
Article:
Han, J., Shenlu, Z., Liu, Y. et al. (2 more authors) (2023) Mitigating modality discrepancies for RGB-T semantic segmentation. IEEE Transactions on Neural Networks and Learning Systems, 35 (7). pp. 9380-9394. ISSN 2162-237X

https://doi.org/10.1109/TNNLS.2022.3233089

Mitigating Modality Discrepancies for RGB-T Semantic Segmentation

Shenlu Zhao, Yichen Liu, Qiang Jiao, Qiang Zhang* and Jungong Han

***Abstract*—Semantic segmentation models gain robustness against adverse illumination conditions by taking advantage of complementary information from visible and thermal infrared (RGB-T) images. Despite its importance, most existing RGB-T semantic segmentation models directly adopt primitive fusion strategies, such as element-wise summation, to integrate multi-modal features. Such strategies, unfortunately, overlook the modality discrepancies caused by inconsistent unimodal features obtained by two independent feature extractors, thus hindering the exploitation of cross-modal complementary information within the multi-modal data. For that, we propose a novel network for RGB-T semantic segmentation, *i.e.* MDRNet+, which is an improved version of our previous work ABMDRNet [1]. The core of MDRNet+ is a brand new idea, termed the strategy of bridging-then-fusing, which mitigates modality discrepancies before cross-modal feature fusion. Concretely, an improved Modality Discrepancy Reduction (MDR+) subnetwork is designed, which first extracts unimodal features and reduces their modality discrepancies. Afterwards, discriminative multi-modal features for RGB-T semantic segmentation are adaptively selected and integrated via several Channel Weighted Fusion (CWF) modules. Furthermore, a Multi-Scale Spatial Context (MSC) module and a Multi-Scale Channel Context (MCC) module are presented to effectively capture the contextual information. Last, we elaborately assemble a challenging RGB-T semantic segmentation dataset *RTSS* for urban scene understanding to mitigate the lack of well-annotated training data. Comprehensive experiments demonstrate that our proposed model surpasses other state-of-the-art models on the MFNet, PST900 and *RTSS* datasets remarkably.**

***Index Terms*—RGB-T Semantic Segmentation, Bridging-then-fusing, Modality Discrepancy Reduction, Contextual Information, Dataset.**

I. Introduction

SEMANTIC segmentation aims to associate pixel-level category labels to the objects in the scenes. As a pivotal yet challenging scene understanding technology, it plays an important role in a variety of computer vision applications, including autonomous driving [2]–[4], agriculture monitoring [5], pathological analysis [6]–[8] and so on.

With the emergence of Fully Convolutional Networks (FCNs) [9], Deep Convolutional Neural Networks (DCNNs) based RGB semantic segmentation methods [10]–[18] have been proposed successively and achieved prominent performance in many challenging large-scale datasets [19], [20]. Unfortunately, their performance usually degrades significantly under much weak or strong illumination conditions due to the light sensitivity of RGB sensors. Specifically, in common urban scenes, the failures of semantic segmentation algorithms based on RGB data are often due to: 1) The illumination condition is good on the whole but there are some indistinguishable low contrast areas, such as objects containing similar colors with the backgrounds (*e.g.*, the green dotted box in the first row of Fig. 1 (a)) or different categories of objects with similar spatial appearances (*e.g.*, the red dotted box in the second row of Fig. 1 (a)). Here, we call these cases 'Confusing Illumination' ('CI'); 2) The overall illumination conditions are so weak or strong that most objects and backgrounds are almost invisible, as shown in the third and fourth rows of Fig. 1 (a), which are named as 'Extreme Illumination' ('EI').

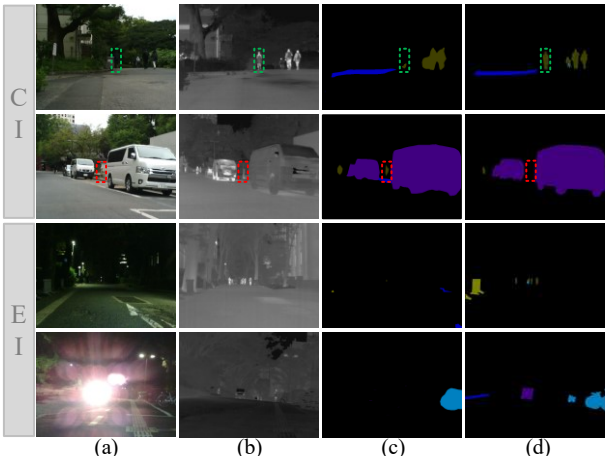

Fig. 1. Two typical categories of failure scenarios for semantic segmentation algorithms based on RGB data. (a) RGB images; (b) TIR images; (c) Semantic segmentation masks induced from only RGB images; (d) Semantic segmentation masks induced from RGB-T image pairs.

To make up for the deficiency of RGB sensors, considerable researches have been devoted to combining information provided by other sensors. Compared to RGB sensors, thermal infrared (TIR) sensors show stronger robustness against illumination or weather changing, which dedicate to providing clear contour and semantic cues for targets with temperatures above absolute zero. In recent years, with the prevalence of TIR sensors, researchers have begun to appreciate the preponderance of TIR sensors to assist RGB sensors for achieving reliable semantic segmentation under diversely adverse illumination conditions. So far, some relevant exploratory works [1], [3],

Shenlu Zhao, Yichen Liu, Qiang Jiao and Qiang Zhang are with the Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an, Shaanxi 710071, China, and are also with the Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China. Email: zhaoshenlu@stu.xidian.edu.cn, xdulyc@163.com, qjiao@xidian.edu.cn and qzhang@xidian.edu.cn.

Jungong Han is with Computer Science Department, Aberystwyth University, SY23 3FL, UK. Email: jungonghan77@gmail.com.

*Corresponding author: Qiang Zhang.

A preliminary version of this work has appeared in CVPR 2021 [1].

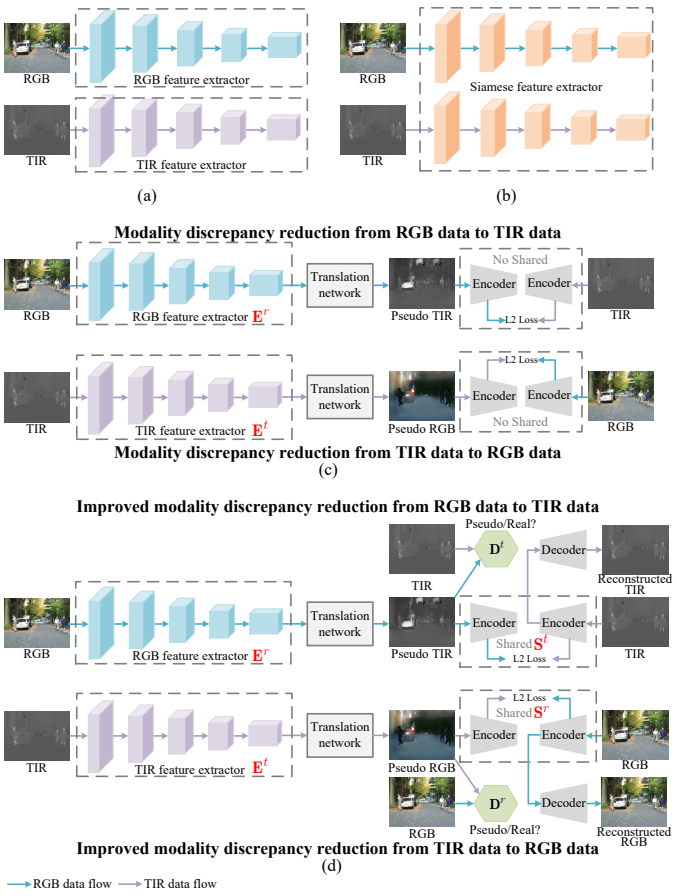

Fig. 2. Different pipelines for unimodal feature extraction. (a) The RGB and TIR images are routinely fed into two independent feature extractors to obtain their unimodal features, respectively, which completely ignores the modality discrepancies; (b) The RGB and TIR images are fed into a Siamese feature extractor to obtain their unimodal features, respectively, which can intuitively reduce modality discrepancies. However, the huge distribution discrepancies between multi-modal data make it difficult for the Siamese feature extractor to well represent the features of different modalities, resulting in inferior prediction performance; (c) The previous modality discrepancy reduction subnetwork presented in ABMDRNet [1] employs a cross-modal image translation based way to reduce modality discrepancies from the perspective of unimodal feature enhancement; (d) The MDR+ subnetwork further ensures the effectiveness and stability of modality discrepancy reduction by jointly introducing Siamese feature extractors, adversarial learning and image reconstruction constraints.

[21]–[27] have been consecutively proposed.

As illustrated in Fig. 2 (a), most of existing RGB-T semantic segmentation methods [3], [21]–[26], [28], [29] routinely adopt two independent feature extractors to extract unimodal features from different modalities separately. For example, [21], [23], [26], [28], [29] adopt two symmetrical ResNets [30] to extract multi-level RGB and TIR features. Unfortunately, they usually ignore the modality discrepancies between multi-modal features. Specifically, the unimodal features extracted by two irrelevant feature extractors with ZERO interaction indicate that the RGB data and their paired TIR data are mapped into two heterogeneous feature spaces, which may result in the lack of comparability and compatibility between the two types of unimodal features, thus hindering the exploitation of cross-modal complementary information. This may be one of the rational reasons why the complementary information between RGB and TIR modalities is difficult to

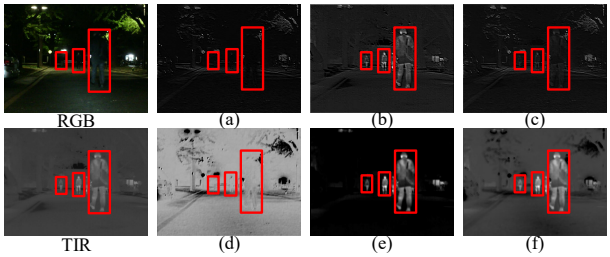

Fig. 3. Illustration of modality discrepancy reduction. (a)-(c): Original RGB features, TIR features and their fused features obtained by element-wise summation, respectively. (d)-(f): RGB features, TIR features and their fused features obtained by element-wise summation after reducing modality discrepancies, respectively. The visualizations illustrated here are all from the same channel of their own features.

exploit crudely. More intuitively, such negligence of modality discrepancies will be reflected in the dramatic weakening of the discriminative complementary information in the cross-modal fused features. For example, the people regions, marked by the red boxes in Fig. 3 (a), have intensity values close to the backgrounds, while the same regions in Fig. 3 (b) have higher intensity values than the backgrounds. If the summation fusion is directly employed without considering the modality discrepancies, the complementary information with higher discriminability provided by TIR data will be undesirably suppressed in the cross-modal fused features, as shown in Fig. 3 (c).

Considering the above issue, an intuitive solution is to extract both RGB and TIR features by employing a Siamese unimodal feature extractor [31] with the same network structures and parameters, as shown in Fig. 2 (b). The Siamese network has been extensively used to map the data from different modalities into a common feature space. However, due to the insurmountable distribution discrepancies between RGB data and TIR data, it is almost impossible to find a suitable common feature space to completely represent the information of the two modalities simultaneously, thus resulting in inferior prediction performance. Alternatively, inspired by some knowledge transfer methods such as [32], our previous work [1] proposed a roundabout strategy to alleviate the problem of modality discrepancies, as shown in Fig. 2 (c). Specifically, RGB (TIR) data is first fed into a feature extractor $\mathbf{E}^r$ ($\mathbf{E}^t$) to capture multi-level RGB (TIR) features, which are then employed to generate the pseudo TIR (RGB) data by a translation network. On top of that, the modality discrepancies will be effectively reduced by minimizing the distance between the features of pseudo TIR (RGB) data and those of matched real TIR (RGB) data. In fact, the core idea of modality discrepancy reduction is to potentially improve the feature mining capability of the unimodal feature extractors (*i.e.,* $\mathbf{E}^r$ and $\mathbf{E}^t$ in Fig. 2 (c)), thus achieving modality discrepancy reduction from the perspective of unimodal feature enhancement. As a result, the extracted unimodal features will certainly contain some discriminative information of another modality in addition to the information of their own modality, as shown in Fig. 3 (d) and Fig. 3 (e). As visualized in Fig. 3 (f), the fused features obtained after reducing modality discrepancies preserve more cross-modal complementary information than those original fused features in Fig. 3 (c).

Despite its effectiveness, our previous modality discrepancy reduction subnetwork used in [1] suffers from two crises. Firstly, the features of pseudo data and those of matched real data with the same modality extracted by two independent feature extractors are still represented in two heterogeneous feature spaces, which makes it unreasonable to measure the distance between them directly. Secondly, minimizing the distance between the features of pseudo data and those of matched real data with the same modality without considering the feature extraction capability of their corresponding feature extractors may lead to the training collapse of modality discrepancy reduction. For example, the values of pseudo data features and those of matched real data features with the same modality may be greatly close but meaningless (*e.g.*, all of them are 0) in some extreme cases. This will lead to the failure of modality discrepancy reduction.

This paper intends to further improve and extend our conference paper [1] by addressing the aforementioned two problems. Based on the above analysis, we present, an MDRNet+ for RGB-T semantic segmentation in this paper, where an improved Modality Discrepancy Reduction (MDR+) subnetwork is proposed to better achieve the modality discrepancy reduction. As shown in Fig. 2 (d), in order to rationalize the measurement of feature distances, we employ a Siamese feature extractor (see $\mathbf{S}^t$ or $\mathbf{S}^r$ in Fig. 2 (d)) to represent the features of pseudo data and those of matched real data with the same modality in a common feature space, instead of using two independent feature extractors. Moreover, in order to further facilitate the Siamese feature extractor ($\mathbf{S}^t$ or $\mathbf{S}^r$) to find a suitable common feature space for well representing the information of pseudo and real data simultaneously, we enforce the distribution of generated pseudo data and that of matched real data with the same modality to be as close as possible via adversarial learning. Finally, we consider image reconstruction as an auxiliary to constrain the feature extraction capability of the Siamese feature extractor ($\mathbf{S}^t$ or $\mathbf{S}^r$). Owing to these improvements, the feature mining capability of unimodal feature extractors (*i.e.,* $\mathbf{E}^r$ and $\mathbf{E}^t$ in Fig. 2 (d)) will be potentially improved by simply but reasonably minimizing the distance between the features of pseudo data and those of matched real data with the same modality, thus effectively and stably reducing modality discrepancies.

Apart from algorithmic problems, the lack of data with annotations, especially for urban scenes, is another major bottleneck to hold back the development of RGB-T semantic segmentation. In this paper, we assemble a challenging RGB-T Semantic Segmentation (*RTSS*) dataset for urban scene understanding. We elaborately select 1880 pairs of matched RGB-T images from KAIST [33], LasHeR [34] and RGBT234 [35], and manually label them at the pixel level. These labels include four categories: human, non-motor, car and background, which focus on the most important objects (usually having higher temperature values) in urban scenes for autonomous driving and traffic dispersion. Similar to the most widely used RGB-T semantic segmentation dataset MFNet [3], we also divide *RTSS* into daytime and nighttime parts. Moreover, we further divide *RTSS* into three categories of scenarios according to illumination conditions, including one category of scenarios with good illumination conditions both in whole and in part (here is named as General Illumination ('GI')) and two typical categories of failure scenarios for semantic segmentation algorithms based on RGB data (*i.e.,* 'CI' and 'EI').

The main contributions of our work are summarized as follows:

- We propose an MDRNet+ based on a novel strategy of bridging-then-fusing, which innovatively mitigates the modality discrepancies before cross-modal feature fusion to effectively exploit cross-modal complementary information with high discriminability for RGB-T semantic segmentation.
- Different from our previous ABMDRNet, in this paper, we propose an improved MDR+ subnetwork, which rationalizes the distance measurement between the features of pseudo data and those of matched real data with the same modality and alleviates the training collapse problem by jointly introducing Siamese feature extractors, adversarial learning and image reconstruction constraints, thus providing a more effective and more stable modality discrepancy reduction process.
- We carefully assemble a challenging RGB-T semantic segmentation dataset *RTSS* for urban scene understanding, which provides a timely supplement for training such autonomous driving algorithms based on RGB-T semantic segmentation.

The remainder of this paper is organized as follows. Section II discusses related works on RGB semantic segmentation and multi-modal semantic segmentation. Section III details the proposed MDRNet+. Section IV provides a more detailed analysis of our *RTSS*. Experimental results, performance evaluations and comparisons are included in Section V. Finally, conclusions are drawn in Section VI.

II. Related Work

A. RGB Semantic Segmentation

Early RGB semantic segmentation methods [36], [37] mainly rely on low-level hand-crafted features to segment objects and backgrounds. Recently, deep learning based semantic segmentation models have become the mainstream and achieved remarkable performance. Fully Convolutional Network (FCN) [9], as a pioneer, arouses the research upsurge of semantic segmentation. Then, Noh *et al.* [10] proposed the first Encoder-Decoder architecture for semantic segmentation, which is simple but efficient and is still one of the mainstream architectures used in semantic segmentation models to date. To preserve more precise spatial information in encoder, Sun *et al.* [13] proposed the HRNet by performing repeated multi-scale fusions to boost the high-resolution representations with the help of the low-resolution representations of the same depth and similar levels. As well, to address the problem of the diversity of objects, Chen *et al.* [14] proposed the Atrous Spatial Pyramid Pooling (ASPP) to capture the discriminative multi-scale contextual information by several atrous convolutional layers with different dilation rates. Although such multi-scale contextual information extraction modules have achieved great successes in semantic segmentation, their receptive fields

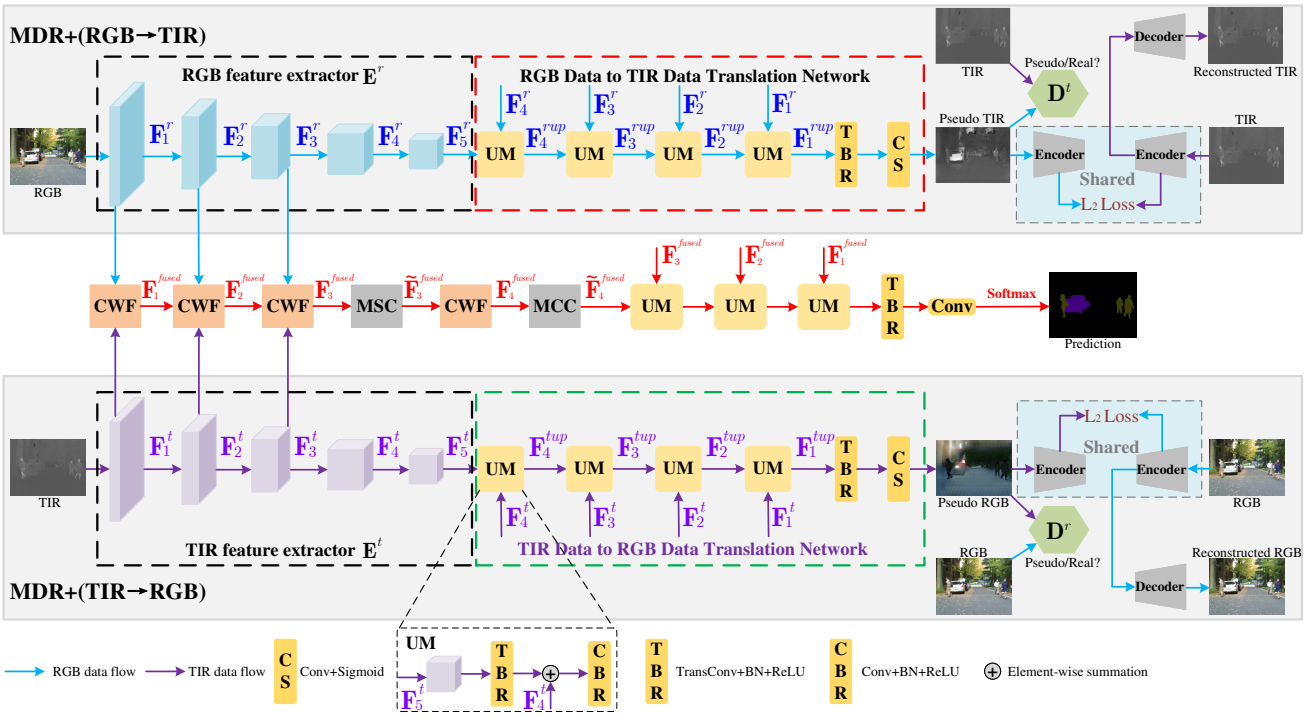

Fig. 4. Overall framework of our proposed MDRNet+. The proposed model consists of the MDR+ subnetwork for unimodal feature extraction as well as for modality discrepancy reduction, several CWF modules for cross-modal feature fusion and the MSC and MCC modules for contextual information exploitation.

are still limited, thus failing to exploit the global contextual information. Recently, many models try to exploit the long-range dependencies to address such an issue and have achieved promising results. For example, Li *et al.* [11] jointly learned the long-range semantic and spatial dependencies to mine the global contexts for semantic segmentation. Ding *et al.* [17] introduced the pixel-wise semantic attention block and the category-wise semantic attention block to capture more explicit contextual dependencies in a low computational complexity. In addition, some works [18], [38] focus on exploring some lightweight frameworks to balance the parameters and accuracy for real-time semantic segmentation.

B. Multi-modal Semantic Segmentation

Recently, with the development of imaging techniques [3], [22], [33]–[35], [39], [40], many studies employ multi-modal data (*e.g.,* RGB-D images and RGB-T images) to address some issues arising from the RGB semantic segmentation.

1) RGB-D Semantic Segmentation: The depth sensors aim to provide rich information of geometrical structures and spatial layouts to improve segmentation performance in complex scenes. To better integrate RGB information and depth information, Hazirbas *et al.* [41] proposed the pioneering work based the encoder–decoder structure, *i.e.,* FuseNet, in which multi-level RGB and depth features are integrated progressively by summation fusion. Hung *et al.* [42] proposed LDFNet to enhance the fused multi-modal features by introducing the luminance information for depth image. Hu *et al.* [43] 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. In addition, unlike the idea of fusing RGB and depth features conventionally, some methods exploit depth information to improve the standard convolution. For example, Wang *et al.* [44] proposed a depth-aware CNN to integrate the geometry information within depth images into CNN by augmenting the standard convolution with a depth similarity term, which does not introduce any parameters and computation complexity but boosts the RGB-D semantic segmentation performance. As well, to solve the problem of noise in depth images, Chen *et al.* [45] proposed a separation-and-aggregation gate module to effectively diminish the influence of noisy depth measurements while incorporating sufficiently complementary information to facilitate RGB-D semantic segmentation. In addition, aiming at the diversity of objects, which also exists in RGB-D semantic segmentation, some methods focus on capturing multi-scale contextual information. Zhang *et al.* [46] proposed the NANet to exploit the non-local context of RGB-D features at multiple stages along the spatial and channel dimensions.

2) RGB-T Semantic Segmentation: Compared to depth sensors, TIR sensors dedicate to providing robust contour and semantic cues under extreme illumination conditions. Recently, with the development of TIR sensors, some RGB-T semantic segmentation methods are gradually proposed. MFNet [3], as the pioneer work for RGB-T semantic segmentation, first adopted a two-stream structure to extract unimodal features and then integrated them through custom shortcut blocks. Sun *et al.* [21] proposed the RTFNet, in which the multi-level RGB features and TIR features are first fused by element-wise summation and then several upsampling inception blocks are inserted to improve decoding performance. Shivakumar *et al.* [22] proposed the PSTNet to enhance the segmentation results by using the RGB segmentation mask as an additional input.

Guo *et al.* [23] proposed the MLFNet to capture contextual information comprehensively by employing multi-level skip connections in decoder. Sun *et al.* [24] proposed the FuseSeg, in which the multi-level RGB and TIR features are integrated through concatenation. Xu *et al.* [25] proposed an Attention Fusion Module (AFM) to improve the contextual correlation between the RGB and TIR features. Deng *et al.* [28] exploited attention mechanisms to excavate and enhance multi-level features from both the channel and spatial views. Zhou *et al.* [29] introduced prior edge information to enhance boundary extraction for finer RGB-T semantic segmentation. Zhou *et al.* [47] jointly used multi-label supervision to optimize the network in terms of semantic, binary and boundary characteristics. Moreover, in addition to the fusion strategy, MMNet [26] and ABMDRNet [1] also consider the semantic gaps between encoders and decoders and the modality discrepancies between multi-modal features, respectively.

Alternatively, in this paper, a novel strategy of bridging-then-fusing is presented to capture the cross-modal features, where the unimodal feature extraction and modality discrepancy reduction are achieved by an MDR+ subnetwork and then the discriminative multi-modal features are adaptively selected and fused by several CWF modules.

III. Our Model

In this section, we describe the overall architecture of the proposed MDRNet+, as illustrated in Fig. 4. Specifically, the proposed model consists of four procedures: 1) Unimodal feature extraction and modality discrepancy reduction by the MDR+ subnetwork; 2) cross-modal feature fusion by several CWF modules; 3) Robust contextual information exploitation by the MSC and MCC modules and 4) Semantic segmentation mask prediction.

A. MDR+ Subnetwork

Although the complementary information within the multi-modal input images can boost the semantic segmentation performance, the modality discrepancies, as discussed in Section I, may hinder the exploitation of cross-modal information. To avoid such a dilemma, we design an MDR+ subnetwork, primarily consisting of four steps (*i.e.*, unimodal feature extraction, cross-modal image translation via adversarial learning, feature extraction of pseudo data and matched real data with the same modality and feature distance minimization), which aims to reduce modality discrepancies via unimodal feature enhancement. It is noted that the steps of MDR+ subnetwork except for the unimodal feature extraction only work during the training phase and do not require any additional computations and parameters during the testing phase.

As shown in the regions marked by the grey boxes of Fig. 4, the MDR+ subnetwork is a bi-directional process, in which two branches (*i.e.*, RGB→TIR and TIR→RGB) are used for the RGB/TIR feature extraction and the modality discrepancy reduction from RGB/TIR data to TIR/RGB data, respectively. Specifically, the unimodal feature extractors are first employed to extract multi-level features from one modality, which are then used to generate the matched pseudo data of another modality. After that, we introduce adversarial learning to constrain the distribution of pseudo data and that of matched real data with the same modality to be as close as possible. Next, the Siamese feature extractors are employed to represent the multi-level features of pseudo data and those of matched real data with the same modality in a common feature space. Moreover, we also perform additional image reconstruction constraints on the Siamese feature extractors to constrain their feature extraction capability. Finally, the modality discrepancies from one modality to another modality are reduced by minimizing the distance between the multi-level features of pseudo data and those of matched real data with the same modality. In the following contents, we will show the details of the branch RGB→TIR that is used for the RGB feature extraction and the modality discrepancy reduction from RGB data to TIR data as an example. The implementation details of MDR+ subnetwork and its advantages over the modality discrepancy reduction subnetwork presented in [1] will also be discussed as follows.

1) Unimodal Feature Extraction: First, given the RGB images $\mathbf{I}^r$ from paired RGB-T images, we choose ResNet-50 [30] as the RGB feature extractor $\mathbf{E}^r$, which contains five residual convolutional blocks. The average pooling layers and the fully connected layers in the original ResNet-50 are removed to maintain more spatial information. Therefore, five levels of RGB features $\{\mathbf{F}_n^r|n{=}1,2,3,4,5\}$ are obtained by $\mathbf{E}^r$, which have the resolutions of 1/2, 1/4, 1/8, 1/16 and 1/32 of the original input sizes, respectively. It is important to note that these RGB features are shared for subsequent cross-modal image translation from RGB images to TIR images and semantic segmentation.

2) Cross-modal Image Translation via Adversarial Learning: For the multi-level RGB features $\{\mathbf{F}_n^r|n{=}1,2,3,4,5\}$ obtained by $\mathbf{E}^r$, an RGB data to TIR data translation network $\mathbf{T}^{r\to t}$ (*i.e.*, the region marked by the red dotted box in Fig. 4) is employed to restore the resolutions of feature maps progressively and generate the pseudo TIR images. Specifically, the translation network $\mathbf{T}^{r\to t}$ consists of two steps: 1) Four Upsample Modules (UMs) with the same structures but different inputs are employed to refine the details and semantics in each level, and 2) A cascaded combination of a 2×2 transposed convolutional layer, a standard 3×3 convolution layer and a Sigmoid activation function is employed to generate pseudo TIR images with one channel. More specifically, the UM consists of a 2×2 transposed convolutional layer, a standard 3×3 convolution layer and a short cut. As a result, the refined RGB features $\{\mathbf{F}_n^{rup}|n{=}1,2,3,4,5\}$ with the same resolutions as their corresponding features in $\{\mathbf{F}_n^r|n{=}1,2,3,4,5\}$ can be obtained by

$$\mathbf{F}_n^{rup}=\begin{cases}\mathrm{Conv}((\mathrm{TConv}(\mathbf{F}_{n+1}^{rup};\alpha_n)+\mathbf{F}_n^r);\beta_n), & n=1,2,3,4\\ \mathbf{F}_5^r, & n=5\end{cases} \tag{1}$$

where $\mathrm{Conv}(*;\alpha_n)$ denotes a 3×3 convolutional layer with its parameters α_n. $\mathrm{TConv}(*;\beta_n)$ denotes a 2×2 transposed convolutional layer with a stride of 2 and β_n denotes its

parameters. Here, the short cut is element-wisely added up. Finally, the output of the first step in the RGB data to TIR data translation network, *i.e.,* $\mathbf{F}_1^{rup}$, is fed into the second step mentioned above to generate the final pseudo TIR images $\mathbf{I}^{pt}$ with one channel, *i.e.,*

$$\mathbf{I}^{pt} = \mathrm{Sigmoid}(\mathrm{Conv}(\mathrm{TConv}(\mathbf{F}_1^{rup};\gamma);\theta)), \tag{2}$$

where $\mathrm{TConv}(*;\gamma)$ denotes a 2×2 transposed convolutional layer with a stride of 2 and γ denotes its parameters. $\mathrm{Conv}(*;\theta)$ denotes another 3×3 convolutional layer with its parameters θ. $\mathrm{Sigmoid}(*)$ denotes the Sigmoid activation function.

In the following content, we will use $\mathbf{G}^{r\to t}$ containing the RGB feature extractor $\mathbf{E}^r$ and the translation network $\mathbf{T}^{r\to t}$ to refer to the cross-modal image translation process from RGB data to TIR data.

Furthermore, in order to ensure that the features of generated pseudo TIR images and those of matched real TIR images can be well represented in a common feature space, we introduce a TIR modality discriminator $\mathbf{D}^t$ to constrain the distribution of generated pseudo TIR images $\mathbf{I}^{pt}$ to be as close as possible to that of matched real TIR images $\mathbf{I}^t$ through adversarial learning. The structure of the TIR modality discriminator is consistent with PatchGAN [48]. Specifically, for the adversarial learning process between $\mathbf{G}^{r\to t}$ and $\mathbf{D}^t$, we first describe the training objective and relevant losses for the TIR modality discriminator. Given the generated pseudo TIR images $\mathbf{I}^{pt}$ and the matched real TIR images $\mathbf{I}^t$, we feed them into $\mathbf{D}^t$ and choose the cross-entropy loss as the TIR modality discriminator loss $\mathcal{L}_d^t$ for the two classes (*i.e.*, real TIR images and pseudo TIR images) to train $\mathbf{D}^t$, *i.e.,*

$$\mathcal{L}_d^t = -\frac{1}{H_1 W_1}\sum_{i=1}^{H_1}\sum_{j=1}^{W_1}(log(\mathbf{D}_{ij}^t(\mathbf{I}^t)) + log(1-(\mathbf{D}_{ij}^t(\mathbf{I}^{pt})))). \tag{3}$$

Here, $\mathbf{D}_{ij}^t(\mathbf{I}^t)$ and $\mathbf{D}_{ij}^t(\mathbf{I}^{pt})$ represent the output of $\mathbf{D}^t$ at position (i,j) with inputs I^t and I^{pt}, respectively. H_1 and W_1 denote the height and width of $\mathbf{D}^t(\mathbf{I}^t)$ or $\mathbf{D}^t(\mathbf{I}^{pt})$.

On the other hand, $\mathbf{G}^{r\to t}$ aims to make the distribution of pseudo TIR images $\mathbf{I}^{pt}$ and that of matched real TIR images $\mathbf{I}^t$ tend to be similar and confuse $\mathbf{D}^t$ by maximizing the probability of generated pseudo TIR images being considered as their matched real TIR images. Therefore, the adversarial loss, *i.e.,* $\mathcal{L}_{adv}^t$, is expressed as:

$$\mathcal{L}_{adv}^t = -\frac{1}{H_1 W_1}\sum_{i=1}^{H_1}\sum_{j=1}^{W_1} log(\mathbf{D}_{ij}^t(\mathbf{I}^{pt})), \tag{4}$$

where $\mathbf{I}^{pt} = \mathbf{G}^{r\to t}(\mathbf{I}^r)$.

3) Feature Extraction of Pseudo Data and Matched Real Data with the Same Modality: Given the generated pseudo TIR images $\mathbf{I}^{pt}$ and the real TIR images $\mathbf{I}^t$, instead of employing two separate feature extractors with the same network structures but different parameters as [1], we employ a Siamese feature extractor $\mathbf{S}^t$ with the same network structures and parameters to simultaneously capture their multi-level features.

Specifically, we still adopt ResNet-50 as the Siamese feature extractor $\mathbf{S}^t$ to extract the multi-level features (*i.e.,* $\{\mathbf{F}_n^{pt}|n{=}1,2,3,4,5\}$ and $\{\mathbf{F}_n^{rt}|n{=}1,2,3,4,5\}$) from $\mathbf{I}^{pt}$ and $\mathbf{I}^t$, respectively. The average pooling layers and the fully connected layers in the original ResNet-50 are also removed.

Moreover, in order to further ensure that the features of pseudo TIR images $\{\mathbf{F}_n^{pt}|n{=}1,2,3,4,5\}$ and those of matched real TIR images $\{\mathbf{F}_n^{rt}|n{=}1,2,3,4,5\}$ contain sufficient effective information, we concisely employ a real TIR image reconstruction constraint to improve the feature representation capability of the Siamese feature extractor $\mathbf{S}^t$. Specifically, we use the high-level features $\mathbf{F}_5^{rt}$ to reconstruct the real TIR images by using five successive 2×2 transposed convolutional layers with strides of 2. Here, the TIR image reconstruction loss $\mathcal{L}_{rec}^t$ employs a binary cross-entropy loss, *i.e.,*

$$\mathcal{L}_{rec}^t = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(\mathbf{I}_{ij}^t log(\mathbf{I}_{ij}^{rt}) + (1-\mathbf{I}_{ij}^t) log(1-\mathbf{I}_{ij}^{rt})), \tag{5}$$

where $\mathbf{I}^{rt}$ denotes the reconstructed TIR images after using the Sigmoid activation function. H and W denote the height and width of $\mathbf{I}^t$. (i,j) represents the position coordinate of a pixel.

4) Feature Distance Minimization: Finally, we minimize the distance between the features of pseudo TIR images (*i.e.,* $\{\mathbf{F}_n^{pt}|n{=}1,2,3,4,5\}$) and those of matched real TIR images (*i.e.,* $\{\mathbf{F}_n^{rt}|n{=}1,2,3,4,5\}$) with a modality discrepancy reduction loss $\mathcal{L}_{mdr}^{r\to t}$, *i.e.,*

$$\mathcal{L}_{mdr}^{r\to t} = \sum_{n=1}^{5}\|\mathbf{F}_n^{rt} - \mathbf{F}_n^{pt}\|_2^2. \tag{6}$$

By minimizing the distance between the features of pseudo TIR images and those of matched real TIR images, the feature mining capability of the RGB feature extractor $\mathbf{E}^r$ can be potentially improved. As a result, the multi-level RGB features $\{\mathbf{F}_n^r|n{=}1,2,3,4,5\}$ will contain some certain discriminative information of TIR modality in addition to the information of their own modality, thus indirectly achieving the information propagation from TIR data to RGB data.

Similarly, for the branch that is used for the TIR feature extraction and the modality discrepancy reduction from TIR data to RGB data, given the TIR images $\mathbf{I}^t$, five levels of TIR features $\{\mathbf{F}_n^t|n{=}1,2,3,4,5\}$ are first obtained by a TIR feature extractor $\mathbf{E}^t$. Here, we also choose ResNet-50 as the TIR feature extractor $\mathbf{E}^t$. These TIR features are also shared for subsequent cross-modal image translation from TIR images to RGB images and semantic segmentation. After that, the refined TIR features $\{\mathbf{F}_n^{tup}|n{=}1,2,3,4,5\}$ and the generated pseudo RGB images $\mathbf{I}^{pr}$ are obtained by a TIR data to RGB data translation network $\mathbf{T}^{t\to r}$ (*i.e.,* the region marked by the green dotted box in Fig. 4) with the same strategy as that in $\mathbf{T}^{r\to t}$. Subsequently, we also introduce an RGB modality discriminator $\mathbf{D}^r$ with the same structure as $\mathbf{D}^t$ to constrain the distribution of generated pseudo RGB images $\mathbf{I}^{pr}$ to be as close as possible to that of matched real RGB images $\mathbf{I}^r$ by adversarial learning. Accordingly, the RGB modality discriminator loss $\mathcal{L}_d^r$ and the corresponding adversarial loss

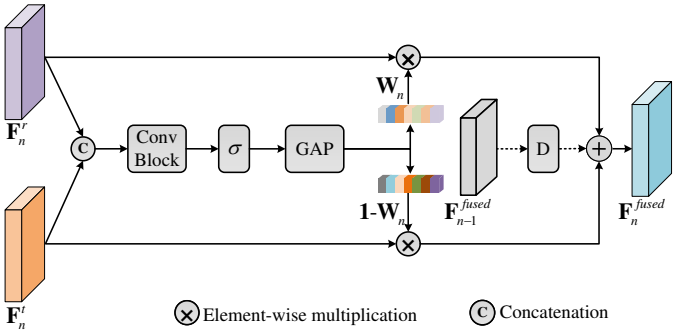

Fig. 5. Structure of the proposed CWF module. The weight vector $\mathbf{W}_n$ is able to weigh the importance of features from the RGB modality along the channel dimension.

i.e., $\mathcal{L}_{adv}^r$, are expressed as:

$$\mathcal{L}_d^r = -\frac{1}{H_1 W_1}\sum_{i=1}^{H_1}\sum_{j=1}^{W_1}(log(\mathbf{D}_{ij}^r(\mathbf{I}^r)) + log(1-(\mathbf{D}_{ij}^r(\mathbf{I}^{pr})))), \tag{7}$$

$$\mathcal{L}_{adv}^r = -\frac{1}{H_1 W_1}\sum_{i=1}^{H_1}\sum_{j=1}^{W_1} log(\mathbf{D}_{ij}^r(\mathbf{I}^{pr})), \tag{8}$$

respectively. Here, $\mathbf{I}^{pr} = \mathbf{G}^{t\to r}(\mathbf{I}^t)$. $\mathbf{D}_{ij}^r(\mathbf{I}^r)$ and $\mathbf{D}_{ij}^r(\mathbf{I}^{pr})$ represent the output of $\mathbf{D}^r$ at position (i,j) with inputs I^r and I^{pr}, respectively. $\mathbf{G}^{t\to r}$ contains the TIR feature extractor $\mathbf{E}^t$ and the translation network $\mathbf{T}^{t\to r}$.

Then, we also employ a ResNet-50 as another Siamese feature extractor $\mathbf{S}^r$ to capture the multi-level features (*i.e.,* $\{\mathbf{F}_n^{pr}|n{=}1,2,3,4,5\}$ and $\{\mathbf{F}_n^{rr}|n{=}1,2,3,4,5\}$) from $\mathbf{I}^{pr}$ and $\mathbf{I}^r$, respectively. Meanwhile, a real RGB image reconstruction constraint is employed to improve the feature representation capability of the Siamese feature extractor $\mathbf{S}^r$. Mathematically, the RGB image reconstruction loss $\mathcal{L}_{rec}^r$ can be expressed by

$$\mathcal{L}_{rec}^r = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(\mathbf{I}_{ij}^r log(\mathbf{I}_{ij}^{rr}) + (1-\mathbf{I}_{ij}^r) log(1-\mathbf{I}_{ij}^{rr})), \tag{9}$$

where $\mathbf{I}^{rr}$ denotes the reconstructed RGB images after using the Sigmoid activation function.

Finally, we minimize the distance between the features of pseudo RGB images (*i.e.,* $\{\mathbf{F}_n^{pr}|n{=}1,2,3,4,5\}$) and those of matched real RGB images (*i.e.,* $\{\mathbf{F}_n^{rr}|n{=}1,2,3,4,5\}$) with another modality discrepancy reduction loss $\mathcal{L}_{mdr}^{t\to r}$, *i.e.*,

$$\mathcal{L}_{mdr}^{t\to r} = \sum_{n=1}^{5}\|\mathbf{F}_n^{rr} - \mathbf{F}_n^{pr}\|_2^2. \tag{10}$$

5) Advantage Analysis: Compared with the modality discrepancy reduction subnetwork designed in the previous version [1], our improved MDR+ subnetwork employs the Siamese feature extractors to extract the features of pseudo data and those of matched real data with the same modality, which can map the pseudo data and the matched real data with the same modality into a common feature space. Before that, an adversarial learning technique is employed to constrain the distribution of pseudo data to be as close as possible to that of matched real data, which can facilitate the Siamese feature extractors to well represent the features of pseudo and real data simultaneously. These two improvements jointly improve the rationality of the distance measurement between the features of pseudo data and those of matched real data with the same modality. Moreover, the introduction of additional image reconstruction constraints can improve the feature representation capability of the Siamese feature extractors, which further ensures the validity and stability of modality discrepancy reduction. The visual and quantitative ablation analysis in Section V will further prove the effectiveness of modality discrepancy reduction and the superiority of our improved modality discrepancy reduction subnetwork over the previous version in [1].

B. CWF

After using MDR+ subnetwork, we obtain the multi-level RGB and TIR features with lower modality discrepancies from the perspective of unimodal feature enhancement. Considering that these unimodal features only contain partial discriminative information of another modality in addition to the information of their own modality, an effective fusion strategy needs to be explored to further exploit the cross-modal complementary information. However, existing RGB-T semantic segmentation methods usually adopt primitive fusion strategies, such as concatenation [3], [22] and element-wise summation [21], [23], [24], which fail to take into account the importance of different channels within multi-modal features. In fact, different channels of features have different class-discriminability, and treating them equally will significantly affect the categorical performance of RGB-T semantic segmentation.

Based on the above analysis, we propose a novel Channel Weighted Fusion (CWF) module to effectively integrate the cross-modal complementary information by reweighting the importance of unimodal features in a channel-dependent way. Specifically, for the obtained multi-level features $\{\mathbf{F}_n^r|n{=}1,2,3,4,5\}$ and $\{\mathbf{F}_n^t|n{=}1,2,3,4,5\}$, we only use the features from the first four levels to reduce the computation complexity of the entire model. Therefore, we will describe the fusion of $\{\mathbf{F}_n^r|n{=}1,2,3,4\}$ and $\{\mathbf{F}_n^t|n{=}1,2,3,4\}$ next. As shown in Fig. 5, $\mathbf{F}_n^r$ and $\mathbf{F}_n^t$ are first concatenated and then fed into two convolutional layers to calculate their relative importance in a channel-wise way. The corresponding importance weight vector $\mathbf{W}_n$ is obtained by

$$\mathbf{W}_n = \mathrm{GAP}(\mathrm{Sigmoid}(\mathrm{BConv}(\mathrm{Cat}(\mathbf{F}_n^r, \mathbf{F}_n^t); \varepsilon))). \tag{11}$$

Here, $\mathrm{BConv}(*;\varepsilon)$ denotes a convolutional block with a $1{\times}1$ convolutional layer and a 3×3 convolutional layer, and ε denotes its parameters. $\mathrm{GAP}(*)$ denotes the global average pooling operation. Intuitively, in our CWF module, $\mathbf{W}_n$ and $\mathbf{1}-\mathbf{W_n}$ are used to represent the importance of RGB features and TIR features along the channel dimension, respectively. Higher values in $\mathbf{W}_n$ indicate that corresponding channels of features in the RGB modality are more likely to contain important information than those corresponding channels of features from the TIR images and vice versa. Here, $\mathbf{1}$ denotes a vector of 1's with the same size of $\mathbf{W}_n$.

After obtaining these channel importance weight vectors $\{\mathbf{W}_n|n{=}1,2,3,4\}$ for the four levels, we fuse $\{\mathbf{F}_n^r|n{=}1,2,3,4\}$ and $\{\mathbf{F}_n^t|n{=}1,2,3,4\}$ in a weighted summation way to select those features with high discriminability from multi-modal data. The fused features $\{\mathbf{F}_n^{fused}|n{=}1,2,3,4\}$ are thus obtained by

$$\mathbf{F}_n^{fused} = \begin{cases} \mathbf{W}_n \odot \mathbf{F}_n^r + (\mathbf{1} - \mathbf{W}_n) \odot \mathbf{F}_n^t, & n = 1 \\ \mathrm{DB}(\mathbf{F}_{n-1}^{fused}) + \mathbf{W}_n \odot \mathbf{F}_n^r + (\mathbf{1} - \mathbf{W}_n) \odot \mathbf{F}_n^t, & n = 2,3,4 \end{cases} \tag{12}$$

where $\odot$ denotes the channel-wise multiplication and $\mathrm{DB}(*)$ denotes a residual block with a stride of 2 for downsampling.

C. MSC and MCC

Contextual information has been proved to be effective for dealing with the diversity of objects in RGB semantic segmentation, but they are still not well exploited in RGB-T semantic segmentation. For that, we propose the MSC and MCC modules, which are performed on $\mathbf{F}_3^{fused}$ and $\mathbf{F}_4^{fused}$, respectively, to capture contextual information by exploiting the interaction between multi-scale information of the cross-modal fused features and their long-range dependencies along the spatial and channel dimensions.

1) MSC: The structure of MSC is shown in Fig. 6. Given the 3-rd level of the fused features $\mathbf{F}_3^{fused} \in \mathbb{R}^{H_2 \times W_2 \times C_2}$, MSC establishes the interaction between the multi-scale information and their long-range dependencies along the spatial dimension to fully mine contextual information.

Specifically, an ASPP-based structure with dilation rates of 1, 6, 12 and 18 is first employed to extract multi-scale information from the input fused features. Then, the features of four scales are concatenated and fed into a 1×1 convolutional layer to reduce their channels, thus obtaining the fused multi-scale features $\mathbf{F}_3^{ms} \in \mathbb{R}^{H_2 \times W_2 \times C_2}$. Subsequently, the multi-scale spatial correlation matrix $\mathbf{M}_{ms} \in \mathbb{R}^{H_2 W_2 \times H_2 W_2}$ of $\mathbf{F}_3^{ms}$ can be computed by a Long-range Dependency Unit (LDU), *i.e.,*

$$\mathbf{M}_{ms} = \mathrm{Norm}(\mathrm{RS}\,(\mathbf{F}_3^{ms}) \cdot (\mathrm{RS}\,(\mathbf{F}_3^{ms}))^{\mathrm{T}}), \tag{13}$$

where $(*)^{\mathrm{T}}$ denotes the matrix transpose and $\mathrm{RS}(*)$ transfers the size of the input matrix from $\mathbb{R}^{H_2 \times W_2 \times C_2}$ to $\mathbb{R}^{H_2 W_2 \times C_2}$. $\mathrm{Norm}(*)$ denotes the Min-Max Normalization operation and $\cdot$ denotes the matrix multiplication. Analytically, the long-range dependencies among the multi-scale features can be regarded as an inter-regional correlation for input features. In addition, the correlations among feature points from input features should also be considered to supplement global contextual information. Therefore, we introduce the original spatial correlation matrix $\mathbf{M}_{os} \in \mathbb{R}^{H_2 W_2 \times H_2 W_2}$, computed by another LDU (*i.e.*, Eq. (14)), and subsequently establish the cooperation between $\mathbf{M}_{ms}$ and $\mathbf{M}_{os}$ by element-wise summation.

$$\mathbf{M}_{os} = \mathrm{Norm}(\mathrm{RS}(\mathbf{F}_3^{fused}) \cdot (\mathrm{RS}(\mathbf{F}_3^{fused}))^{\mathrm{T}}). \tag{14}$$

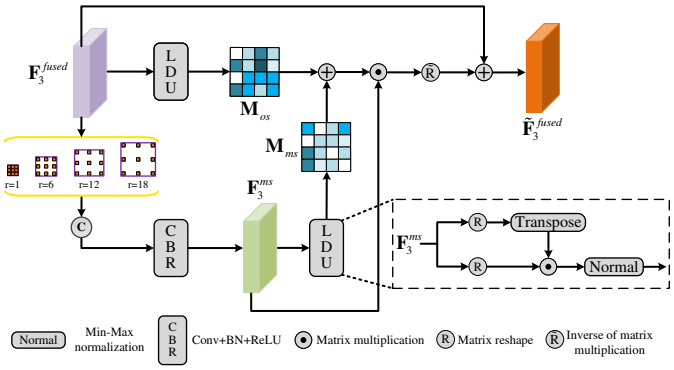

Fig. 6. Structure of the proposed MSC module.

After that, the output features with rich spatial contextual information are obtained by

$$\widetilde{\mathbf{F}}_3^{fused} = \widetilde{\mathrm{RS}}((\mathbf{M}_{ms} + \mathbf{M}_{os}) \cdot \mathrm{RS}(\mathbf{F}_3^{ms})) + \mathbf{F}_3^{fused}, \tag{15}$$

where $\widetilde{\mathrm{RS}}(*)$ denotes the inverse process of $\mathrm{RS}(*)$.

2) MCC: Similarly, given the 4-th level of the fused features $\mathbf{F}_4^{fused} \in \mathbb{R}^{H_3 \times W_3 \times C_3}$, MCC aims to establish the interaction between the multi-scale information and their long-range dependencies along the channel dimension to further mine contextual information.

Specifically, MCC first uses the same method as MSC to obtain the multi-scale features $\mathbf{F}_4^{ms} \in \mathbb{R}^{H_3 \times W_3 \times C_3}$. Then, the multi-scale channel correlation matrix $\mathbf{M}_{mc} \in \mathbb{R}^{C_3 \times C_3}$ and the original channel correlation matrix $\mathbf{M}_{oc} \in \mathbb{R}^{C_3 \times C_3}$ are, respectively, computed by

$$\mathbf{M}_{mc} = \mathrm{Norm}((\mathrm{RS}\,(\mathbf{F}_4^{ms}))^{\mathrm{T}} \cdot (\mathrm{RS}\,(\mathbf{F}_4^{ms}))), \tag{16}$$

$$\mathbf{M}_{oc} = \mathrm{Norm}((\mathrm{RS}(\mathbf{F}_4^{fused}))^{\mathrm{T}} \cdot (\mathrm{RS}(\mathbf{F}_4^{fused}))). \tag{17}$$

Finally, the output features with rich channel contextual information are obtained by

$$\widetilde{\mathbf{F}}_4^{fused} = \widetilde{\mathrm{RS}}(\mathrm{RS}(\mathbf{F}_4^{ms}) \times (\mathbf{M}_{mc} + \mathbf{M}_{oc})) + \mathbf{F}_4^{fused}. \tag{18}$$

3) Analysis of Order: Theoretically, in the semantic segmentation model, lower-level features provide richer detail information with larger resolutions, while higher-level features provide richer semantic information with larger channel numbers. Thus, when performing MSC on the 3-rd level of cross-modal fused features, we can explore richer spatial dependencies and significantly reduce the computation compared to performing it on the 1-st or 2-nd level. While, performing MCC on the 4-th level of features can explore richer channel dependencies, thus capturing more semantic cues of targets. This underlying principle motivates us to apply MSC and MCC to $\mathbf{F}_3^{fused}$ and $\mathbf{F}_4^{fused}$, respectively.

D. Semantic Segmentation Mask Prediction

Given $\widetilde{\mathbf{F}}_4^{fused}$ with rich contextual information, we employ three UMs to progressively recover details and semantics. Finally, a series of operations, involving a $2{\times}2$ transposed convolutional layer, a standard $3{\times}3$ convolution layer and a Softmax activation function, are employed to predict the final semantic segmentation mask.

Fig. 7. Sample examples from *RTSS*, including three groups, *i.e.,* 'GI', 'CI' and 'EI' mentioned in the earlier Section I.

E. Loss Function

Considering the imbalance among pixels of each class, we employ the weighted cross-entropy loss as the semantic segmentation loss, which is defined by

$$\mathcal{L}_{seg} = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{W}(x_{ij})\mathbf{P}\left(x_{ij}\right)\log\left(\mathbf{Q}\left(x_{ij}\right)\right), \quad (19)$$

where $\mathbf{W}(x_{ij})$ represents the coefficient of the category to which the pixel x_{ij} belongs. $\mathbf{P}(x_{ij})$ and $\mathbf{Q}(x_{ij})$ represent the ground truth label and the prediction on this pixel, respectively.

Finally, we adopt a joint loss function $\mathcal{L}_{joint}$ that combines semantic segmentation loss (*i.e.,* $\mathcal{L}_{seg}$), adversarial losses (*i.e.,* $\mathcal{L}^{t}_{adv}$ and $\mathcal{L}^{r}_{adv}$), modality discrepancy reduction losses (*i.e.,* $\mathcal{L}^{r\to t}_{mdr}$ and $\mathcal{L}^{t\to r}_{mdr}$) and image reconstruction losses (*i.e.,* $\mathcal{L}^{t}_{rec}$ and $\mathcal{L}^{r}_{rec}$) to train the entire network, *i.e.,*

$$\begin{aligned}\mathcal{L}_{joint} =& \lambda_0\mathcal{L}_{seg} + \lambda_1(\mathcal{L}^{t}_{adv} + \mathcal{L}^{r}_{adv}) + \lambda_2(\mathcal{L}^{r\to t}_{mdr} + \mathcal{L}^{t\to r}_{mdr}) \\ &+ \lambda_3(\mathcal{L}^{t}_{rec} + \mathcal{L}^{r}_{rec}),\end{aligned} \quad (20)$$

where λ_0, λ_1, λ_2 and λ_3 denote four hyper-parameters for controlling the tradeoff among the loss functions. We finally set λ_0, λ_1, λ_2 and λ_3 to 1, 0.01, 1 and 0.1 in our experiments, respectively.

The ultimate optimization objective becomes the following min-max criterion:

$$\max_{\{\mathbf{D}^t,\mathbf{D}^r\}}\min_{\{\mathbf{G}^{r\to t},\mathbf{G}^{t\to r}\}}\mathcal{L}_{joint}. \quad (21)$$

IV. RTSS Dataset

The lack of data with annotations is one of the major bottlenecks restricting the development of RGB-T semantic segmentation. With this in mind, we elaborately assemble a challenging RGB-T semantic segmentation dataset *RTSS* as a supplement to the autonomous driving field. We will provide the details of *RTSS* in terms of two key aspects, including image collection and category annotation and dataset analysis.

A. Image Collection and Category Annotation

We elaborately select the RGB-T image pairs from some existing datasets for other computer vision tasks, *e.g.,* KAIST [33] for RGB-T pedestrian detection, and LasHeR [34] and RGBT234 [35] for RGB-T tracking. Our proposed *RTSS* dataset contains 1880 RGB-T image pairs under various urban scenes, including 1496 pairs of daytime and 384 pairs of nighttime. In addition to the classification of daytime and nighttime as that in MFNet dataset [3], *RTSS* also divides these RGB-T image pairs into three groups, *i.e.,* 'GI', 'CI' and 'EI' as mentioned in the earlier Section I. Sample examples from *RTSS* are shown in Fig. 7.

Meanwhile, pixel-level category annotations are essential components for training a supervised RGB-T semantic segmentation network. To this end, we manually provide four pixel-level categories of annotations for *RTSS*, including car, human, non-motor and background. This pixel-level annotation costs 10 minutes per image pair on average.

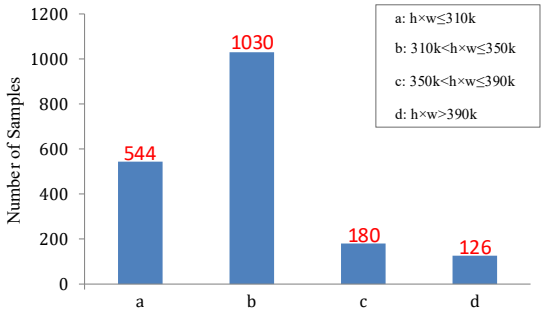

Fig. 8. Image resolution distribution of *RTSS* dataset. We divide the images into four resolution ranges (*i.e.,* $h \times w \leq 310k$, $310k < h \times w \leq 350k$, $350k < h \times w \leq 390k$ and $h \times w > 390k$) for statistics.

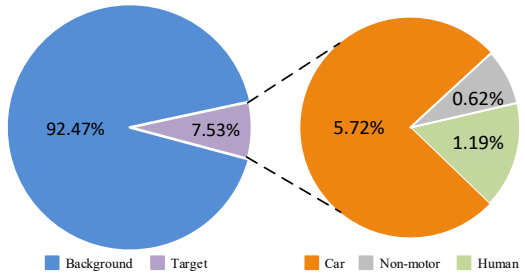

Fig. 9. Percentage of the number of pixels for different classes in *RTSS* dataset.

B. Dataset Analysis

1) Resolution Distribution: When we select RGB-T image pairs from existing datasets for other computer vision tasks, we keep their original resolutions unchanged. Fig. 8 displays the resolution distribution of *RTSS*, in which we divide them into four resolution ranges (*i.e.,* $h \times w \leq 310k$, $310k < h \times w \leq 350k$, $350k < h \times w \leq 390k$ and $h \times w > 390k$) for statistics. In *RTSS*, scenarios are more complex and target sizes are more diverse, which is more challenging than [3].

2) Category Label Analysis: *RTSS* contains four classes, including one background class and three target classes (*i.e.,* human, non-motor and car), which focus on the most important objects (usually having higher temperature values) in urban scenes for autonomous driving and traffic dispersion. Fig. 9 presents the label distribution of different categories in *RTSS*. In our dataset, only a few target categories in each image are labelled, so the background pixels occupy the majority, which

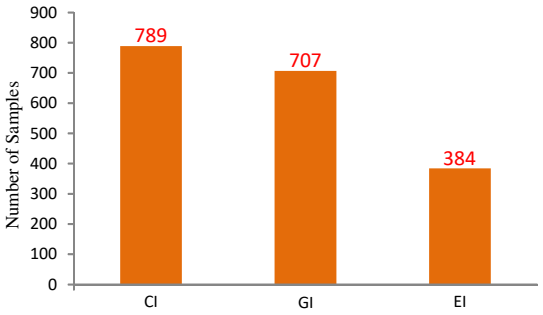

Fig. 10. Statistics of sample sets (*i.e.,* 'GI', 'CI' and 'EI') in *RTSS* dataset.

is similar to that in MFNet [3]. It is worth mentioning that the intuitive reason why we choose the three categories of targets for labelling is that, in urban scenes, their temperatures are usually significantly in contrast to those of their surrounding environments. For these targets, TIR data can provide more discriminative information.

3) Sample Set Division: The RGB-T image pairs in *RTSS* are manually divided into three different sample sets, including 'GI', 'CI' and 'EI'. Fig. 10 shows the statistics of sample sets in *RTSS*. Compared with the existing RGB-T semantic segmentation dataset [3] for urban scene parsing, *RTSS* contains more samples that need to effectively combine the information from the two modalities for accurate segmentation, which is beneficial to improve the exploitation ability of cross-modal information for an RGB-T semantic segmentation model.

V. Experiment and Analysis

A. Datasets

We evaluate the superiority of our model on three RGB-T semantic segmentation datasets, including two public datasets (*i.e.*, MFNet [3] and PST900 [22] datasets) and our proposed *RTSS* dataset. Details of them are described as follows.

MFNet [3] contains 1569 RGB and TIR image pairs with 9 classes of pixel-level labels from urban scenes, in which 820 image pairs are taken at daytime and 749 image pairs are taken at nighttime. All of the images in MFNet dataset [3] have the same resolution of 480 $\times$ 640. PST900 contains 894 aligned RGB and TIR image pairs with 5 classes of pixel-level labels. Details of *RTSS* are given in Section IV. We divide our proposed *RTSS* dataset into the training set and the testing set. The training set includes 75% of images and the testing set contains 25% of the images in the entire dataset, respectively.

B. Evaluation Criteria

We evaluate the performance of our model and other state-of-the-art methods from both visual and quantitative perspectives. For the quantitative experiments, we adopt two widely used evaluation metrics, *i.e.,* mean Accuracy per class (mAcc) and mean Intersection over Union per class (mIoU), to evaluate the semantic segmentation performance of different models. They can be computed by

$$\mathrm{mAcc} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{m=1}^{M}\eta_{ii}^{m}}{\sum_{m=1}^{M}\eta_{ii}^{m}+\sum_{m=1}^{M}\sum_{j=1,j\neq i}^{N}\eta_{ij}^{m}}, \tag{22}$$

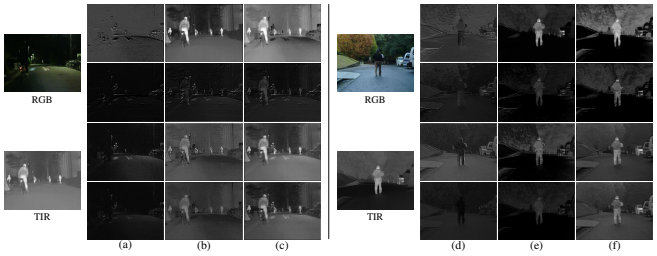

Fig. 11. Visual comparisons of the cross-modal fused features with or without reducing modality discrepancies. (a) and (d) denote the cross-modal fused features obtained by summation fusion; (b) and (e) denote the cross-modal fused features, which are obtained by summation fusion after reducing the modality discrepancies between unimodal features by the modality discrepancy reduction subnetwork presented in ABMDRNet [1]; (c) and (f) denote the cross-modal fused features, which are obtained by summation fusion after reducing the modality discrepancies between unimodal features by the improved MDR+ subnetwork proposed in our MDRNet+.

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{m=1}^{M}\eta_{ii}^{m}}{\sum_{m=1}^{M}\eta_{ii}^{m}+\sum_{m=1}^{M}\sum_{j=1,j\neq i}^{N}\eta_{ji}^{m}+\sum_{m=1}^{M}\sum_{j=1,j\neq i}^{N}\eta_{ij}^{m}}, \tag{23}$$

respectively. Here, η_{ii}^{m}, η_{ij}^{m} and η_{ji}^{m} are the number of pixels of class i that are correctly classified as class i, the number of pixels of class i that are wrongly classified as class j and the number of pixels of class j that are wrongly classified as class i in the m-th image, respectively. M denotes the number of images. N represents the number of classes, which is 9 in MFNet dataset [3] and 4 in *RTSS* dataset.

C. Implementation Details

The proposed network is implemented by Python 3.7.1 + PyTorch 1.7.1 + Torchvision 0.8.2 on an NVIDIA GTX 3090 Ti GPU. The stochastic gradient descent (SGD) optimization algorithm 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 0.001, which is decreased by the 'poly' policy during training. Moreover, the training data are augmented by using random flipping, cropping and noise injecting techniques. We train the network about 500 epochs to its convergence. Especially, for the training on the *RTSS* dataset, we resize the input resolutions to 480$\times$640 to be consistent.

D. Ablation Analysis

In this section, we validate the effectiveness of each component in our proposed model on the MFNet dataset [3]. The proposed MDR+ subnetwork, CWF, MSC and MCC modules are first removed from our model as the baseline (denoted by 'BS').

1) MDR+ Subnetwork: We first verify the effectiveness of modality discrepancy reduction. Moreover, to further explore the superiority of our improved MDR+ subnetwork, we provide visual and quantitative comparisons with the previous version of the modality discrepancy reduction subnetwork proposed in ABMDRNet [1]. Finally, we also verify the effectiveness of each component in MDRNet+ subnetwork. Here, 'MDR†' and 'MDR' denote the MDR+ subnetwork

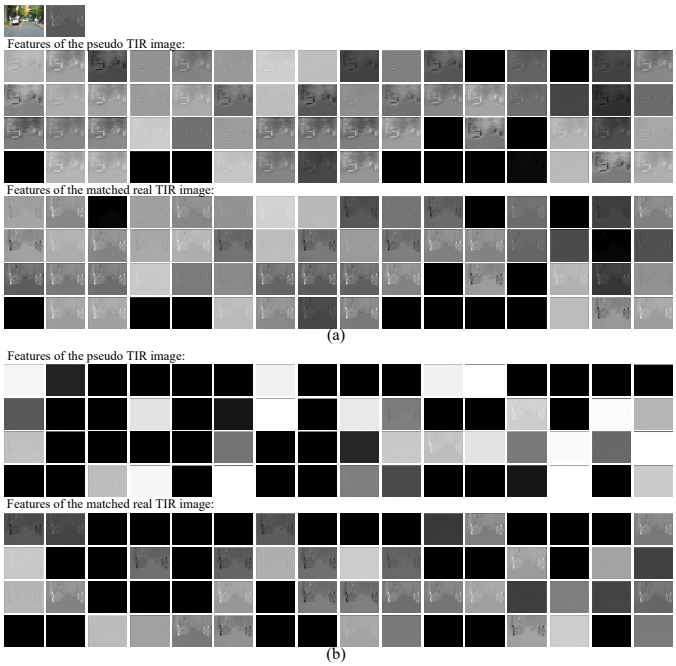

Fig. 12. For a given RGB-T image pair (shown in the top row), the shallowest layers of pseudo and real TIR features obtained by our improved MDR+ subnetwork and those obtained by the previous version of the modality discrepancy reduction subnetwork in [1] are shown in (a) and (b), respectively.

and the previous modality discrepancy reduction subnetwork proposed in [1], respectively. 'adv' and 'recon' denote the adversarial learning and the image reconstruction constraints used in MDR+ subnetwork, respectively.

The quantitative and visual experimental results are shown in Table I and Fig. 11, respectively. The results of 'BS', 'BS+MDR' and 'BS+MDR†' indicate that reducing the modality discrepancies between multi-modal features benefits the exploitation of cross-modal complementary information, thus boosting the performance of RGB-T semantic segmentation greatly. Meanwhile, by comparing Fig. 11 (a) (or Fig. 11 (d)) with Fig. 11 (b) (or Fig. 11 (e)) and Fig. 11 (c) (or Fig. 11 (f)), it can be observed that the cross-modal fused features obtained by summation after reducing modality discrepancies are more discriminative than those obtained by summing directly. Furthermore, from the results of 'BS+MDR' and 'BS+MDR†' in Table I, it can be observed that the improved modality discrepancy reduction subnetwork MDR+ proposed in MDRNet+ shows better performance than the original modality discrepancy reduction subnetwork presented in ABMDRNet [1], which benefits from our constraints on the distribution of pseudo data as well as on the extraction of the features from pseudo data and those from the matched real data with the same modality. In addition, the visual comparisons between Fig. 11 (b) (or Fig. 11 (e)) and Fig. 11 (c) (or Fig. 11 (f)) also verify the advantages of MDR+ subnetwork in cross-modal complementary information exploitation. More intuitively, the visual comparisons between the features of pseudo and real TIR images obtained by the previous version of the modality discrepancy reduction subnetwork in [1] and those obtained by our MDR+ subnetwork is taken as an example, as shown in Fig. 12. Owing to our improvements for reducing modality discrepancy, more channels of features from pseudo and real TIR images contain effective information. This will further facilitate the RGB feature extractor $\mathbf{E}^r$ to capture more discriminative TIR information in addition to the information of their own RGB modality, thus achieving a more effective and more stable modality discrepancy reduction from the perspective of unimodal feature enhancement.

TABLE I
QUANTITATIVE RESULTS (%) OF ABLATION ANALYSIS FOR THE IMPROVED MODALITY DISCREPANCY REDUCTION SUBNETWORK.

Methods	mAcc	mIoU
BS	57.30	47.99
BS+MDR	62.37	51.98
BS+MDR†	66.90	53.44
BS+MDR†+w/o adv+w/o recon	64.92	50.34
BS+MDR†+w/o adv	66.84	52.18
BS+MDR†+w/o recon	65.18	51.25

TABLE II
QUANTITATIVE RESULTS (%) OF ABLATION ANALYSIS FOR THE CWF MODULE EMPLOYED IN THE FUSION STAGE.

Methods	mAcc	mIoU
BS+MDR†+Sum	66.90	53.44
BS+MDR†+Concat	66.58	52.95
BS+MDR†+CW	67.25	53.67
BS+MDR†+CWF	67.11	54.06

Furthermore, in order to further verify the contributions of each component in MDR+ subnetwork, we also conduct ablation analysis on the adversarial learning technique and image reconstruction constraints used in MDR+ subnetwork. As shown in the 4th-6th rows of Table I, it can be observed from the results among 'BS+MDR†+w/o adv+w/o recon', 'BS+MDR†+w/o adv' and 'BS+MDR†+w/o recon' that the introduction of adversarial learning for constraining the distribution of pseudo data and image reconstruction constraints for constraining the effectiveness of the features of pseudo and real images can both facilitate modality discrepancy reduction.

2) CWF: To explore the superiority of our proposed CWF module, we fully compare it with some existing RGB-T feature fusion strategies (*e.g.*, 'Sum' [21], [23], [24], 'Concat' [3], [22] and 'CW' [49]). The quantitative experimental results are shown in Table II. It can be observed that, compared with other fusion strategies (*e.g.,* 'BS+MDR†+Sum', 'BS+MDR†+Concat', BS+MDR†+CW), our proposed CWF module can more effectively select those discriminative information from multi-modal features, which can further facilitate the exploitation of cross-modal information for RGB-T semantic segmentation.

3) MSC and MCC: **MSC:** We verify the effectiveness of MSC module by embedding it into the baseline model 'BS' mentioned in Table I and the model 'BS+MDR†+CWF' men-

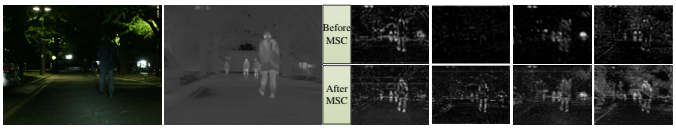

Fig. 13. Visual comparisons of the cross-modal fused features before and after employing MSC.

TABLE III
QUANTITATIVE RESULTS (%) OF ABLATION ANALYSIS FOR MSC AND MCC MODULES. THE BOLD FONT HIGHLIGHTS THE BEST RESULTS.

Methods	mAcc	mIoU
BS+MSC	65.75	52.86
BS+MDR†+CWF+MSC	68.61	54.01
BS+MCC	69.19	53.66
BS+MDR†+CWF+MCC	72.02	55.00
BS+MSC+MCC	70.87	54.93
BS+MDR†+CWF+MSC+MCC (MDRNet+)	**74.65**	**56.78**

TABLE IV
QUANTITATIVE RESULTS (%) OF DIFFERENT HYPER-PARAMETERS OF LOSS FUNCTIONS. THE BOLD FONT HIGHLIGHTS THE BEST RESULTS.

λ_0:λ_1:λ_2:λ_3	mIoU	λ_0:λ_1:λ_2:λ_3	mIoU	λ_0:λ_1:λ_2:λ_3	mIoU
1:0:0:0	55.70	1:0.01:1:0	56.23	1:0.01:1:0.01	56.29
1:0:0.2:0	55.94	1:0.02:1:0	56.20	1:0.01:1:0.02	56.45
1:0:0.4:0	55.95	1:0.04:1:0	56.09	1:0.01:1:0.04	56.50
1:0:0.8:0	56.01	1:0.08:1:0	56.18	1:0.01:1:0.08	56.59
1:0:1:0	56.08	1:0.1:1:0	56.18	1:0.01:1:0.1	**56.80**

tioned in Table II, respectively, thus obtaining 'BS+MSC' and 'BS+MDR†+CWF+MSC'. As shown in Table III, the results of 'BS+MSC' and 'BS+MDR†+CWF+MSC' indicate that the introduction of the multi-scale information of cross-modal fused features and their long-range dependencies along the spatial dimension may provide more crucial spatial contextual information for semantic segmentation. Fig. 13 illustrates the visual comparisons of cross-modal fused features before and after employing MSC, which indicates the discriminability of the cross-modal fused features can be greatly boosted by introducing those contextual information captured by MSC.

MCC: To validate the effectiveness of MCC, we embed it into the baseline model 'BS' mentioned in Table I and the model 'BS+MDR†+CWF' mentioned in Table II, respectively. Obviously, the results of 'BS+MCC' and 'BS+MDR†+CWF+MCC' in Table III verified that the semantic contextual information obtained by introducing the multi-scale information of cross-modal fused features and their long-range dependencies along the channel dimension can further boost the RGB-T semantic segmentation performance. Fig. 14 provides the visual comparisons of cross-modal fused features before and after employing MCC module. It can also be observed that more discriminative contextual information containing rich target semantics is effectively mined by MCC.

Finally, we also embed the MSC and MCC modules into the baseline model 'BS' mentioned in Table I and the model 'BS+MDR†+CWF' mentioned in Table II. As shown in the 5th-6th rows of Table III, the results of 'BS+MSC+MCC' and 'BS+MDR†+CWF+MSC+MCC (MDRNet+)' indicate that the joint mining of multi-scale information of cross-modal fused features and their long-range dependencies along the spatial and channel dimensions can further facilitate the exploitation of contextual information for RGB-T semantic segmentation.

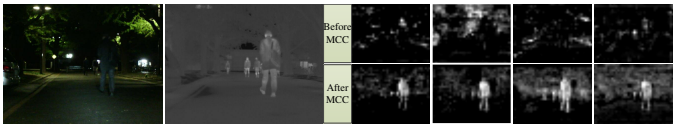

Fig. 14. Visual comparisons of the cross-modal fused features before and after employing MCC.

4) Hyper-Parameter Settings: In this section, we conduct extensive experiments to determine the optimal hyper-parameters to control the tradeoff among the loss functions during the training phase, *i.e.,* λ_0 for the semantic segmentation loss $\mathcal{L}_{seg}$, λ_1 for the adversarial losses $\mathcal{L}_{adv}^{t}$ and $\mathcal{L}_{adv}^{r}$, λ_2 for the modality discrepancy reduction losses $\mathcal{L}_{mdr}^{r\rightarrow t}$ and $\mathcal{L}_{mdr}^{t\rightarrow r}$, and λ_3 for the image reconstruction losses $\mathcal{L}_{rec}^{t}$ and $\mathcal{L}_{rec}^{r}$. According to the results from Table IV, we finally set λ_0, λ_1, λ_2 and λ_3 to 1, 0.01, 1 and 0.1, respectively, in our proposed model for better performance.

E. Comparison with the State-of-the-Art Methods

In this section, we compare our model with 16 state-of-the-art (SOTA) methods, including 3 deep learning based RGB semantic segmentation methods (*i.e.,* DUC [12], DANet [16] and HRNet [13]), 10 RGB-T semantic segmentation approaches (*i.e.,* MFNet [3], RTFNet-50/152 [21], PSTNet [22], MLFNet [23], MMNet [26], AFNet [25], FuseSeg-50/161 [24], ABMDRNet [1]), FEANet [28], EGFNet [29] and 3 RGB-D semantic segmentation models (*i.e.,* LDFNet [42], ACNet [43] and SA-Gate [45]). Visual and quantitative comparisons are both taken into account for comprehensive comparisons.

1) Evaluation on the MFNet Dataset: To evaluate the effectiveness of the proposed MDRNet+, we provide comprehensive comparisons between our model and other existing methods on the MFNet dataset [3]. The quantitative results are shown in Table V, which demonstrates that our method outperforms other SOTA methods by a large margin on the MFNet dataset [3]. This indicates that our method can well exploit the complementary information from RGB-T image pairs by using the MDR+ subnetwork and CWF module and fully mine the contextual information by the proposed MSC and MCC modules, thus significantly facilitating RGB-T semantic segmentation. In addition, we are especially excited to obtain a noticeable gain (*i.e.,*+5.2% and +2.0% in terms of mAcc and mIoU, respectively) over the previous version ABMDRNet [1], which benefits from our improvements for modality discrepancy reduction.

To further explain the superiority of our proposed model, Fig. 15 provides some visual comparisons of different models. As shown in the 1st-3rd rows of Fig. 15, our proposed method achieves significant superiorities over other SOTA models under poor lighting conditions. This owes to the exploitation of cross-modal information by the proposed MDR+ subnetwork and CWF module. In addition, as shown in the 1st-6th rows of Fig. 15, our method still outperforms other SOTA models under complex scenes with multiple objects. This may benefit from the robust contextual information of cross-modal fused features and their long-range dependencies along the spatial and channel dimensions mined by our proposed MSC and MCC modules.

2) Evaluation on the PST900 Dataset: We also report the quantitative experimental results of our model and other

TABLE V
QUANTITATIVE RESULTS (%) OF DIFFERENT MODELS ON THE MFNET DATASET [3]. THE VALUE 0.0 REPRESENTS THAT THERE ARE NO TRUE POSITIVES AND '—' DENOTES THAT THE CORRESPONDING RESULTS ARE UNPUBLISHED. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN AND BLUE, RESPECTIVELY.

Methods	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		
DUC [12]	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 [16]	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 [13]	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 [42]	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 [43]	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 [45]	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 [3]	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-50 [21]	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
RTFNet-152 [21]	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
PSTNet [22]	—	76.8	—	52.6	—	55.3	—	29.6	—	25.1	—	15.1	—	39.4	—	45.0	—	48.4
MLFNet [23]	—	82.3	—	68.1	—	67.3	—	27.3	—	30.4	—	15.7	—	55.6	—	40.1	—	53.8
MMNet [26]	—	83.9	—	69.3	—	59.0	—	43.2	—	24.7	—	4.6	—	42.2	—	50.7	62.7	52.8
AFNet [25]	91.2	86.0	76.3	67.4	72.8	62.0	49.8	43.0	35.3	28.9	24.5	4.6	50.1	44.9	61.0	56.6	62.2	54.6
FuseSeg-50 [24]	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	65.8	53.1
FuseSeg-161 [24]	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
ABMDRNet [1]	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
FEANet [28]	93.3	87.8	82.7	71.1	76.7	61.1	65.5	46.5	26.6	22.1	70.8	6.6	66.6	55.3	77.3	48.9	73.2	55.3
EGFNet [29]	95.8	87.6	89.0	69.8	80.6	58.8	71.5	42.8	48.7	33.8	33.6	7.0	65.3	48.3	71.1	47.1	72.7	54.8
Ours	95.2	87.1	92.5	69.8	76.2	60.9	72.0	47.8	42.3	34.2	66.8	8.2	64.8	50.2	63.5	55.0	74.7	56.8

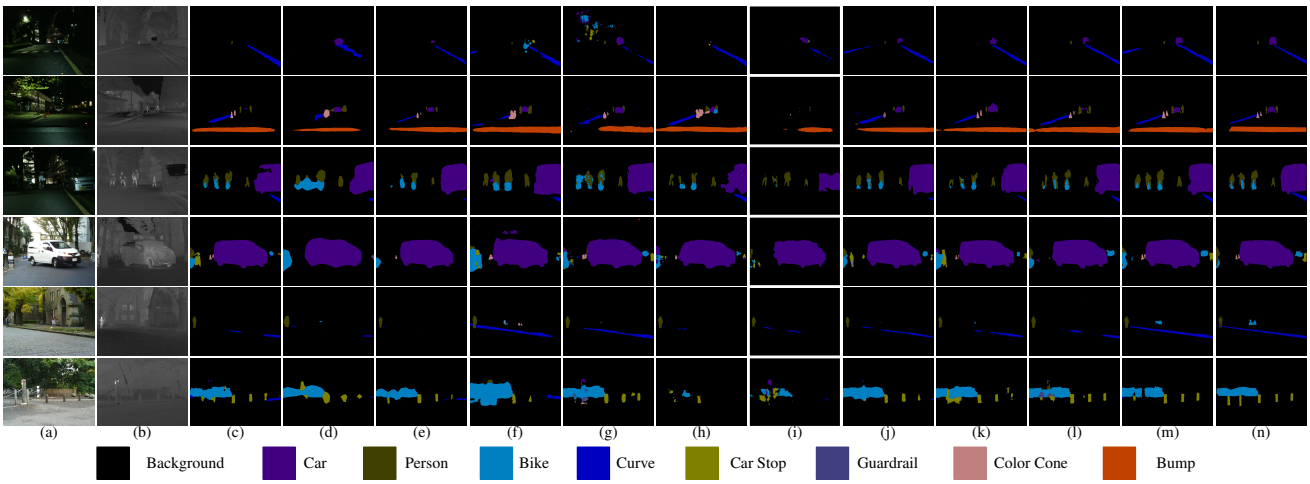

Fig. 15. Visual comparisons of different methods. (a) RGB images; (b) TIR images; (c) DUC [12]; (d) DANet [16]; (e) HRNet [13]; (f) LDFNet [42]; (g) ACNet [43]; (h) SA-Gate [45]; (i) MFNet [3]; (j) RTFNet [21]; (k) MMNet [26]; (l) ABMDRNet [1]; (m) Ours; (n) Ground truth.

TABLE VI
QUANTITATIVE RESULTS OF DIFFERENT MODELS (%) ON THE PST900 DATASET. '—' DENOTES THAT THE CORRESPONDING RESULTS ARE UNPUBLISHED. THE BOLD FONT HIGHLIGHTS THE BEST RESULTS.

Methods	Background		Fire-Extinguisher		Backpack		Hand-Drill		Survivor		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
MFNet [3]	98.89	97.85	68.59	4.87	77.49	63.02	57.15	37.31	15.40	8.63	63.50	50.34
RTFNet-50 [21]	99.74	98.89	69.78	54.46	73.08	67.91	64.07	52.24	58.19	54.11	72.97	65.52
PSTNet [22]	—	98.85	—	70.12	—	69.20	—	53.60	—	50.03	—	68.36
ABMDRNet [1]	99.78	99.00	76.49	66.22	71.41	67.91	81.24	61.52	66.40	62.02	79.06	71.33
Ours	99.30	99.07	90.17	63.04	93.53	76.27	86.63	63.47	85.56	71.26	**91.04**	**74.62**

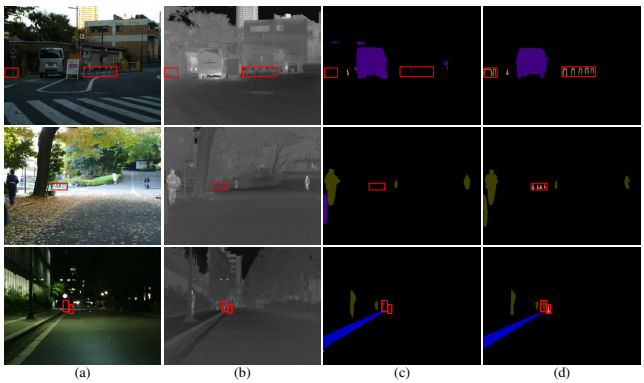

Fig. 16. Some failure cases of our proposed method. (a) RGB images; (b) TIR images; (c) Ours; (d) Ground truth.

TABLE VII
QUANTITATIVE RESULTS (%) OF DIFFERENT MODELS ON THE *RTSS* DATASET. THE BOLD FONT HIGHLIGHTS THE BEST RESULTS.

Methods	Background		Human		Car		Non-motor		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
MFNet [3]	99.4	97.9	56.2	47.5	84.9	78.7	55.9	47.9	74.1	68.0
RTFNet-50 [21]	99.4	98.0	46.0	39.2	91.0	84.6	50.1	44.1	71.6	66.5
PSTNet [22]	99.1	97.9	56.3	44.6	88.2	79.5	64.2	49.9	76.9	68.0
ABMDRNet [1]	99.4	98.5	66.9	54.9	91.7	85.5	65.4	56.9	80.8	74.0
Ours	99.5	98.5	64.6	56.0	91.2	85.7	73.8	61.0	**82.3**	**75.3**

RGB-T semantic segmentation models on the PST900 dataset. As shown in Table VI, our proposed MDRNet+ still out-

TABLE VIII
COMPLEXITIES AND INFERENCE SPEEDS OF DIFFERENT MODELS ON AN GTX 1080TI GPU FOR AN RGB-T IMAGE PAIR OF SIZE 480×640. '—' DENOTES THAT THE CORRESPONDING RESULTS ARE UNPUBLISHED.

Methods	Backbone	FLOPs/G	Params/M	Size/M	FPS
DUC [12]	ResNet-50	211.22	246.06	939.30	62.83
DANet [16]	ResNet-50	144.08	359.57	1372.29	63.82
HRNet [13]	DCNN	222.58	132.72	507.61	11.87
LDFNet [42]	ERFNet	27.69	2.41	9.33	74.58
ACNet [43]	ResNet-50	123.81	116.60	445.70	36.87
SA-Gate [45]	ResNet-50	193.25	110.85	198.38	54.36
MFNet [3]	DCNN	8.39	0.72	3.01	229.86
RTFNet [21]	ResNet-152	337.04	254.51	1024.00	34.07
AFNet [25]	DeepLab	—	—	—	79.30
FuseSeg [24]	DenseNet-161	193.40	141.52	568.30	30.01
ABMDRNet [1]	ResNet-50	194.33	64.60	901.00	40.14
Ours	ResNet-50	194.33	64.60	901.00	40.14

performs other models. This demonstrates the effectiveness of our proposed model on different datasets. What's more, compared with the previous version ABMDRNet [1], our model improves the semantic segmentation performance by 11.98% and 3.29% in terms of mAcc and mIoU, respectively, which further verifies the validity of our improvements for reducing modality discrepancies.

3) Evaluation on the RTSS Dataset: To further evaluate the effectiveness of the proposed MDRNet+, we also compare our model with some RGB-T semantic segmentation methods on our proposed *RTSS* dataset. The quantitative experimental results are shown in Table VII. Our model is also superior to other SOTA methods remarkably. This may result from the fact that scenarios in the challenging *RTSS* dataset have higher dependencies on multi-modal data and the strategy of bridging-then-fusing can better facilitate the exploitation of those discriminative cross-modal complementary information. Notably, the mAcc and mIoU value improvements (*i.e.,* mAcc from 80.8% to 82.3% and mIoU from 74.0% to 75.3%) of MDRNet+ over the previous version ABMDRNet [1] also demonstrate that the improved MDR+ subnetwork can better reduce modality discrepancies, thus improving semantic segmentation performance.

4) Complexities and Inference Speed Analysis: We measure the inference speed of our proposed MDRNet+ and other SOTA models on an NVIDIA GeForce GTX 1080Ti GPU. Furthermore, the number of parameters (Params), the FLoating point OPerations (FLOPs) and the memory sizes of these RGB-T semantic segmentation models are also reported. As shown in Table VIII, our model achieves competitive inference speeds, *i.e.*, 40.14 FPS, with other models. Noticeably, the inference speed, Params, FLOPs and memory sizes of our method is exactly the same to those of the previous version ABMDRNet [1], because the modality discrepancy reduction only occurs in the training phase and no additional parameters and computational costs are introduced in the testing phase.

F. Failure Cases

In this section, we report some failure cases for our proposed model. As shown in the red boxes of Fig. 16, some tiny objects, which usually occupy less than 1000 pixels in the scenes, are difficult to be precisely perceived by our model. Even though the proposed MSC and MCC modules can deal with objects with different sizes to some extent, the deepest layers of features extracted by the hierarchically structured CNN encoders almost lose the full details and semantics of those tiny objects. As a result, those tiny objects may not be well perceived when just using such features of the deepest layers to predict semantic segmentation masks. We will study this issue in the future work, which may be solved by strengthening the exploitation of multi-level features.

VI. CONCLUSION

In this paper, a novel MDRNet+ has been presented for RGB-T semantic segmentation, where the modality discrepancy reduction, cross-modal feature fusion and contextual information mining are simultaneously considered. By virtue of the improved MDR+ subnetwork, the discrepancies between RGB data and TIR data can be greatly reduced in a stable way. On top of that, the discriminative multi-modal features can be adaptively selected and fused by the proposed CWF module. As a result, our bridging-then-fusing strategy can obtain higher discriminative cross-modal 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 contextual information can be well exploited by exploring the interaction between multi-scale information of cross-modal fused features and their long-range dependencies along the spatial and channel dimensions. Thanks to that, the issue of objects diversity in semantic segmentation can be addressed to a large extent. In addition, we also provide a challenging RGB-T semantic segmentation dataset *RTSS* with pixel-level labels on four categories for urban scene understanding. Extensive evaluations have verified the effectiveness and superiorities of our proposed RGB-T semantic segmentation model. With the collaboration of these subnetwork and modules, our proposed RGB-T semantic segmentation model achieves competitive results with other SOTA models on the MFNet, PST900 and *RTSS* datasets.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant No.61773301 and by the Shaanxi Innovation Team Project 2018TD-012.

REFERENCES

[1] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2633–2642.

[2] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2547–2560, 2020.

[3] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE Int. Conf. Intell. Rob. Syst.* IEEE, 2017, pp. 5108–5115.

[4] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3069–3082, 2020.

[5] A. Milioto, P. Lottes, and C. Stachniss, "Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns," in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2018, pp. 2229–2235.

[6] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, 2020.

[7] H. Du, J. Wang, M. Liu, Y. Wang, and E. Meijering, "Swinpa-net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[8] R. Wang, C. Ji, Y. Zhang, and Y. Li, "Focus, fusion, and rectify: Context-aware learning for covid-19 lung infection segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 12–24, 2021.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[10] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[11] Z. Li, Y. Sun, L. Zhang, and J. Tang, "Ctnet: Context-based tandem network for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[12] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Winter Conference on Applications of Computer Vision*. IEEE, 2018, pp. 1451–1460.

[13] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[15] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[16] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[17] X. Ding, C. Shen, T. Zeng, and Y. Peng, "Sab net: A semantic attention boosting framework for semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[18] M. Shi, J. Shen, Q. Yi, J. Weng, Z. Huang, A. Luo, and Y. Zhou, "Lmffnet: A well-balanced lightweight network for fast and accurate semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[20] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[21] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, 2019.

[22] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "Pst900: Rgb-thermal calibration, dataset and segmentation network," in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2020, pp. 9441–9447.

[23] Z. Guo, X. Li, Q. Xu, and Z. Sun, "Robust semantic segmentation based on rgb-thermal in variable lighting scenes," *Measurement*, vol. 186, p. 110176, 2021.

[24] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "Fuseseg: semantic segmentation of urban scenes based on rgb and thermal data fusion," *IEEE Trans. Autom. Sci. Eng.*, 2020.

[25] J. Xu, K. Lu, and H. Wang, "Attention fusion network for multi-spectral semantic segmentation," *Pattern Recognit. Lett.*, vol. 146, pp. 179–184, 2021.

[26] X. Lan, X. Gu, and X. Gu, "Mmnet: Multi-modal multi-stage network for rgb-t image semantic segmentation," *Appl. Intell.*, pp. 1–13, 2021.

[27] S. Zhao and Q. Zhang, "A feature divide-and-conquer network for rgb-t semantic segmentation," *IEEE Trans. Circuit Syst. Video Technol.*, 2022.

[28] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, and T. L. Lam, "Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation," in *Proc. IEEE Int. Conf. Intell. Rob. Syst.* IEEE, 2021, pp. 4467–4473.

[29] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for rgb–thermal scene parsing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 3571–3579.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[31] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1. IEEE, 2005, pp. 539–546.

[32] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, "Few-shot image recognition with knowledge transfer," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 441–449.

[33] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1037–1045.

[34] C. Li, W. Xue, Y. Jia, Z. Qu, B. Luo, J. Tang, and D. Sun, "Lasher: A large-scale high-diversity benchmark for rgbt tracking," *IEEE Trans. Image Process.*, 2021.

[35] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "Rgb-t object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, p. 106977, 2019.

[36] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[37] J.-L. Starck, M. Elad, and D. L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1570–1582, 2005.

[38] N. Paluru, A. Dayal, H. B. Jenssen, T. Sakinis, L. R. Cenkeramaddi, J. Prakash, and P. K. Yalavarthy, "Anam-net: Anamorphic depth embedding-based lightweight cnn for segmentation of anomalies in covid-19 chest ct images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 3, pp. 932–946, 2021.

[39] P. Zhang, J. Zhao, D. Wang, H. Lu, and X. Ruan, "Visible-thermal uav tracking: A large-scale benchmark and new baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8886–8895.

[40] Q. Liu, X. Li, Z. He, C. Li, J. Li, Z. Zhou, D. Yuan, J. Li, K. Yang, N. Fan *et al.*, "Lsotb-tir: A large-scale high-diversity thermal infrared object tracking benchmark," in *ACM Int. Conf. Multimedia*, 2020, pp. 3847–3856.

[41] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Proc. Asian Conf. Comput. Vis.* Springer, 2016, pp. 213–228.

[42] S.-W. Hung, S.-Y. Lo, and H.-M. Hang, "Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation," in *IEEE Int. Conf. Image Process.* IEEE, 2019, pp. 2374–2378.

[43] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *IEEE Int. Conf. Image Process.* IEEE, 2019, pp. 1440–1444.

[44] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 135–150.

[45] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 561–577.

[46] G. Zhang, J.-H. Xue, P. Xie, S. Yang, and G. Wang, "Non-local aggregation for rgb-d semantic segmentation," *IEEE Sign. Process. Letters*, vol. 28, pp. 658–662, 2021.

[47] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021.

[48] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[49] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for rgb-t salient object detection," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 31, no. 5, pp. 1804–1818, 2021.