larger weights to the more difficult task, the two tasks are dynamically balanced during training. Doing so ensures that the learning difficulty and convergence phase of both tasks are well balanced, thus making the two tasks learning more sufficient. As a contrast, our reference method DWA [18] adapts the task weighting over time by considering the rate of change of the loss for each task. In terms of computational complexity, both DTB and DWA only need to calculate the loss value of different iterations, so neither will increase the number of network parameters. However, DWA overlooks the different loss magnitudes of different tasks and requires balancing them to be similar manually at the beginning of the training. In contrast, DTB can dynamically balance the task magnitude during the training process.

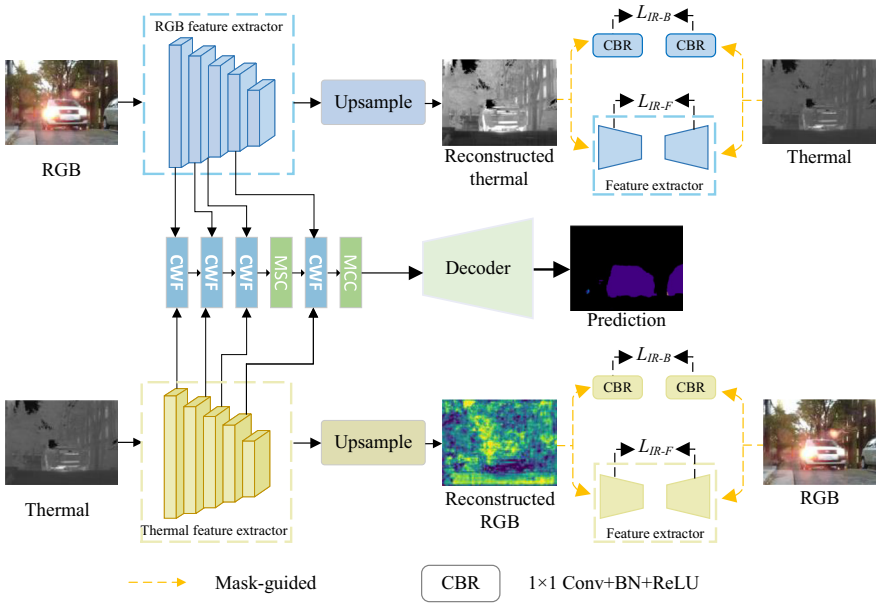


**Fig. 1.** Overall framework of our proposed model.

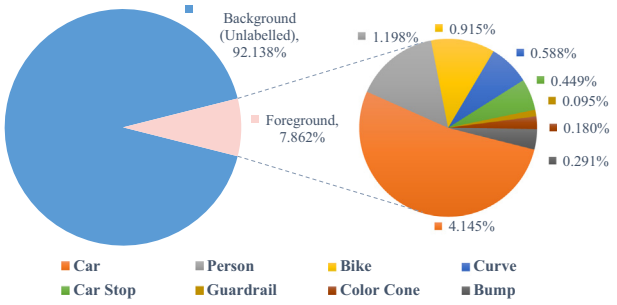


**Fig. 2.** Percentage of the number of pixels in each class in the MFNet dataset [14].

### 3.2. Mask-guided modality difference reduction

As shown in Fig. 1, to better reduce the modality difference and fuse the RGB and thermal modalities, we employ an image mask in the modality difference reduction stage. The image mask is generated by treating all the to-be-segmented regions of different classes as foreground (pixel value set as 1), while the pixels in the background region are set as 0.

In the mask-guided image-to-image translation, we perform pixel-wise multiplication between the generated image mask and the image to obtain the foreground image and background image. The operation details are shown by taking the process of transferring the RGB image to the thermal image as an example. Firstly, the ResNet50 without the fully connected layers and average pooling layers is utilized to extract the five-level RGB feature maps $\left\{F_n^{RGB}|n=1,2,3,4,5\right\}$. Secondly, the bilinear interpolation method is employed to upsample the last four-level feature maps $\left\{F_n^{RGB}|n=2,3,4,5\right\}$ to the original image size, and then they are added to obtain the reconstructed thermal image $F^{rec-T}$. After that, we multiply $F^{rec-T}$ with the image mask to obtain the foreground image and background image of $F^{rec-T}$ respectively, which is formulated as:

$$F^{rec-T-F} = F^{rec-T} * Mask, \tag{1}$$

$$F^{rec-T-B} = F^{rec-T} * (1 - Mask), \tag{2}$$

similarly, we also obtain the foreground image and background image of the corresponding ground-truth thermal image $F^{real-T}$ of $F^{rec-T}$ by image mask as follows:

$$F^{real-T-F} = F^{real-T} * Mask, \tag{3}$$

$$F^{real-T-B} = F^{real-T} * (1 - Mask), \tag{4}$$

where $*$ means pixel-wise multiplication. $F^{rec-T-F}$ and $F^{real-T-F}$ denotes the foreground image of $F^{rec-T}$ and $F^{real-T}$, respectively. $F^{rec-T-B}$ and $F^{real-T-B}$ denotes the background image of $F^{rec-B}$ and $F^{real-B}$, respectively.

After separating the foreground and background region of the reconstructed thermal image $F^{rec-T}$, we employ ResNet18 without the fully connected layers and average pooling layers as the feature extractor of $F^{rec-T-F}$, to extract its five-level features $\left\{F_n^{rec-T-F}|n=1,2,3,4,5\right\}$. To ensure that the reconstructed thermal image is similar to the corresponding ground-truth thermal image, we need to ensure that both the foreground and background regions of the reconstructed thermal image are similar to the corresponding regions of the ground-truth thermal image. With the aforementioned analysis, we need to pay more attention to the foreground region, so we employ a strategy that extracts simple features from the background region separately. Specifically, we process the background using the convolution $1 \times 1$ on $F^{rec-T-B}$, to obtain $\left\{F_n^{rec-T-B}|n=1\right\}$.

In order to ensure that the modality difference between RGB image and thermal image can be well reduced, we employ the following modality difference reduction loss $L_{IR}$ for supervision, i.e.,

$$L_{IR} = L_{IR-F} + L_{IR-B}, \tag{5}$$

where, $L_{IR-F}$ denotes the foreground loss and $L_{IR-B}$ denotes the background loss, respectively. $L_{IR-F}$ and $L_{IR-B}$ can be calculated as follows:

$$L_{IR-F} = \sum_{n=1}^{5}\left|F_n^{real-T-F} - F_n^{rec-T-F}\right| + \sum_{n=1}^{5}\left|F_n^{real-RGB-F} - F_n^{rec-RGB-F}\right|, \tag{6}$$

$$L_{IR-B} = \left| F_n^{real-T-B} - F_n^{rec-T-B} \right| + |F_n^{real-RGB-B} - F_n^{rec-RGB-B}|. \quad (7)$$

By minimizing the background features and foreground features between the reconstructed image and those of the matched real image with the above loss, the modality difference between the RGB image and the thermal image can be effectively reduced, thus improving the mining capability of RGB and thermal image feature extractors. As a result, the single modality features will contain more discriminative information of the other modality, which can further reduce the difference between the RGB feature and thermal features to a certain extent.

### 3.3. Multi-task learning optimization

Our method involves two tasks, the modality difference reduction task and the semantic segmentation task, which is a multi-task learning network. Since different tasks have different learning processes during the training, it is necessary to dynamically adjust the weights of the tasks for balancing the different tasks at the same learning level. However, fixed weights are utilized in the original ABMDRNet [16]. In this section, we propose a novel method DTB to solve this problem. The DWA [18] method learns to balance the task weighting by considering the learning rate of the different tasks to be similar. But whilst DWA requires adjusting the magnitude of each task loss to be consistent manually at the beginning of training, our DTB proposal adjusting the weight of each task over time by considering the magnitude of the loss of each task.

The weighting $\gamma_k$ for task $k$ is defined as:

$$\gamma_k(t) = K \frac{L_k(t)}{\sum_i L_i(t)}, \quad (8)$$

where $K$ ($K = 2$ here) is the number of tasks, is designed to ensure $\sum_i \gamma_i(t) = K$, ($i = 1, 2$ here). $t$ is the iteration index. $L_k(t)$ is the loss value, and in our experiment, it is calculated as the loss of the task in the current iteration, which can ensure that the weighting $\gamma_k(t)$ can truly fits the current training data.

The total loss function $L_{total}$ for training our model is composed of the image reconstruction task loss $L_{IR}$ and the semantic segmentation task loss $L_{seg}$.

$$L_{total} = \gamma_1 L_{Seg} + \gamma_2 L_{IR}, \quad (9)$$

where, $\gamma_1$ and $\gamma_2$ are calculated by DTB. $L_{seg}$ is calculated by the Cross Entropy Loss function.

## 4. Results and discussion

### 4.1. Datasets

Our model is verified on the MFNet dataset [14] and the PST900 dataset [30], which are the only two public datasets for RGB-T semantic segmentation.

MFNet dataset contains a total of 1569 annotated RGB and thermal image pairs, of which 820 are acquired during the day and 749 are acquired at night. Eight semantic classes of obstacles commonly encountered during driving (bike, person, car, curve, guardrail, car stop, bump and color cone) and an unlabelled background class are included in a total of nine classes. The dataset has been divided into three parts: a training set contains 50% daytime images and 50% nighttime images, a validation set and a test set each contains 25% daytime images and 25% nighttime images. All of the images are resized to the same resolution of 480 × 640.

PST900 dataset is designed for the DARPA Subterranean Challenge, containing 894 matched RGB and thermal natural image pairs. The sensor head of the mobile robot platform for data collection consists of a Stereolabs Zed Mini stereo RGB camera, a FLIR Boson 320 camera, and an active illumination setup. This dataset includes four visible artifacts (fire-extinguisher, backpack, handdrill, survivor: thermal mannequin and human) of pixel-level human annotations.

### 4.2. Training details

We implement our proposed network using PyTorch 0.10.0 with the CUDA 10.2 and cuDNN 7.6.5 libraries. Our network is trained on a single NVIDIA Tesla V100 graphics card. For fair comparison, the experimental settings are the same as our baseline ABMDRNet. Specifically, the Stochastic Gradient Descent (SGD) optimization solver is used for training, with a momentum of 0.9 and a weight decay of 0.005. The initial learning rate is set to 0.01, with an exponential decay scheme adopted to gradually decrease it and the batch size is set as 2. Before each epoch, the input image is randomly shuffled. In addition, the adopted data augmentation methods include random flipping, cropping and noise injecting techniques. We train the network until its loss no longer decreases.

### 4.3. Evaluation measures

We use the widely-used evaluation measures, Accuracy (Acc) and Intersection-over-Union (IoU), to evaluate the segmentation performance. Specifically, Acc is calculated as the ratio of the true and the predicted values for each class. IoU is calculated as the ratio of the intersection and union of the two sets of true and predicted values for each class. The mean Acc (mAcc) and mean IoU (mIoU) are global evaluation measures, calculated by averaging Acc and IoU across all classes.

$$mAcc = \frac{1}{N} \sum_{i=1}^{N} \frac{p_{ii}}{\sum_{j=1}^{N} p_{ij}}, \quad (10)$$

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{p_{ii}}{\sum_{j=1}^{N} (p_{ij} + p_{ji}) - p_{ii}}, \quad (11)$$

where $N$ denotes the number of the object classes, $p_{ii}$ denotes the number of pixels for class $i$ correctly classified as class $i$, $p_{ij}$ denotes the number of pixels for class $i$ wrongly classified as class $j$ and $p_{ji}$ denotes the number of pixels for class $j$ wrongly classified as class $i$.

Additionally, the $F_1$ measure is used to comprehensively consider both mAcc and mIoU.

$$F_1 = 2 \cdot \frac{mAcc \cdot mIoU}{mAcc + mIoU}. \quad (12)$$

### 4.4. Ablation studies

In this section, we validate the effectiveness of our proposed modality difference reduction strategy and multi-task learning optimization method in the proposed network on the MFNet dataset.

*1) The Effectiveness of Mask-guided Modality Difference Reduction:* To validate the effectiveness of the proposed mask-guided modality difference reduction strategy, we utilize the ABMDRNet as the baseline (denoted as 'BS'). Then we implement the proposed mask-guided strategy on the baseline ('BS + Mask-guided'). The experimental results are shown in Table 1 where the results of 'BS' are cited from [16]. 'BS + Mask-guided' indicates that our proposed mask-guided modality difference reduction strategy can fur-

**Table 1**
Results of ablation experiments for mask-guided modality difference reduction strategy. Boldface values indicates the better results.

| Variants | mAcc | mIoU | F1 |
|---|---|---|---|
| BS | 69.5 | **54.8** | 61.3 |
| BS + Mask-guided | **71.5** | 54.7 | **62.0** |

ther reduce the modality differences between RGB features and thermal features. The reduction of the modality differences benefits the mining capability of the single-modality feature extractors, thus, more discriminative information of cross-modality can be extracted and fused for RGB-T semantic segmentation.

*2) The Effectiveness of DTB Method:* We implement the proposed DTB method on the baseline, and the results are given in Table 2. With the proposed DTB method, the modality difference reduction task and the semantic segmentation task can be balanced well, so that the network can be trained sufficiently, thus boosting the RGB-T semantic segmentation.

We also compare the DTB and the existing DWA method. We add them to the 'BS' and 'BS + Mask-guided' respectively. The results of ('BS + DTB' and 'BS + DWA') and ('BS + Mask-guided + DTB' and 'BS + Mask-guided + DWA') indicate that the employing of the multi-task optimization method enables each task to be learned sufficiently, thus boosting the network to obtain better performance. Meanwhile, the DTB method outperforms over DWA method.'BS + Mask-guided + DTB' indicates that separating the foreground and background of the reconstructed image and processing them accordingly can effectively reduce the difference between cross-modalities, and dynamically balancing different tasks helps to promote the two tasks learning more sufficient. Combined with the 'Mask-guided' strategy and the 'DTB' method, our performance improved by 1.9% over 'BS'.

### 4.5. Comparsion with state-of-the-art methods

#### 4.5.1. Evaluation on MFNet dataset

*1) Overall Results:* In this section, we compare the proposed MMDRNet with DUC [28], DANet [36], HRNet [37], LDFNet [38], ACNet [39], SA-Gate (ResNet-50) [40], D-CNN [41], MFNet [14], FuseNet [42], RTFNet [13], PSTNet [30], FuseSeg [15] and ABMDRNet [16]. The results of DUC, DANet, HRNet, LDRNet, ACNet, SA-Gate are obtained from [16] and the results of D-CNN and MFNet are obtained from [15] for comparison.

The quantitative results are shown in Table 3. Compared with the state-of-the-art (SOTA) methods, our proposed MMDRNet achieves competitive results in all categories, especially in small object detection and segmentation. This performance may owe to that the proposed modality difference reduction strategy can make the difference caused by the imaging mechanism between modalities further reduced. After that, the ability of the single-modality extractor to extract cross-modality discriminative information is improved, and thus more useful complementary information such as boundary and contour information can be extracted. In addition, the application of the DTB method ensures the network training is more sufficient, so the learning ability of the model is improved, and the ability to distinguish small targets is significantly enhanced. Compared with our baseline ABMDRNet, our predicted Guardrail class has the improvement of 22.9% in Acc and 2.6% in IoU. At the same time, the Bump class has the improvement of 6.8% in Acc and 3.4% in IoU. And our proposed method also performs well in other classes.

**Table 2**
Results of ablation experiments for DTB method. Boldface values indicates the best results.

| Variants | mAcc | mIoU | F1 |
|---|---|---|---|
| BS | 69.5 | 54.8 | 61.3 |
| BS + DTB | 71.2 | 55.1 | 62.1 |
| BS + DWA | 71.8 | 54.3 | 61.8 |
| BS + Mask-guided + DTB | 72.4 | **56.0** | **63.2** |
| BS + Mask-guided + DWA | **72.8** | 54.6 | 62.4 |

Fig. 3 gives the segmentation results on some typical images, which visually showing that our proposed MMDRNet outperforms most of the other methods on the MFNet dataset, especially in small object detection. As shown in the fourth column, only our method can segment the small Color Cone class (indicated by the red box).

*2) Daytime and Nighttime Results:* To further evaluate these methods, we test these methods on daytime set and nighttime set of the MFNet dataset, respectively. The experimental results are shown in Table 4, where the results of FuseNet [42], MFNet [14], RTFNet [13], FuseSeg [15] are obtained from [15]. The quantitative results show that our model achieves competitive results under different lighting conditions. In addition, it can be found that the overall effect of daytime is worse than that of nighttime, indicating that at night, the complementary information of the two modalities is better utilized. In the case of sufficient lighting conditions during the day, the possible registration errors [14] between the two modalities may result in poor performance of semantic segmentation.

#### 4.5.2. Evaluation on PST900 dataset

To further evaluate the effectiveness of our proposed model, we also conduct the quantitative analysis with some RGB-T models on the PST900 dataset [30]. The results are summarized in Table 5, where the results of MFNet [14], RTFNet [13] and PSTNet [30] are from [30], and the quantitative experimental results also demonstrate that our method is superior to other RGB-T semantic segmentation models.

## 5. Conclusions

In this paper, we have proposed a novel MMDRNet for RGB-T semantic segmentation. Firstly, we utilize the mask-guided modality difference reduction strategy to reduce the differences between RGB modality and thermal modality caused by different imaging mechanism, thus benefiting the capability of single-modality extractors to mining the discriminative complementary information from the cross-modality. Then, we propose a DTB method for balancing the modality difference reduction task and semantic segmentation task dynamically, to make both tasks sufficiently learned. Our experimental results confirmed that the proposed method is better than the baseline, and is very competitive with other SOTA methods. In the future, we would like to explore more efficient modality difference reduction methods for segmentation improvement.

## CRediT authorship contribution statement

**Wenli Liang:** Conceptualization, Methodology, Software, Validation, Writing - original draft. **Yuanjian Yang:** Conceptualization, Methodology, Validation, Writing - original draft. **Fangyu Li:** Conceptualization, Methodology, Writing - review & editing. **Xi Long:** Conceptualization, Methodology, Writing - review & editing. **Caifeng Shan:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration.

**Table 3**
Quantitative results (%) of different models on the test set of the MFNet dataset. The best three results for the corresponding column are highlighted in red, green and blue.

| Methods | Car | | Person | | Bike | | Curve | | Car Stop | | Guardrail | | Color Cone | | Bump | | mAcc | mIoU | F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | | | |
| DUC [28] | 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 | 55.5 |
| DANet [36] | 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 | 47.2 |
| HRNet [37] | 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 | 54.6 |
| D-CNN [41] | 85.2 | 77.0 | 61.7 | 53.4 | 76.0 | 56.5 | 40.2 | 30.9 | 41.3 | 29.3 | 22.8 | 8.5 | 32.9 | 30.1 | 36.5 | 32.3 | 55.1 | 46.1 | 50.2 |
| LDFNet [38] | 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 | 51.3 |
| ACNet [39] | 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 | 53.8 |
| SA-Gate [40] | 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 | 51.3 |
| FuseNet [42] | 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 | 48.8 |
| MFNet [14] | 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 | 25.5 | 30.0 | 27.7 | 45.1 | 39.7 | 42.2 |
| RTFNet [13] | 93.0 | 87.4 | 79.3 | 70.3 | 76.8 | 62.7 | 60.7 | 45.3 | 38.5 | 29.8 | 0.0 | 0.0 | 45.5 | 29.1 | 74.7 | 55.7 | 63.1 | 53.2 | 57.7 |
| PSTNet [30] | – | 76.8 | – | 52.6 | – | 55.3 | – | 29.6 | – | 25.1 | – | 15.1 | – | 39.4 | – | 45.0 | – | 48.4 | – |
| FuseSeg [15] | 93.1 | 87.9 | 81.4 | 71.7 | 78.5 | 64.6 | 68.4 | 44.8 | 29.1 | 22.7 | 63.7 | 6.4 | 55.8 | 46.9 | 66.4 | 47.9 | 70.6 | 54.5 | 61.5 |
| ABMDRNet [16] | 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 | 61.3 |
| Ours | 93.4 | 85.7 | 89.3 | 70.3 | 74.7 | 61.5 | 65.7 | 46.9 | 42.7 | 32.7 | 53.9 | 7.7 | 59.9 | 48.2 | 73.0 | 53.4 | 72.4 | 56.0 | 63.2 |

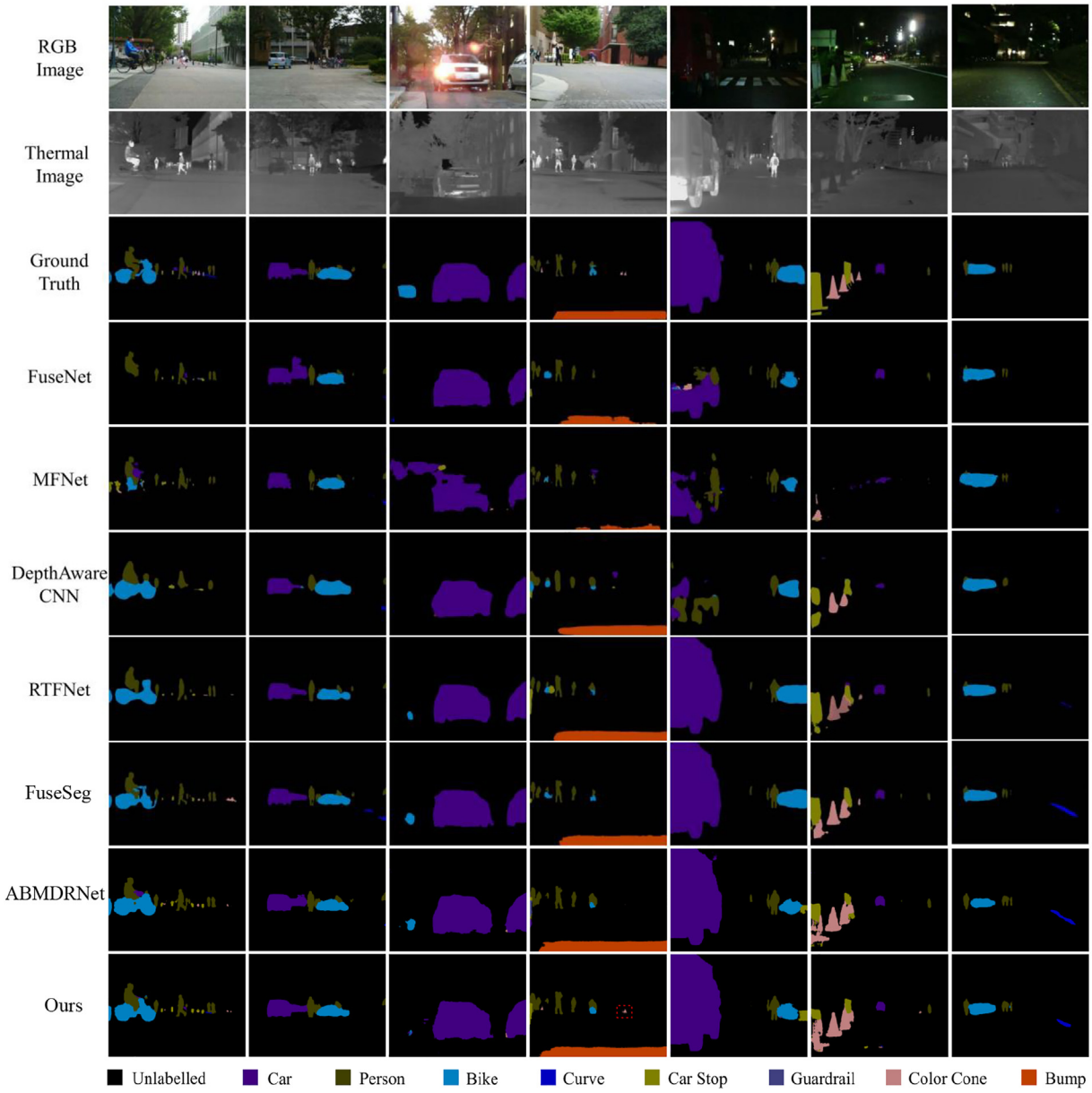


**Fig. 3.** Segmentation results of different methods in typical daytime images (left four columns) and nighttime images (right three columns).

**Table 4**
Quantitative results (%) of different models on the daytime set and nighttime set of [14]. The boldface value indicates the best results for the corresponding column.

| Methods | Daytime | | | Nighttime | | |
|---|---|---|---|---|---|---|
| | mAcc | mIoU | F1 | mAcc | mIoU | F1 |
| FuseNet [42] | 49.5 | 41.0 | 44.9 | 48.9 | 43.9 | 46.3 |
| MFNet [14] | 42.6 | 36.1 | 39.1 | 41.4 | 36.8 | 39.0 |
| RTFNet [13] | 60.0 | 45.8 | 51.9 | 60.7 | 54.8 | 57.6 |
| FuseSeg [15] | 62.1 | 47.8 | 54.0 | 67.3 | 54.6 | 60.3 |
| ABMDRNet [16] | 65.7 | 46.7 | 54.6 | 68.3 | 55.5 | 61.2 |
| Ours | **70.3** | **48.1** | **57.1** | **69.9** | **56.7** | **62.6** |

**Table 5**
Quantitative results (%) of different models on the test set of the PST900 dataset. The boldface value indicates the best results for the corresponding column.

| Methods | Background | | Fire-Extinguisher | | Backpack | | Hand-Drill | | Survivor | | mAcc | mIoU | F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | Acc | IoU | | | |
| MFNet [14] | – | 98.6 | – | **60.4** | – | 64.3 | – | 41.1 | – | 20.7 | – | 57.0 | – |
| RTFNet [13] | – | **98.9** | – | 52.0 | – | **75.3** | – | 25.4 | – | 36.4 | – | 57.6 | – |
| PSTNet [30] | – | 98.9 | – | 70.1 | – | 69.2 | – | 53.6 | – | 50.0 | – | 68.4 | – |
| ABMDRNet [16] | 99.4 | 98.7 | **89.4** | 54.9 | **89.9** | 72.9 | 67.8 | **57.6** | 24.4 | 24.1 | 74.2 | 61.6 | 67.3 |
| Ours | **99.5** | **98.9** | 77.9 | 52.4 | 79.1 | 71.1 | **77.1** | 40.6 | **69.8** | **62.3** | **81.3** | **68.7** | **74.5** |

## Data availability

The data that has been used is publicly available.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

This work is supported by the Natural Science Foundation of Shandong Province (Grant No. ZR2022MF322) and the Talent Introduction Program for Youth Innovation Teams of Shandong Province.

## References

[1] J.Y. Li, F.L. Jiang, J. Yang, B. Kong, M. Gogate, K. Dashtipour, A. Hussain, Lane-deeplab: Lane semantic segmentation in automatic driving scenarios for high-definition maps, Neurocomputing 465 (2021) 15–25.
[2] X. Wang, H.M. Ma, S.D. You, Deep clustering for weakly-supervised semantic segmentation in autonomous driving scenes, Neurocomputing 381 (2020) 20–28.
[3] C. Chen, K. Debattista, J. Han, Semi-supervised object detection via virtual category learning, arXiv preprint arXiv:2207.03433.
[4] Y.S. Ye, M.R. Chen, H.L. Zou, B.B. Yang, G.Q. Zeng, Gid: Global information distillation for medical semantic segmentation, Neurocomputing.
[5] Q. Wang, Y.K. Du, H.J. Fan, C. Ma, Towards collaborative appearance and semantic adaptation for medical image segmentation, Neurocomputing 491 (2022) 633–643.
[6] A. Song, Y. Kim, Deep learning-based hyperspectral image classification with application to environmental geographic information systems, Korean J. Remote Sens. 33 (2017) 1061–1073.
[7] K. Alkiek, A. Othman, H. Rizk, M. Youssef, Deep learning-based floor prediction using cell network information, AGIS, 2020, pp. 663–664.
[8] T.Y. Wu, S. Tang, R. Zhang, J. Cao, Y.D. Zhang, Cgnet: A light-weight context guided network for semantic segmentation, IEEE Trans. Image Process. 30 (2020) 1169–1179.
[9] Q. Wang, J. Gao, X. Li, Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes, IEEE Trans. Image Process. 28 (9) (2019) 4376–4386.
[10] Y. Liu, D. Zhang, Q. Zhang, J. Han, Part-object relational visual saliency, IEEE Transactions on Pattern Analysis and Machine Intelligence.
[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, CVPR, 2016, pp. 3213–3223.
[12] H. Fu, M. Gong, C. Wang, D. Tao, Moe-spnet: A mixture-of-experts scene parsing network, Pattern Recogn. 84 (2018) 226–236.
[13] Y. Sun, W. Zuo, M. Liu, Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes, IEEE Robot. Autom. Lett. 4 (3) (2019) 2576–2583.
[14] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, T. Harada, Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes, IROS, 2017, pp. 5108–5115.
[15] Y. Sun, W. Zuo, P. Yun, H. Wang, M. Liu, Fuseseg: semantic segmentation of urban scenes based on rgb and thermal data fusion, IEEE Trans. Autom. Sci. Eng. 18 (3) (2020) 1000–1011.
[16] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, J. Han, Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation, CVPR, 2021, pp. 2633–2642.
[17] Z. Shao, J. Han, D. Marnerides, K. Debattista, Region-object relation-aware dense captioning via transformer, IEEE Transactions on Neural Networks and Learning Systems.
[18] S. Liu, E. Johns, A.J. Davison, End-to-end multi-task learning with attention, CVPR, 2019, pp. 1871–1880.
[19] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, CVPR, 2015, pp. 3431–3440.
[20] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25.
[21] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, ICCV, 2015, pp. 1520–1528.
[22] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[23] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, MICCAI, 2015, pp. 234–241.
[24] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062.
[25] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2017) 834–848.
[26] L. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587.
[27] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, CVPR, 2017, pp. 2881–2890.
[28] P.Q. Wang, P.F. Chen, Y. Yuan, D. Liu, Z.H. Huang, X.D. Hou, G. Cottrell, Understanding convolution for semantic segmentation, WACV, 2018, pp. 1451–1460.
[29] Z. Luo, Z. Jia, Z. Yuan, J. Peng, Hdc-net: Hierarchical decoupled convolution network for brain tumor segmentation, IEEE J. Biomed. Health Inform. 25 (3) (2020) 737–745.
[30] S.S. Shivakumar, N. Rodrigues, A. Zhou, I.D. Miller, V. Kumar, C.J. Taylor, Pst900: Rgb-thermal calibration, dataset and segmentation network, ICRA, 2020, pp. 9441–9447.
[31] F.Q. Deng, H. Feng, M.J. Liang, H.M. Wang, Y. Yang, Y. Gao, J.F. Chen, J.J. Hu, X.Y. Guo, T.L. Lam, Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation, IROS, 2021, pp. 4467–4473.
[32] Z. Guo, X. Li, Q. Xu, Z. Sun, Robust semantic segmentation based on rgb-thermal in variable lighting scenes, Measurement 186 (2021).

[33] M. Crawshaw, J. Košecká, Slaw: Scaled loss approximate weighting for efficient multi-task learning, arXiv preprint arXiv:2109.08218.
[34] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, CVPR, 2018, pp. 7482–7491.
[35] Z. Chen, V. Badrinarayanan, C. Lee, A. Rabinovich, Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks, ICML, 2018, pp. 794–803.
[36] J. Fu, J. Liu, H.J. Tian, Y. Li, Y.J. Bao, Z.W. Fang, H.Q. Lu, Dual attention network for scene segmentation, CVPR, 2019, pp. 3146–3154.
[37] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, CVPR, 2019, pp. 5693–5703.
[38] S. Hung, S. Lo, H. Hang, Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation, ICIP, 2019, pp. 2374–2378.
[39] X. Hu, K. Yang, L. Fei, K. Wang, Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation, ICIP, 2019, pp. 1440–1444.
[40] X. Chen, K. Lin, J. Wang, W. Wu, C. Qian, H. Li, G. Zeng, Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation, ECCV, 2020, pp. 561–577.
[41] W. Wang, U. Neumann, Depth-aware cnn for rgb-d segmentation, ECCV, 2018, pp. 135–150.
[42] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, ACCV, 2016, pp. 213–228.

**Wenli Liang** received the B.Eng. degree from the Shandong Jiaotong University, Jinan, China, in 2018 and the M.Eng. degree from the Shandong University of Science and Technology, Qingdao, China, in 2021. She is currently pursuing the Ph.D. degree in Control Science and Engineering with the Shandong University of Science and Technology, Qingdao, China. Her research interests include computer vision and deep learning.

**Yuanjian Yang** received the B.S degree from the College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao, China, in 2021. He is currently pursuing the M.S. degree at the College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao, China. His research interest is cross-modal image fusion for computer vision.

**Fangyu Li** received the Ph.D. in Computational Geophysics from The University of Oklahoma in 2017. His Master (2013) and Bachelor (2009) degrees were both in Electrical Engineering, obtained from Tsinghua University and Beihang University, respectively. From 2017 to 2020, he was a postdoctoral fellow with the College of Engineering, University of Georgia. From 2020 to 2021, he was an assistant professor with the Department of Electrical and Computer Engineering at Kennesaw State University. Dr. Li is currently a full professor and PhD supervisor with the Faculty of Information Technology at Beijing University of Technology. His research interests include complex signal processing, machine learning, deep learning, distributed computing, complex system modeling and monitoring, Internet of things (IoT), and cyber-physical systems (CPS).

**Xi Long** received the B.Eng. degree in electronic information engineering from Zhejiang University, Hangzhou, China, in 2006, and the M.Sc. and the Ph.D. (cum laude) degrees in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 2009 and 2015, respectively. From 2010 to 2011, he was with Tencent, China, on data mining and user research. He has more than ten years of Research and Development experience in healthcare industry, with Philips Research, Eindhoven. He is currently an Associate Professor with the Eindhoven University of Technology and a Senior Scientist with Philips Research. His research interests include signal processing, data analytics, and machine learning in healthcare and medical applications. He has published more than 100 articles and reports in these fields and his inventions led to more than ten patent applications.

**Caifeng Shan** received the B.Eng. degree from the University of Science and Technology of China (USTC), the M.Eng. degree from the Institute of Automation, Chinese Academy of Sciences, and the PhD degree from Queen Mary, University of London. His research interests include computer vision, pattern recognition, medical image analysis, and related applications. He has authored 140 papers and 80 granted patents. He has served as Associate Editor for scientific journals including IEEE Journal of Biomedical and Health Informatics, Neurocomputing, and IEEE Transactions on Circuits and Systems for Video Technology. He is a Senior Member of IEEE.