SpiderMesh: Spatial-aware Demand-guided Recursive Meshing for RGB-T Semantic Segmentation

Siqi Fan[1] Zhe Wang[1] Yan Wang[1*] Jingjing Liu[1*]

Abstract— For semantic segmentation in urban scene understanding, RGB cameras alone often fail to capture a clear holistic topology in challenging lighting conditions. Thermal signal is an informative additional channel that can bring to light the contour and fine-grained texture of blurred regions in low-quality RGB image. Aiming at practical RGB-T (thermal) segmentation, we systematically propose a Spatial-aware Demand-guided Recursive Meshing (SpiderMesh) framework that: 1) proactively compensates inadequate contextual semantics in optically-impaired regions via a demand-guided target masking algorithm; 2) refines multimodal semantic features with recursive meshing to improve pixel-level semantic analysis performance. We further introduce an asymmetric data augmentation technique M-CutOut, and enable semi-supervised learning to fully utilize RGB-T labels only sparsely available in practical use. Extensive experiments on MFNet and PST900 datasets demonstrate that SpiderMesh achieves state-of-the-art performance on standard RGB-T segmentation benchmarks.

a.1 Nighttime scene   a.2 Over-exposure scenario

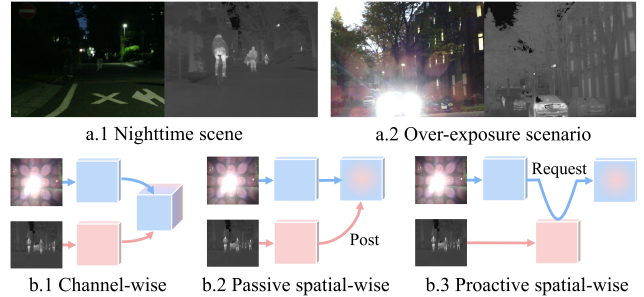b.1 Channel-wise   b.2 Passive spatial-wise   b.3 Proactive spatial-wise

Fig. 1. Some RGB regions are blacked out or blurred in nighttime or over-exposure scenes, while the corresponding thermal images are robust to varying illumination conditions (*a*, images from MFNet dataset [1]). Compared with channel-wise and passive spatial-wise fusion, proactive spatial-wise fusion can fully exploit extra semantic feature for targeted region (*b*). Code is available at GitHub.

I. INTRODUCTION

To realize robust pixel-wise scene understanding in real-world urban environment, RGB-based semantic segmentation is often inadequate due to low image quality from poor lighting conditions, such as nighttime scenes and over-exposure scenarios (as illustrated in Figure 1 a).

One effective way to overcome this is to dynamically adjust the exposure time of cameras. However, this introduces new challenges such as motion blur under long exposure time. To amend this, RGB-T brings in thermal sensors that are relatively robust to variation in illumination conditions [1]–[13]. Unlike visual cameras that react to visible light spectrum, thermal sensors capture infrared radiations emitted by objects [14], which can bring to light the nuanced texture information about environmental surroundings even in challenging lighting conditions.

Existing methods on RGB-T segmentation mainly focus on two aspects: cross-modal feature interaction between RGB and thermal images, and pixel-wise semantic analysis. To exploit cross-modality information from RGB-T pairs, early approaches adopt simple operations such as summation and concatenation [1]–[3], [6]. Recent attention-based fusion methods [4], [5], [7]–[13] integrate RGB and thermal features by taking into consideration their relative context. However, channel-wise fusion directly overlaps RGB and thermal features, blind to their relative spatial positions (Figure 1 b.1). The '*passive*' spatial-wise integration strategy directly provides features without asking for real needs (Figure 1 b.2). To achieve fine-grained fusion for semantic

analysis, multi-supervision has been applied to semantically ambiguous regions [5], [6], but boundary supervision highly relies on accurate pixel-level labeling, which induces high-cost in practical applications. To better leverage the intrinsic contextual relativity between paired thermal signals and RGB features (e.g., to specifically and automatically target darkened areas in RGB images), a more target-guided proactive integration strategy is needed (Figure 1 b.3).

In order to achieve this goal, we need to answer the following questions: 1) How to proactively use thermal signals to compensate the inadequate contextual semantics in optically-impaired regions? 2) How to utilize a small set of RGB-T pairs with limited annotations to reach preferable segmentation performance for practical use?

To address the first challenge, we design a demand-guided target masking algorithm to enhance the features of poorly captured regions in RGB images in a '*request*' manner, via proactive region-level masking with the learned compensation needs. To alleviate information loss incurred in encoding and maximize feature utility, we propose a spatial-aware recursive meshing method, which enhances cross-modal RGB-T features iteratively for pixel-wise semantic analysis. To fully exploit limited labeled pairwise data, we further propose an asymmetric data augmentation technique, named *mono-modal CutOut* (M-CutOut), which creates artificial optically-impaired regions and encourages the network to learn more compensated features from thermal signals. An architecture design with semi-supervised learning capability is also introduced to utilize both natural and artificial regional complementarity in RGB-T. Extensive experiments on the popular MFNet and PST900 benchmarks demonstrate that SpiderMesh achieves state-of-the-art

[1] Institute for AI Industry Research (AIR), Tsinghua University, {fansiqi, wangyan}@air.tsinghua.edu.cn;

that our proposed framework, *Spatial-aware Demand-guided Recursive Meshing* (SpiderMesh), achieves state-of-the-art performance on RGB-T semantic segmentation.

Our contributions are summarized as follows:

- We propose a systematic framework, termed Spider-Mesh, for RGB-T semantic segmentation. Specifically, a demand-guided target masking algorithm is proposed to directly meet the real needs of RGB-T feature compensation in a proactive 'request' manner, and a spatial-aware recursive meshing method to iteratively refine multimodal semantic features.
- To fully leverage the limited labeled pairwise data, we propose a data augmentation technique for RGB-T pairs and firstly extend the task to the semi-supervised setting.
- SpiderMesh not only achieves state-of-the-art performance, but also effectively addresses practical concerns such as computational complexity, robustness to signal loss, and manual labeling cost.

II. RELATED WORK

In this section, we briefly review two related topics, image semantic segmentation and RGB-T segmentation.

A. Image Semantic Segmentation

Image semantic segmentation is a pixel-level scene understanding task. Since FCN [15] performed learning-based segmentation, early RGB methods [16], [17] usually adopted the encoder-decoder network architecture. Deeplabv3 [18] proposed atrous spatial pyramid pooling (ASPP) to apply parallel atrous convolutions with different dilation rates, and SegFormer [19] further boosted the performance. Although RGB segmentation methods have achieved promising progress in recent years, most methods are still susceptible to challenging lighting conditions with poor image quality. RGB-D (depth) methods [20]–[24] leverage depth map to either enhance the whole RGB signal with depth value or highlight RGB features of foreground regions based on depth. However, RGB-D still falls short when additional semantic context for targeted areas is needed for strengthening poorly captured RGB regions.

B. RGB-T Segmentation

Thermal images can provide complementary information for those less informative regions in RGB images. Most of RGB-T fusion employed an explicit aggregation operation [8]–[13]. MFNet [1] collected an RGB-T semantic segmentation dataset and proved significantly performance improvement by utilizing thermal images. Two identical encoders were employed in RTFNet [2] and FuseSeg [3], and the thermal features were gradually integrated into RGB features. FEANet [4] refined the detail features using attention mechanism to deal with small objects. MFFENet [6] used spatial attention to emphasize foreground objects. The multimodal features are fused coarsely with simple operations (e.g., summation, concatenation) in these approaches. GMNet [5] proposed different fusion strategies for shallow and deep features to integrate multi-level features.

Some recent works explored to utilize the power of transformer. MFTNet [25] used modified transformer to learn intraspectral correlations and interspectral interaction, but introduced additional computational complexity. To further boost the performance, alignment-based fusion is utilized via domain adaptation techniques [7], [26]. Different from existing methods, we propose a systematic demand-guided approach addressing not only performance but also practical concerns. We enhance the less informative regions in RGB images with thermal features via proactive spatial-wise interaction. Instead of applying multi-supervision [5], [6], we only utilize semantic supervision considering labeling cost, and enable the extension of our framework to semi-supervised semantic segmentation.

III. SPIDERMESH FRAMEWORK

In this section, we describe the proposed SpiderMesh framework, which consists of demand-guided target masking, spatial-aware recursive meshing, a novel M-CutOut technique for data augmentation, and a mutual learning strategy for semi-supervised adaptation.

A. Overall Architecture

As illustrated in Figure 2, RGB-T semantic segmentation is technically resolved into cross-modal feature interaction and pixel-level feature refinement in SpiderMesh framework. RGB-T pairs are fed into corresponding branches for each modality. Each branch adopts an encoder-decoder structure and employs ResNet [27] as backbone for feature extraction. The number of input channels in the first convolutional layer of the thermal branch is set to 1. Five encoder layers are utilized consecutively to extract features. A DTM (demand-guided target masking) module is embedded after each layer. Data scale is gradually decreased from $H \times W$ to $\frac{H}{32} \times \frac{W}{32}$. Next, a SRM (spatial-aware recursive meshing) module is used as the decoder to enhance unsampled features with fine-grained multimodal semantic features. We utilize bi-linear interpolation for upsampling. Although the two branches are treated equally during encoding and decoding, we regard the RGB branch as the main branch for generating final predictions. Thus, the enhanced thermal feature f_{the}^{e} is introduced to the RGB branch and added with enhanced RGB feature f_{rgb}^{e}, which is further fed to a classifier. Meanwhile, f_{the}^{e} is also fed to a classifier to output an auxiliary prediction. A Convolutional layer is used as the classifier.

B. Demand-guided Target Masking

To better leverage the regional complementary texture feature across RGB and thermal signals, we propose a DTM module, a proactive spatial-wise fusion component whose architecture is illustrated in Figure 3. The overall feature interaction is guided by the demand map dynamically learned via spatial-wise attention. Technically, there are various implementation options for attention-based operations. We adopt a statistical approach [28] as an example to demonstrate our insight of proactive spatial-wise fusion.
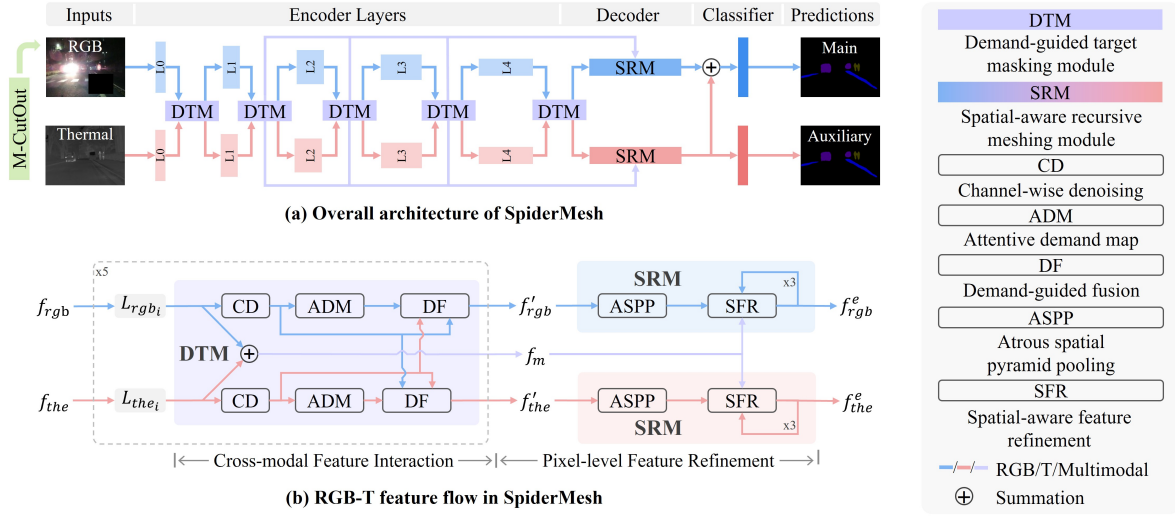
Fig. 2. *Top*: Overall architecture of SpiderMesh; *Bottom*: RGB-T feature flow in SpiderMesh. L_i denotes different layer of the backbone.
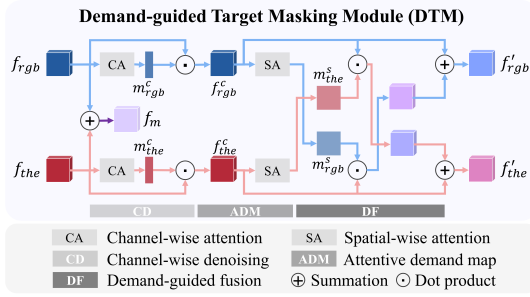

Fig. 3. Architecture of DTM module. The less informative regions are complemented via demand-guided target masking in a 'request' manner.

Take the RGB modality as an example, demand-guided target masking includes the following steps:

Channel-wise Denoising. To eliminate the inevitable camera noise (e.g., over-exposure lighting) before the demand map generation, $f_{rgb} \in \mathbb{R}^{H \times W \times C}$ is first denoised via channel-wise attention. The attention map $m_{rgb}^c \in \mathbb{R}^{1 \times 1 \times C}$ is used to weigh f_{rgb} by element-wise multiplication, resulting in $f_{rgb}^c \in \mathbb{R}^{H \times W \times C}$.

$$f_{rgb}^c = m_{rgb}^c \cdot f_{rgb} = CA(f_{rgb}) \cdot f_{rgb} \qquad (1)$$

where $CA(\cdot)$ is channel-wise attention operator.

Attentive Demand Map. To generate spatial-wise demand for thermal signal complementation, max-pooling and mean-pooling are utilized for spatial-wise statistics. The pooled features are concatenated and forwarded to a convolution operation with a filter in the size of 7×7 for further region-level statistics. After a sigmoid operation, the attentive demand map $m_{rgb}^s \in \mathbb{R}^{H \times W \times 1}$ is obtained, representing the demand of spatial-wise complementation for f_{rgb}^c.

Demand-guided Fusion. The thermal feature f_{the}^c is spatial-wise weighted according to the adaptive demand represented by m_{rgb}^s. Then, f_{rgb}^c is integrated with f_{the}^c

attentively in a 'request' manner:

$$\begin{aligned} f_{rgb}^{'} &= f_{rgb}^c + m_{rgb}^s \cdot f_{the}^c \\ &= f_{rgb}^c + SA(f_{rgb}^c) \cdot f_{the}^c \end{aligned} \qquad (2)$$

where $SA(\cdot)$ is spatial-wise attention operator.

For the thermal modality, $f_{the}^{'}$ can be obtained likewise. In addition, input features f_{rgb} and f_{the} are also fused using summation operation to generate multimodal feature f_m for detailed semantic feature refinement in later stage.

C. Spatial-aware Recursive Meshing

Semantic segmentation is a pixel-level scene understanding task, which relies on fine-grained semantic features for pixel-wise classification. However, detailed information loss caused by downsampling is inevitable during encoding. To compensate information loss and refine the fine-grained features, we propose a SRM module that leverages the fused multimodal features in a recursive manner with spatial awareness. SRM module is composed of a modified ASPP block [18] and three spatial-aware feature refinement blocks, as shown in Figure 4.
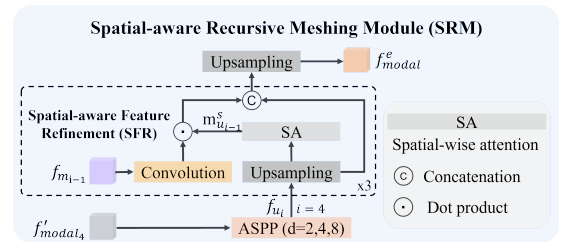

Fig. 4. Architecture of SRM module. *'modal'* is replaced with 'rgb' or 'the' according to which branch it applied in.

For RGB modality, the encoded feature map $f_{rgb_4}^{'}$ is first embedded with more global features via atrous spatial pyramid pooling. We use three dilation rates ($d = 2, 4, 8$), and the number of channels is reduced to 256 after this block.

To refine the fine-grained semantic features, the features of different receptive fields are recursively introduced via skip-connection. Instead of simple concatenation of features, we perform spatial-aware feature refinement to proactively mesh upsampled features with multimodal features f_{m_i} which contain rich detailed semantic information from both RGB-T signals. Considering the complexity, channel reduction is applied to f_{m_i} via convolutional operation. For features with a scale index i, the refinement step can be formulated as:

$$
\begin{aligned}
f_{u_{i-1}} &= E(f_{u_i}, f_{m_{i-1}}) \\
&= Up(f_{u_i}) \oplus (Conv(f_{m_{i-1}}) \cdot m^s_{u_{i-1}}) \\
&= Up(f_{u_i}) \oplus (Conv(f_{m_{i-1}}) \cdot SA(Up(f_{u_i})))
\end{aligned} \quad (3)
$$

where '$\oplus$' is the concatenation operator, '$SA(\cdot)$' is the spatial-wise attention operator, and '$Up(\cdot)$' and '$Conv(\cdot)$' denote the upsampling and convolution operators, respectively. $m^s_{u_i}$ is the spatial-aware attentive mask generated via spatial-wise attention operation, which indicates where and how much the feature needs to be compensated.

The input RGB feature is compensated recursively, which can be represented by:

$$
f^e_{rgb} = Up(E(E(E(ASPP(f'_{rgb_4}), f_{m_3}), f_{m_2}), f_{m_1})) \quad (4)
$$

where '$ASPP(\cdot)$' is the atrous spatial pyramid pooling. Similarly, the encoded thermal feature is compensated recursively to generate f^e_{the}.

D. M-CutOut Augmentation

The key to RGB-T segmentation is to fully exploit the regional complementarity of thermal signals on optically-invisible regions. The Model is supposed to learn the intrinsic contextual relativity between RGB and thermal signals. However, normal CutOut [29] masks all modalities. Different from that, M-CutOut cuts out part of the RGB image with randomly positioned mask M and encourages the model to recover the masked RGB information with thermal signals, which is more in line with the nature of the task. To facilitate the learning of the adaptive cross-modal regional compensation, the artificial optically-impaired regions are randomly created. An example is shown in Figure 5.
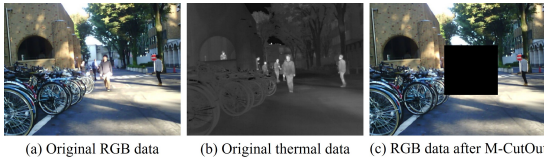

(a) Original RGB data   (b) Original thermal data   (c) RGB data after M-CutOut

Fig. 5. An example of M-CutOut. An optically-impaired region is created.

E. Semi-supervised Learning

Benefiting from dual-branch architecture and M-CutOut, SpiderMesh can be easily extended to semi-supervised segmentation task by leveraging both natural and artificial regional complementarity in RGB-T (Figure 6).

Both labeled and unlabeled data are provided in semi-supervised learning tasks. Without loss of generality, let G
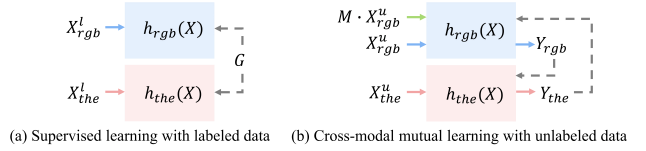

(a) Supervised learning with labeled data   (b) Cross-modal mutual learning with unlabeled data

Fig. 6. Framework for semi-supervised semantic segmentation.

denote the label for labeled data pairs (X^l_{rgb}, X^l_{the}) and (X^u_{rgb}, X^u_{the}) denote unlabeled data. When using labeled data, the network is trained in a supervised manner:

$$
\mathcal{L}_{\mathcal{S}} = \text{CE}(G, h_{rgb}(X^l_{rgb})) + \text{CE}(G, h_{the}(X^l_{the})) \quad (5)
$$

where '$h_{rgb}(\cdot)$' and '$h_{the}(\cdot)$' denote the two branches in SpiderMesh, and '$\text{CE}(\cdot)$' is the cross-entropy loss function. Inspired by the applications of pseudo supervision in RGB segmentation task [30], [31], the network is trained via cross-modal mutual learning with unlabeled data for multimodal alignment-based fusion. We first generate the pseudo label Y_{rgb} and Y_{the} using predictions of original input data with weak augmentations only:

$$
Y_{rgb} = \arg\max_y h_{rgb}(y|X^u_{rgb}) \quad (6)
$$

Y_{the} can be obtained likewise. Then, we apply cross-modal pseudo supervision between the generated pseudo label and the predictions of augmented data using M-CutOut. The cross-modal supervision is conducive to modality alignment:

$$
\mathcal{L}_{\mathcal{U}} = \text{CE}(Y_{the}, h_{rgb}(M \cdot X^u_{rgb})) + \text{CE}(Y_{rgb}, h_{the}(X^u_{the})) \quad (7)
$$

The losses for supervised and unsupervised training are combined to form the final training objective:

$$
\mathcal{L} = \mathcal{L}_{\mathcal{S}} + \mathcal{L}_{\mathcal{U}} \quad (8)
$$

IV. EXPERIMENTS

In this section, we compare our model with other methods and provide detailed ablation studies on standard RGB-T semantic segmentation datasets.

A. Experimental Setting

Datasets We evaluate the proposed SpiderMesh on public datasets of both urban scenes (MFNet [1]) and underground scenes (PST900 [37]). MFNet Dataset [1] is the only public dataset on RGB-T semantic segmentation for urban traffic scenes. It contains $1,569$ pairs of RGB and thermal images captured simultaneously, which comprises 820 daytime and 749 nighttime paired images. For fair comparison, we follow the same splitting scheme as in previous work. Batch size is 6, and the input is resized to a fixed size of 480×640.

PST900 is a challenging underground environment dataset proposed for the DARPA Subterranean Challenge [37]. It contains 894 aligned RGB-T pairs collected from diverse environments with varying lighting conditions. We adopt the same splitting scheme as in [37] for fair comparison. Input data is resized to 720×1280, and batch size is 2.

TABLE I

QUANTITATIVE EVALUATION ON MFNET DATASET. RESULTS OF RGB AND RGB-D SEMANTIC SEGMENTATION METHODS ARE OBTAINED FROM [2], [6], [7], [25]. THE BEST VALUES ARE MARKED BY BOLD, AND THE SECOND ARE MARKED BY UNDERLINE. ALL SCORES ARE IN %.

Category	Methods	mIoU	Car	Person	Bike	Curve	Car-stop	Guardrail	Color-cone	Bump
	UNet [17]	45.1	66.2	60.5	46.2	41.6	17.9	1.8	30.6	44.2
RGB	SwinT [32]	49.0	85.2	57.6	61.0	33.2	28.0	2.4	42.7	33.5
	BiSeNet [33]	50.0	84.1	63.2	60.1	36.7	25.3	5.0	42.2	35.9
RGB-D	SA-Gate [34]	45.8	73.8	59.2	51.3	38.4	19.3	0.0	24.5	48.8
	ACNet [35]	46.3	79.4	64.7	52.7	32.9	28.4	0.8	16.9	44.4
	FuseSeg [3]	54.5	87.9	71.7	<u>64.6</u>	44.8	22.7	6.4	46.9	47.9
	ABMDRNet [7]	54.8	84.8	69.6	60.3	45.1	<u>33.1</u>	5.1	47.4	50.0
	FEANet [4]	55.3	87.8	71.1	61.1	46.5	22.1	6.6	55.3	48.9
RGB-T	MFFENet-single [6]	55.5	87.1	<u>74.4</u>	61.3	45.6	30.6	5.2	**57.0**	40.5
	MFTNet [25]	57.3	87.9	66.8	64.4	47.1	**36.1**	<u>8.4</u>	<u>55.5</u>	62.2
	DooDLENet [36]	57.3	86.7	72.2	62.5	46.7	28.0	5.1	50.7	**65.8**
	SpiderMesh-152 (Ours)	<u>57.9</u>	<u>88.1</u>	72.8	63.7	<u>48.4</u>	28.2	**8.8**	48.2	<u>64.2</u>
	SpiderMesh-B4 (Ours)	**58.4**	**89.9**	**75.3**	**64.8**	**51.5**	31.4	4.5	54.5	55.9

TABLE II

QUANTITATIVE EVALUATION ON PST900 DATASET. RESULTS OF COMPARED BASELINES ARE OBTAINED FROM [6]. THE BEST VALUES ARE MARKED BY BOLD, AND THE SECOND ARE MARKED BY UNDERLINE. ALL SCORES ARE IN %.

Category	Methods	mIoU	Survivor	Hand-drill	Backpack	Fire-extinguisher	Background
RGB	UNet [17]	52.8	31.6	38.3	52.9	43.0	98.0
RGB-D	ACNet [35]	71.8	<u>65.2</u>	51.5	<u>83.3</u>	60.0	<u>99.3</u>
	RTFNet [2]	57.6	36.4	25.4	75.3	52.0	98.9
	PSTNet [37]	68.4	50.0	53.6	69.2	70.1	98.9
RGB-T	ABMDRNet [7]	71.3	62.0	61.5	67.9	66.2	99.0
	MFFENet-single [6]	<u>77.1</u>	63.0	<u>66.8</u>	76.6	**79.8**	<u>99.3</u>
	SpiderMesh-152 (Ours)	**82.3**	**71.9**	**79.7**	**84.0**	<u>76.6</u>	**99.4**

Implementation Details We mainly employ ResNet-152 as backbone. The encoder is initialized with the pre-trained weights provided by PyTorch. The initial learning rate is set to 10^{-2}, and exponential decay scheme is adopted to gradually decrease the learning rate. We use SGD optimizer with momentum for training. The momentum and weight decay are set as 0.9 and 5×10^{-4}. The network is trained until convergence (200 epochs). For training, we apply several data augmentation methods, including random flipping, random cropping, and the proposed M-CutOut. Experiments are implemented in PyTorch on a server with NVIDIA A30.

B. Main Results

Table I reports the comparison results on MFNet dataset. Besides ResNet-152, we also report the performance with MiT-B4 as a reference point for transformer-based approach. SpiderMesh-B4 achieves the best performance on mIoU (58.4%), and outperforms baselines on 4 categories (car, person, bike, and curve). Among them, cars, pedestrians and bikes are the three most common objects in urban scenes. SpiderMesh-152 is 0.5% lower than SpiderMesh-B4, with the best IoU on guardrail and the second best on 3 categories (car, curve, and bump). Most of RGB-T methods outperforms both RGB and RGB-D techniques, which demonstrates the importance of tackling RGB-T segmentation in a task-specific way to leverage the regional complementary. The reported performance of MFFENet-single [6] is only under semantic supervision for fair comparison. Although the full-version GMNet can achieve 57.3 % mIoU utilizing multi-supervision, its performance drops to 53.9% when bound-

ary supervision is not applied [5]. There are performance fluctuations on the *guardrail* category since the unbalanced class distribution (the proportion of the *guardrail* class is 0.095% [6]). Although SpiderMesh-B4 performs better, the complexity is also higher than the ResNet-152 version, so we choose SpiderMesh-152 for the remaining experiments considering practical application.

To further evaluate the proposed SpiderMesh, we also compare it with baselines on PST900 dataset, as reported in Table II. In line with expectations, SpiderMesh consistently outperforms others on mIoU under diverse underground scenes. For 4 foreground categories, it achieves the best performance on 3 (survivor, hand-drill, and backpack). SpiderMesh achieves a performance increase of 8.9% on *survivor* class over MFFENet-single, by effectively leveraging regional complementary features from thermal images.

C. Ablation Study

To better understand SpiderMesh, we conduct several groups of experiments for ablation study on MFNet dataset.

Effect of Each Component. In the baseline network, multimodal features are only fused at the classifier in the RGB branch, and normal feature upsampling operations are adopted. As we can see from Table III, the overall gain of the three components is 5.5%. Among them, performance improvement from DTM and M-CutOut are more significant, thanks to their regional complementation for RGB regions with thermal images. SRM further compensates the spatial-wise information loss with fused multimodal features and leads to an improvement of 0.7%.

TABLE III

ABLATION STUDY ON SPIDERMESH.

Ablated SpiderMesh		mIoU (%)
-	Baseline	52.4
+ Fusion	Summation during encoding	53.4
	Cross-modal weighted fusion	54.5
	DTM	55.2
+ Refinement	SRM w/ self-modal feature only	55.4
	SRM w/ cross-modal feature only	55.4
	SRM	55.9
+ Data aug.	normal CutOut	56.0
	M-CutOut	57.9

TABLE IV

ROBUSTNESS ANALYSIS ON MODALITY SIGNAL LOSS (%).

Input modality	Branch	Daytime	Nighttime	All time
RGB + thermal	RGB	**52.0**	**56.0**	**57.9**
	Thermal	51.0	55.7	57.3
RGB only	RGB	**40.1**	**32.5**	**39.6**
	Thermal	39.8	32.4	39.2
Thermal only	RGB	**41.7**	**51.1**	**50.5**
	Thermal	41.6	50.8	50.2

facing input signal loss.

Complexity Analysis We obtain several versions of SpiderMesh by replacing the backbone. The complexity and corresponding performances are summarized in Table V. SpiderMesh-B4 achieves the best performance with the highest complexity. Considering the computational cost in practice, SpiderMesh-152 achieves a better tradeoff and has advantages on both performance and complexity compared with the two representative methods also using ResNet-152. The model can be more lightweight by employing lighter backbone, and even SpiderMesh-50 (54.4%) can be on par with other RGB-T approaches, like FuseSeg [3] (54.5%).

TABLE V

COMPLEXITY ANALYSIS ON DIFFERENT VERSIONS OF SPIDERMESH.

Version	Backbone	GFlops	mIoU (%)
RTFNet [2]	ResNet-152	290.6	53.2
MFTNet [25]	ResNet-152	330.6	57.3
SpiderMesh-50	ResNet-50	168.2	54.4
SpiderMesh-101	ResNet-101	214.0	56.1
SpiderMesh-152	ResNet-152	259.8	57.9
SpiderMesh-B4	MiT-B4	398.8	58.4

Architecture Design. Firstly, we explore different RGB-T fusion methods in DTM. The simple feature summation during encoding can improve mIoU to 53.4%, but the features are fused without distinction. The cross-modal weighted fusion is in a passive 'post' manner, where one modality is spatial-wise weighted via self-attention and then integrated to the other modality. It performs better than summation, but is still not as good as the proactive manner in DTM. Figure 7 shows that the less informative an area is, the higher its demand for fusion is. For example, the dark and overexposed areas of RGB images usually have higher compensation demands, corroborating our hypothesis on the regional complementary power of thermal signals. Secondly, we study the design choices for SRM, and the refinement with fused multimodal feature is a better choice since it contains rich detailed semantic information from both RGB-T signals. Besides, the benefit of normal CutOut is limited, and the replacement to M-CutOut results in a lift of 1.9%.
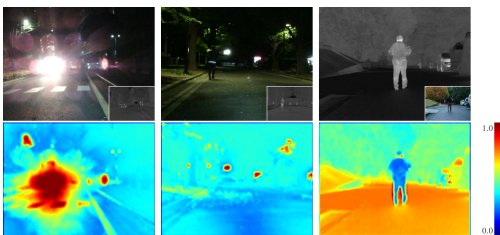

Fig. 7. Visualization of demand map for requesting cross-modal regional complementary information. *Left* and *Middle* are RGB demand maps in dazzling-light and nighttime. *Right* is the thermal demand map in daytime.

Robustness Analysis. It is important to analyze the robustness of multimodal approaches, as signal loss due to software/hardware failure is common in practice. To test SpiderMesh, we manually set the input of a modality to 0 to simulate the signal loss, and evaluate the performance in daytime and nighttime scenarios. The results of both branches are reported in Table IV for analysis. Overall, the performance with inputs from both modalities is the best, which demonstrates the benefit of leveraging multimodal information. The performance differences between RGB-only and thermal-only input indicate that thermal is the more dominant modality due to its high reliability under poor lighting conditions. The slight advantage of the RGB branch comes from further feature fusion at the classifier. As we can see, SpiderMesh can still generate valid predictions when

D. Evaluation on Semi-supervision

We further evaluate SpiderMesh under semi-supervision setting. The 784 labeled images are randomly split into two subsets, 392 images are regarded as unlabeled subset. We make sure that each class appears in the labeled subset. Making use of unlabeled data, SpiderMesh yields a performance lift of **2.1%** (from **53.2%** to **55.3%**). The performance in low-data regime is comparable with that of other methods under full supervision as reported in Table I. Considering the randomness in splitting scheme, we conduct 5 experiments with different partitions, and the gain is $2.1\pm0.1\%$ leveraging unlabeled pairwise RGB-T data.

V. CONCLUSION

In this paper, we manage to fully exploit the regional complementarity of thermal signals on optically-invisible regions. The systematic SpiderMesh framework is proposed. DTM proactively compensates the features of less informative regions via demand-guided target masking in a 'request' manner, SRM recursively refines detailed semantic features for segmentation task. M-CutOut is proposed to create new optically-impaired regions and encourage the model to learn compensated features. Besides, a semi-supervised setting is first explored by leveraging both natural and artificial regional complementarity, which deserves further study.

REFERENCES

[1] 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 *IROS*, 2017, pp. 5108–5115. 1, 2, 4

[2] Y. Sun, W. Zuo, and M. Liu, "RTFNet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE RAL*, vol. 4, no. 3, pp. 2576–2583, 2019. 1, 2, 5, 6

[3] 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 TASE*, vol. 18, no. 3, pp. 1000–1011, 2021. 1, 2, 5, 6

[4] 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 *IROS*, 2021, pp. 4467–4473. 1, 2, 5

[5] 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 TIP*, vol. 30, pp. 7790–7802, 2021. 1, 2, 5

[6] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, "MFFENet: Multiscale feature fusion and enhancement network for rgb-thermal urban road scene parsing," *IEEE TMM*, vol. 24, pp. 2526–2538, 2022. 1, 2, 5

[7] 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 *CVPR*, 2021, pp. 2633–2642. 1, 2, 5

[8] S. Zhao, Y. Liu, Q. Jiao, Q. Zhang, and J. Han, "Mitigating modality discrepancies for rgb-t semantic segmentation," *IEEE TNNLS*, 2023. 1, 2

[9] X. He, M. Wang, T. Liu, L. Zhao, and Y. Yue, "Sfaf-ma: Spatial feature aggregation and fusion with modality adaptation for rgb-thermal semantic segmentation," *IEEE TIM*, vol. 72, pp. 1–10, 2023. 1, 2

[10] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "Rgb-t semantic segmentation with location, activation, and sharpening," *IEEE TCSVT*, vol. 33, no. 3, pp. 1223–1235, 2022. 1, 2

[11] S. Zhao and Q. Zhang, "A feature divide-and-conquer network for rgb-t semantic segmentation," *IEEE TCSVT*, vol. 33, no. 6, pp. 2892–2905, 2023. 1, 2

[12] W. Zhou, S. Dong, J. Lei, and L. Yu, "Mtanet: Multitask-aware network with hierarchical multimodal fusion for rgb-t urban scene understanding," *IEEE TIV*, vol. 8, no. 1, pp. 48–58, 2022. 1, 2

[13] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen, "Delivering arbitrary-modal semantic segmentation," in *CVPR*, June 2023, pp. 1136–1147. 1, 2

[14] M. Vollmer and K.-P. Möllmann, *Infrared Thermal Imaging: Fundamentals, Research and Applications*. Wiley-VCH, 2010. 1

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440. 2

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE TPAMI*, vol. 39, pp. 2481–2495, 2017. 2

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241. 2, 5

[18] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv:1706.05587*, 2017. 2, 3

[19] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *NIPS*, vol. 34, pp. 12 077–12 090, 2021. 2

[20] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *ACCV*, 2017, pp. 213–228. 2

[21] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *ECCV*, 2018, pp. 135–150. 2

[22] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation," in *CVPR*, 2017, pp. 1475–1483. 2

[23] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. H. Lau, and T. S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in *CVPR*, 2019, pp. 2869–2878. 2

[24] Z. Xiong, Y. Yuan, N. Guo, and Q. Wang, "Variational context-deformable convnets for indoor scene parsing," in *CVPR*, 2020, pp. 3992–4002. 2

[25] H. Zhou, C. Tian, Z. Zhang, Q. Huo, Y. Xie, and Z. Li, "Multispectral fusion transformer network for rgb-thermal urban scene semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022. 2, 5, 6

[26] Y.-H. Kim, U. Shin, J. Park, and I. S. Kweon, "MS-UDA: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation," *IEEE Robot. Autom.*, vol. 6, no. 4, pp. 6497–6504, 2021. 2

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 2

[28] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19. 2

[29] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv:1708.04552*, 2017. 4

[30] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *CVPR*, 2021, pp. 2613–2622. 4

[31] S. Fan, F. Zhu, Z. Feng, Y. Lv, M. Song, and F.-Y. Wang, "Conservative-progressive collaborative learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:2211.16701*, 2022. 4

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *CVPR*, 2021, pp. 10 012–10 022. 5

[33] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *ECCV*, 2018, pp. 325–341. 5

[34] 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 *ECCV*, 2020, pp. 561–577. 5

[35] X. Hu, K. Yang, L. Fei, and K. Wang, "ACNet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *ICIP*, 2019, pp. 1440–1444. 5

[36] O. Frigo, L. Martin-Gaffe, and C. Wacongne, "DooDLeNet: Double deeplab enhanced feature fusion for thermal-color semantic segmentation," in *CVPRW*, 2022, pp. 3021–3029. 5

[37] 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 *ICRA*, 2020, pp. 9441–9447. 4, 5