Spatial Balancing for RGB-Thermal Semantic Segmentation in Autonomous Driving: A Study from Analysis to Improvement

Haotian Li, Henry K. Chu, and Yuxiang Sun, *Senior Member*, *IEEE*

***Abstract*—Semantic segmentation based on RGB-Thermal (RGB-T) data fusion has made great progress in the field of autonomous driving. However, we find that most existing RGB-T semantic segmentation methods exhibit inferior performance in image central regions, in which segmentation performance is critical for driving safety. We refer to this phenomenon as spatial bias. To discover the reason for spatial bias, we design a series of experiments. The results challenge the common knowledge that more training data lead to better segmentation performance, and reveal a close causal relationship between segmentation performance and object complexity as well as image quality. We also provide a theoretical interpretation for the causal relationship using information theory and feature space analysis. Based on the findings, we propose a Gaussian-guided regional balancing masking method to balance segmentation performance across different image regions. Moreover, we introduce a spatial-weighted loss to further enhance the overall segmentation performance. Experimental results on two public datasets demonstrate the effectiveness of our method in mitigating spatial bias and improving balanced performance.**

***Index Terms*—Semantic Segmentation, Spatial Balancing, RGB-Thermal Fusion, Autonomous Driving.**

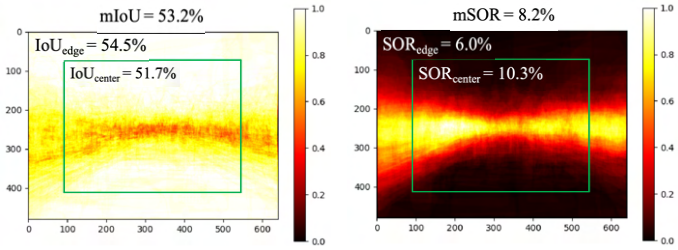
I. Introduction

In recent years, semantic segmentation based on RGB-Thermal (RGB-T) data has attracted increasing attention in the field of autonomous driving, because thermal images can see objects more clearly than RGB images under challenging lighting conditions [1]–[8], such as dim light, nighttime, total darkness, and oncoming headlights. Although existing RGB-T semantic segmentation networks [9]–[11] have made significant progress in terms of segmentation accuracy, they overlook the variations of segmentation performance across different regions of an image. We observe that objects critical to driving safety, such as vehicles, pedestrians, and bicycles, mostly distribute in image central regions. If a model fails to accurately segment objects in the central regions, the performance of downstream tasks [12], [13] could be degraded.

Manuscript received August 10, 2025; Revised December 14, 2025; Accepted March 6, 2026. This paper was recommended for publication by Editor Rafael Murrieta-Cid upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in part by Hong Kong Research Grants Council under Grant 15222523, and in part by City University of Hong Kong under Grant 9610675. *(Corresponding author: Yuxiang Sun.)*

Haotian Li and Henry K. Chu are with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.

Yuxiang Sun is with the Department of Mechanical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong (e-mail: yx.sun@cityu.edu.hk, sun.yuxiang@outlook.com).



(a) Pixel-level IoU distribution (b) Sample Occupancy Rate

Fig. 1. (a) Visualized heatmap of pixel-level IoU distribution, based on the segmentation results from test images of MFNet dataset [7]. IoU$_{center}$ and IoU$_{edge}$ are respectively the IoUs at central and edge regions of the image. mIoU is the mean IoU over the whole image; (b) Visualized heatmap of Sample Occupancy Rate (SOR), calculated from the training set of MFNet dataset. SOR is defined in detail in Sec. III-B1, indicating how frequently pixels are occupied by a non-background label, higher SOR indicates more training samples at a pixel location. SOR$_{center}$ and SOR$_{edge}$ are respectively the SOR at central and edge regions of the image. mSOR is the mean SOR over the whole image. The color scale of the heatmaps ranges from 0 to 1, with darker colors representing lower values and brighter colors higher values.

To study the performance variations of RGB-T semantic segmentation in autonomous driving scenarios across different regions of an image, we visualize the segmentation results of an existing network (i.e., RTFNet-152 [8]) as a heatmap (see Fig. 1(a)). We adopt a ring-shaped partitioning strategy to evaluate the segmentation performance in different regions, and conduct a pixel-level statistical analysis by calculating the average Intersection-over-Union (IoU) across all classes at each pixel location. The heatmap shows that the segmentation performance in central regions is generally worse than the performance in the other regions. Note that a similar phenomenon of such performance variations across different regions can also be found in object detection, in which it is known as spatial bias [14]. So, we also name the phenomenon here as spatial bias. But unlike object detection in which the performance degrades in edge regions, within the context of autonomous driving scenarios, we observe that in semantic segmentation the performance degrades in central regions. Note that such phenomenon may not exist in other domains (e.g., aerial imagery or remote sensing), because the spatial and semantic distributions are different.

It is commonly known that class imbalance is a factor leading to inferior segmentation performance. Class imbalance means that classes with fewer training samples compared to the other classes in a dataset would exhibit inferior performance [15]–[17]. To verify whether the inferior performance in spatial bias is also caused by the same reason, we calculate the Sample Occupancy Rate (SOR) of target objects from the

MFNet dataset (see Fig. 1(b)). The SOR represents the probability of a target object appearing at each pixel location within the training samples. The target objects are `Car`, `Person`, `Bike`, `Curve`, `Car Stop`, `Guardrail`, `Color Cone`, and `Bump`. The heatmap in Fig. 1(b) shows that the target objects are generally distributed densely in central regions, and relatively scarce in edge regions. Comparing Fig. 1(a) and (b), we can see that the regions with higher SOR (i.e., more training data) do not necessarily exhibit better segmentation performance. This is inconsistent with the common reason for inferior performance caused by class imbalance (i.e., less training data leading to the inferior performance). So, class imbalance is not a reason leading to spatial bias.

To discover the underlying reason causing spatial bias, we first studied the evaluation metrics of semantic segmentation. We find that the metrics, such as IoU, only measure the segmentation results over a whole image. To effectively evaluate the segmentation performance in different regions, we then propose a region-based evaluation metric and design a series of experiments. From the experimental results, we discover that spatial bias is closely related to object complexity (e.g., density, overlap) and image quality. By analyzing the spatial characteristics of RGB-T images, we find that the signal-to-noise ratio (SNR) in the central regions of the RGB images is lower and the object complexity is higher, whereas the SNR in the central regions of the thermal images is higher and the object complexity is lower. A possible explanation is the perspective geometry of forward-facing cameras, where distant objects on roads are concentrated in image central regions. These objects appear smaller, which may increase the object complexity at these regions and make accurate segmentation challenging [18], [19]. So, we identify object complexity and image quality as key factors leading to spatial bias. This implies that central regions suffer from inferior quality in RGB modality due to noise and complexity, whereas thermal data retain higher quality.

With the discovered reason, we propose to mitigate the spatial bias issue by designing a Gaussian-guided Regional Balancing Masking (GRBM) method to prioritize thermal features to compensate for RGB ambiguity in these challenging areas. In addition, we propose a spatial-weighted loss to further enhance the overall segmentation performance through spatial enhancement. The experimental results demonstrate the effectiveness of our method. Our code is publicly available[1]. The contributions of this work are summarized as follows:

1) We systematically identify the spatial bias phenomenon in RGB-T semantic segmentation, distinguishing it from the classic class imbalance problem. Through in-depth experiments and theoretical analysis based on information entropy, we reveal that this bias is fundamentally caused by the uneven distribution of object complexity and image SNR.
2) We propose a Gaussian-guided Regional Balancing Masking (GRBM) method. By incorporating a spatial Gaussian prior, this method acts as a spatially-aware information gate that explicitly prioritizes reliable thermal features in complex central regions while retaining RGB details in edge regions.
3) We introduce a novel spatial-weighted loss to further optimize the learning process. By strategically re-weighting pixel contributions based on regional image quality, this loss functions as a regularization term that encourages the model to learn robust features from cleaner edge data, thereby enhancing overall segmentation accuracy.

The remainder of this paper is organized as follows. Section II reviews the related work. Section III analyzes the causes of spatial bias and defines evaluation metrics. Section IV presents the details of our proposed method. Section V discusses the experiments on two benchmark datasets in autonomous driving scenario. Conclusions and future work are drawn in the last section.

II. Related Work

A. RGB-T Semantic Segmentation

In recent years, RGB-T data have attracted increasing attention in various visual perception tasks [11], [20]–[25], especially in semantic segmentation [26]–[31]. Several studies [11], [28]–[30] have proposed various strategies to effectively utilize the complementary features of RGB and thermal images. Shin et al. [11] introduced a complementary random masking strategy and self-distillation loss to mitigate over-reliance on a single modality. Li et al. [28] proposed IGFNet, an illumination-guided fusion network that uses a weight mask from an illumination estimation module to effectively fuse the complementary features from RGB and thermal images. CAINet [29] effectively captures and leverages the complementary relationships between RGB and thermal data through context-aware interactions. Similarly, CACFNet [30] exploits the complementary information between RGB and thermal data while incorporating pixel-region relationships and an enhanced loss function to improve accuracy in RGB-T semantic urban scene understanding.

Several studies have explored other directions. Liang et al. [20] proposed the EAEF method, which considers specific scenarios during fusion. It effectively leverages each type of data to enhance feature extraction and addresses the issue of insufficient representations. Dong et al. [31] introduced EGFNet, which employs edge-aware guidance in combination with multi-modal fusion strategies. This strategy is designed to improve boundary extraction while utilizing high-level semantic information for better representation. Feng et al. [2] also utilized edge information to enhance the performance of a thermal-only semantic segmentation network. SegMiF [32] explicitly generate a fused RGB-T image before semantic segmentation, typically by learning a mapping that integrates RGB and thermal signals at the pixel level. These methods aim to enhance visual quality and exploit the dual-modal correlation for downstream tasks. While pre-fusion offers interpretability and benefits for multi-task learning, it often requires a complete reconstruction pipeline including an encoder-decoder architecture before segmentation, which introduces additional computational overhead and potential redundancy.

[1]Our code is available at: https://github.com/lab-sun/SpatialSeg

Furthermore, RGB-X semantic segmentation [33] expands the feasibility of integrating other modalities of information with RGB images. This is particularly evident in the domain of RGB-D fusion, where the geometric cues from depth sensors provide rich complementary information, a principle that has been successfully leveraged to enhance tasks like salient object detection [34]–[36]. Some studies [9], [10], [37], [38] focus on designing more efficient fusion modules, while others [39], [40] leverage the prior knowledge learned by large models to improve segmentation performance. Although these studies have effectively enhanced the overall performance of RGB-T semantic segmentation, none of them consider the varying performances across different regions in semantic segmentation.

B. Class Imbalance in Semantic Segmentation

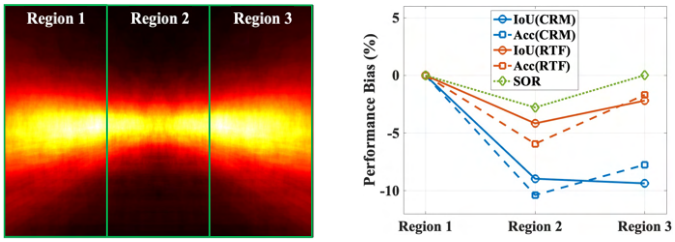
The class imbalance problem [16] has been widely discussed in existing semantic segmentation studies. It arises from the long-tail distribution of samples across different classes in the dataset, which suppresses the segmentation performance of the tail classes. A common solution to this issue is cost-sensitive learning, which improves the performance of tail classes by adopting different strategies in the loss function. Qiu et al. [15] categorized these strategies into class-based loss [41]–[43], pixel-based loss [44]–[46], region-based loss [16], boundary-based loss [47], and compound loss [48].

For instance, Chen et al. [41] introduced importance-aware loss, which prioritizes the accurate segmentation of critical objects by assigning different weights to classes based on their importance for safe driving. The Loss Max-Pooling (LMP) method [44] adaptively re-weights pixel contributions based on observed losses to handle uneven training data distributions. Recall Loss [16] dynamically adjusts class weights based on recall performance, addressing class imbalance while reducing false positives. The Hausdorff distance loss [47] computes the boundary loss by applying distance transforms to both the ground-truth labels and the output predictions. Combo Loss [48] addresses both input and output imbalances by combining the Dice similarity coefficient and cross-entropy, aiming to improve segmentation performance.

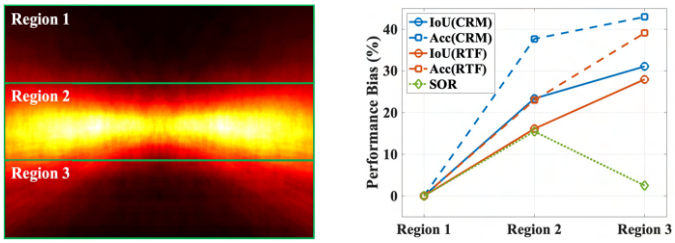
Beyond loss function-based improvements, He et al. [49] proposed a stochastic and deterministic sampling-based attention network, SDSANet, to capture long-range dependencies and refine spatial details for per-pixel segmentation. Additionally, the copy-paste augmentation strategy [50] pastes objects from one image to another to generate a large number of new training samples. While the class imbalance problem differs from the spatial bias, methods designed to solve class imbalance can still provide valuable insights for alleviating spatial bias.

C. Local Evaluation Strategy

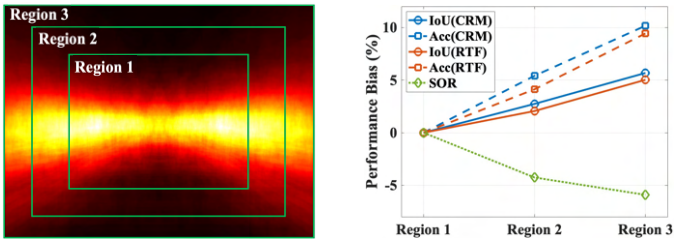
In autonomous driving perception, existing RGB-T semantic segmentation methods typically employ global evaluation strategies to assess segmentation performance [7], [9], [27], [28]. However, local evaluation strategies have been widely used in image quality assessment, such as in image restoration [51], [52] and image super-resolution [53], [54], because they



(a) Vertical Partitioning



(b) Horizontal Partitioning



(c) Ring-shaped Partitioning

Fig. 2. Visualization of the three partitioning strategies and their corresponding performance variations in different regions. CRM-B [11] and RTFNet-152 [8] are used as the segmentation networks. The segmentation performance is represented by the average IoU and average accuracy (Acc) for each region, denoted as IoU (CRM), IoU (RTF), Acc (CRM), and Acc (RTF). The sample size of each region is represented by the Sample Occupancy Rate (SOR). We subtract the value of region 1 from the value of each region to reflect the bias between different regions. The color scale of the heatmaps ranges from 0 to 1 (same as Fig. 1), with darker colors representing lower values and brighter colors higher values.

align more closely with the working principle of human visual system [55]–[58].

For instance, Wang et al. [58] introduced the Structural Similarity Index (SSIM) for local evaluation, computing local SSIM across sliding windows and averaging the results to assess image distortion. Sun et al. [59] proposed a weighted-to-spherically-uniform method that ensures equal influence of pixels based on their mapped spherical area, providing a more accurate and reliable quality assessment. Xue et al. [60] introduced Gradient Magnitude Similarity Deviation (GMSD), which efficiently assesses perceptual image quality by leveraging local gradient magnitude similarities and employing a novel pooling strategy based on the standard deviation of the gradient map. Additionally, Bosse et al. [61] proposed a deep neural network that jointly learns local quality and weights, enabling both no-reference and full-reference image quality assessment.

The human visual system tends to prioritize important areas in an image over global information. In autonomous driving, significant defects in the segmentation accuracy of key areas could lead to safety hazards. So, designing an effective local evaluation strategy is crucial to ensure the safety of autonomous driving systems.

TABLE I
THE COMPARISON BETWEEN INPUT-LEVEL MASKING AND LOSS-LEVEL WEIGHTING.

Category	Method	Core Strategy	Primary Objective
Input-level masking	Random masking [11]	Spatially uniform, stochastic masking	General modality robustness
	GRBM (Ours)	**Spatially aware, prior-driven masking**	**Targeted spatial bias mitigation**
Loss-level weighting	Edge-weighted loss	Increased weight on boundary pixels	Improve boundary segmentation
	Spatial-weighted loss (Ours)	**Region-based weighting guided by SNR/complexity**	**Learning robust features via spatial prior**

D. Differences from Existing Methods

While our proposed GRBM and spatial-weighted loss involve masking and loss weighting, they are fundamentally different in their motivation and mechanism from the existing methods. To clarify our novel contributions, we provide a comparison in Tab. I.

Our GRBM method can be contrasted with the random masking strategy employed by CRM [11]. Random masking applies a spatially-uniform and stochastic mask to prevent the model from over-relying on a single modality. In contrast, GRBM's masking is not uniformly random; its probability distribution is deterministically shaped by a spatial prior derived from domain knowledge in autonomous driving. It is a targeted solution designed specifically to mitigate the spatial bias we identify in Sec. III.

Similarly, our spatial-weighted loss should be distinguished from conventional edge-weighted losses that simply increase weights on object boundaries. Our loss weighting is also directly guided by a spatial prior derived from our analysis of regional noise and complexity. By systematically up-weighting regions with higher signal-to-noise ratios, it encourages the model to ground its learning in more reliable data, leading to a more robust feature representation. This emphasis on regional signal quality, rather than solely on geometric boundaries, is the key difference.

III. SPATIAL BIAS IN SEGMENTATION

A. Differences from Class Imbalance

Class imbalance caused by long-tail distribution may lead to inferior segmentation performance for tail classes. Let $D = \{(x_i, y_i)|i = 1, 2, ..., N_s\}$ denote a dataset, where x_i represents the i-th sample, y_i is the corresponding ground truth, and N_s is the total number of samples. To describe class imbalance, we divide the dataset into subsets. Each subset corresponds to a specific class. For each class $c \in \{1, 2, ..., C\}$, the set D_c is defined as $D_c = \{(x_i, y_i)|y_i = c, i = 1, 2, ..., N_c\}$, where C is the total number of classes, and N_c is the total number of samples belonging to class c. The performance of different classes vary significantly. For example, in the MFNet dataset, the IoU of `Car` is 78% higher than that of `Guardrail`, which is reported in [27].

Class imbalance is reflected by the discrepancies in segmentation results across different classes, while spatial bias refers to uneven segmentation performance across different spatial regions in an image. Similarly, we define $S = \{(x_i, y_i)|i = 1, 2, ..., M_s\}$, where x_i represents the i-th sample, y_i is the corresponding ground-truth label, and M_s is the total number of samples. A whole image can be partitioned into different spatial regions. For each region $r \in \{1, 2, ..., K\}$, the set S_r is defined as $S_r = \{(x_i, y_i)|\text{region}(x_i) = r, i = 1, 2, ..., M_r\}$, where K is the total number of spatial regions, and M_r is the total number of samples in region r. The spatial bias is reflected by the differences of segmentation performance observed from the sample-label pairs S_r in different spatial regions.

B. Reasons Causing Spatial Bias

To thoroughly investigate the reason for spatial bias, we first partition an image into distinct regions and study the performance variations across them. There are several candidate strategies for partitioning. For example, vertical and horizontal partitioning may reveal directional biases along the image axes, while ring-shaped partitioning is particularly suitable for analyzing central-versus-edge performance differences in autonomous driving scenes. This design is motivated by the perspective geometry of forward-facing cameras, where distant objects on the road are projected near the image center and nearby objects are more likely to appear around edge regions [18], [19]. In this section, we therefore compare vertical, horizontal, and ring-shaped partitioning, and show that the ring-based strategy best reveals the spatial bias phenomenon.

1) Sample Size: We first test a hypothesis that spatial bias is caused by different sample sizes (i.e., the number of training samples). We analyze the MFNet dataset [7] using the aforementioned three partitioning strategies. As illustrated in Fig. 2, we compare the segmentation performance (IoU and Acc) against the Sample Occupancy Rate (SOR) for each region. The SOR metric is proposed to measure the number of samples within each region. Specifically, we first calculate the normalized pixel occupancy rate $P(x, y)$. The occupancy rate for each pixel (x, y) is computed by counting how many times the label at that position is non-zero across all ground-truth mask images. Let N_{mask} represent the number of ground-truth mask images $\left\{ L_1, L_2, ..., L_{N_{mask}} \right\}$. For each ground-truth mask image L_i, the label value at position (x, y) is denoted as $L_i(x, y)$. The normalized pixel occupancy rate $P(x, y)$ is defined as:

$$P(x, y) = \frac{1}{N_{mask}} \sum_{i=1}^{N_{mask}} \mathbb{I}\{ L_i(x, y) \neq 0 \}, \tag{1}$$

where $\mathbb{I}\{ L_i(x, y) \neq 0 \}$ is an indicator function which equals 1 if the label $L_i(x, y)$ at position (x, y) is non-zero, and 0 otherwise. $P(x, y)$ reflects the frequency of the pixel being occupied by a non-background label across the dataset.

Then, we calculate the average SOR for each region. Assume that the image is partitioned into K regions

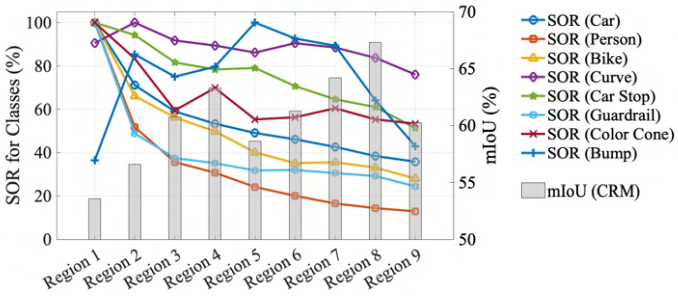


Fig. 3. The SOR (%) distribution for different classes across various regions and the mIoU of all classes for each region. Region 1 represents the innermost region, while region 9 represents the outermost region. We normalize the SOR for each class in different regions to show the sample proportion of each region relative to the region with the highest distribution. We use CRM-B as the segmentation network to obtain the IoU (%) for each region.

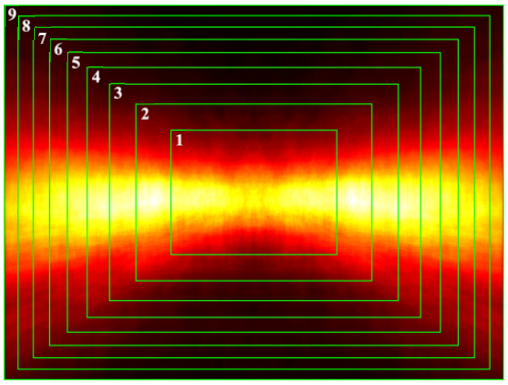


Fig. 4. Visualization of the nine ring-shaped regions with equal areas. Regions are numbered from 1 to 9. Region 1 is the innermost region. Region 9 is the outermost region. The color scale of the heatmap ranges from 0 to 1 (same as Fig. 1), with darker colors representing lower values and brighter colors higher values.

$\{R_1, R_2, ..., R_K\}$, where each region R_k consists of a set of pixels. The average SOR for the region R_k is defined as the mean occupancy rate of all pixels within the region:

$$\mathrm{SOR}_k = \frac{1}{|R_k|} \sum_{(x,y)\in R_k} P(x, y), \tag{2}$$

where $|R_k|$ denotes the number of pixels in region R_k, and $P(x, y)$ is the normalized occupancy rate of the pixel (x, y). This value indicates how frequently pixels in that region are occupied by a non-background label.

From Fig. 2, we can see that there is a positive correlation between segmentation performance and sample size in vertical partitioning, indicating that the number of samples in different regions may have an impact on segmentation performance. In contrast, the results from horizontal partitioning show that while the SOR is highest in the middle region, the corresponding segmentation performance improves gradually from top to bottom, which differs from the results from vertical partitioning. The results of the ring-shaped partitioning are particularly interesting: although the SOR decreases from the inner to outer regions, the segmentation performance improves progressively, which is completely contrary to our expectations. This strongly suggests that the primary factor causing spatial bias in RGB-T semantic segmentation is different from that in the class imbalance problem, in which the dominant reason is the different numbers of training samples for different classes. So, we employ the ring-shaped partitioning strategy in the following analyses, as the observed negative correlation between segmentation performance and sample size across different regions makes it particularly well-suited for examining the underlying reason of spatial bias. Furthermore, as aforementioned, the perspective geometry of forward-facing cameras on vehicles projects distant objects around image central regions, which also justifies the use of concentric rings as a heuristic for spatial performance analysis.

2) Class Distribution: It is known that class imbalance [27] can lead to segmentation performance variances for different classes. We hypothesize that spatial bias is related to class distribution across different regions. In other words, class imbalance combined with uneven distribution of classes across different regions lead to spatial bias. For instance, if low-performing classes (e.g., `Guardrail`) are mainly located in image central regions, and well-performing classes (e.g., `Car` and `Person`) are mainly distributed around edge regions, spatial bias may occur.

To verify this hypothesis, we display the SOR of different classes in each region and the IoU of each region (see Fig. 3). To better illustrate the spatial distribution of SOR and IoU, we partition an image into nine ring-shaped regions of equal area, as shown in Fig. 4. We can see that not only difficult classes are concentrated in the central regions, but also easy classes (e.g., `Car`, `Person`, and `Bike`) are likewise concentrated in the central regions. So, the results negate our hypothesis.

Note that the existing experiments on real-world datasets can only demonstrate that class distribution is not the determinant factor causing spatial bias, but they cannot reveal which other factors might be responsible. This is because analyzing real-world datasets makes it difficult to isolate the effect of a single factor on the results. So, to investigate whether any other spatially-related factors contribute to the bias, we identify a class with a relatively balanced spatial distribution and observe its segmentation performance across different regions. We further illustrate the IoU of different classes across various regions in Tab. II. It is evident that classes with a more evenly distributed sample size (e.g., `Curve`) have significantly lower IoU in central regions compared with the other regions. We observe that most classes, except for `Bump`, exhibit a similar central-concentrated distribution. In addition, classes with better segmentation performance and a higher concentration in the central regions (e.g., `Car`, `Person`, and `Bike`) do not exhibit improved segmentation performance. So, we conclude that there may be some factors in the central regions negatively impacting the segmentation performance and causing spatial bias.

3) Target Object Complexity: From previous analysis, we have already found that the image central regions contain more objects but exhibit inferior segmentation performance. We hypothesize that the inferiority is caused by the target object complexity (e.g., density, overlap), which results in lower IoU despite having a larger number of training samples in the central regions.

To quantitatively measure object complexity, we adopt

TABLE II
THE IOU (%) OF DIFFERENT CLASSES ACROSS VARIOUS REGIONS. REGION 1 REPRESENTS THE INNERMOST REGION, WHILE REGION 9 REPRESENTS THE OUTERMOST REGION. CRM-B [11] IS USED AS THE SEGMENTATION NETWORK. FOR EACH CLASS, THE TWO LOWEST IOU VALUES IN DIFFERENT REGIONS ARE HIGHLIGHTED RESPECTIVELY IN **BOLD** AND UNDERLINED.

Classes	IoU across Regions								
	1	2	3	4	5	6	7	8	9
Car	87.8	90.7	91.9	93.3	93.0	92.2	92.6	92.0	**85.5**
Person	72.7	**72.4**	72.8	75.7	74.4	78.9	79.4	81.7	74.1
Bike	**63.2**	67.3	63.5	70.2	68.2	68.3	68.2	68.1	66.1
Curve	**32.1**	36.0	45.4	50.3	50.6	51.5	51.5	50.9	44.2
Car Stop	49.0	43.9	34.7	**26.4**	31.0	41.3	52.7	48.0	44.2
Guardrail	17.9	5.2	2.9	10.0	11.0	11.4	10.2	8.0	**0.3**
Color Cone	**45.9**	53.5	58.7	69.1	56.1	53.5	61.0	77.9	61.5
Bump	**17.2**	42.8	78.6	76.3	44.7	55.6	63.5	80.3	68.1

Laplacian Variance (LV) as a metric. The rationale for this choice is well founded in image processing literature [62], [63]. The Laplacian operator, as a second-order derivative, is highly responsive to high-frequency components such as edges, lines, and textures. So, the magnitude and variability of the Laplacian response provide a good estimate of object complexity in a region. This principle is widely used in applications like auto-focusing to quantify image sharpness [62]. More importantly, the effectiveness of this type of operator has been well studied. In a detailed review, Pertuz et al. [63] compared many focus measure operators and showed that Laplacian-based methods are reliable and accurate for estimating textures and details in images.

Based on these findings, we adopt Laplacian Variance (LV) to measure object complexity within a region. It is calculated as follows:

$$\sigma^2_{\text{Laplacian}} = \frac{1}{W_{img} H_{img}} \sum_{i=1}^{W_{img}} \sum_{j=1}^{H_{img}} \left[\mathcal{L}(i,j) - \mu_{\text{Laplacian}}\right]^2, \quad (3)$$

where W_{img} and H_{img} are respectively the width and height of an image, $\mathcal{L}(i,j)$ represents the pixel values after the Laplacian transform, and $\mu_{\text{Laplacian}}$ is the mean value of the Laplacian-transformed image.

Higher LV values indicate higher complexity in the regions, in which there are more edges and details in general. So, LV reflects the texture details and complexity of objects. Fig. 5 displays the LV values of RGB and thermal images during daytime and nighttime from the MFNet dataset. The conclusions are listed as follows:

- Both RGB and thermal images have higher LV values in the central regions, indicating that objects in the central regions are more complex, which confirms our hypothesis that spatial bias is related to object complexity in different regions.
- The LV values of RGB images are higher compared with those of thermal images, which is consistent with the widely-known fact that RGB images have more visual textures.

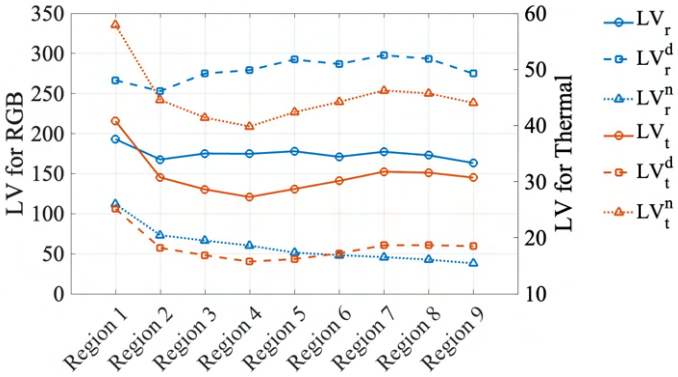


Fig. 5. The Laplacian Variance (LV) for each region. Region 1 represents the innermost region, while region 9 represents the outermost region. LV$_r$, LV$_r^d$, and LV$_r^n$ represent the LV of all RGB images, daytime RGB images, and nighttime RGB images in the dataset, respectively. LV$_t$, LV$_t^d$, and LV$_t^n$ represent the LV of all thermal images, daytime thermal images, and nighttime thermal images in the dataset, respectively.

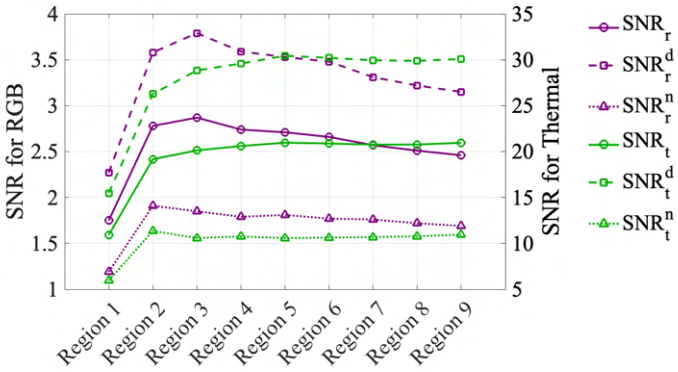


Fig. 6. The Signal-to-Noise Ratio (SNR) for each region. Region 1 represents the innermost region, while region 9 represents the outermost region. SNR$_r$, SNR$_r^d$, and SNR$_r^n$ represent the SNR of all RGB images, daytime RGB images, and nighttime RGB images in the dataset, respectively. SNR$_t$, SNR$_t^d$, and SNR$_t^n$ represent the SNR of all thermal images, daytime thermal images, and nighttime thermal images in the dataset, respectively.

- The LV values of daytime RGB images are higher than those of nighttime RGB images, indicating that daytime RGB images have more visual textures.
- The LV values of daytime thermal images are lower than those of nighttime thermal images, indicating that nighttime thermal images have more visual details, thus can provide a better complement for RGB images.

4) Image Quality: Although in Fig. 5 we can use LV to measure texture details and object complexity, a critical concern is that LV would be sensitive to high-frequency details from both objects and image noise. To disentangle the two potential effects, we also conduct a cross-analysis by comparing images captured under different lighting conditions (e.g., daytime and nighttime), which have different noise characteristics. Here, we use the Signal-to-Noise Ratio (SNR) to measure image quality in different regions. It is calculated as follows:

$$\text{SNR} = \frac{\mu_{\text{signal}}}{\sigma_{\text{noise}} + \epsilon}, \quad (4)$$

where μ_{signal} is the average signal value (i.e., the mean pixel value) of an image, and σ_{noise} is the standard deviation of the image noise, reflecting the intensity of noise in an image. ϵ is

a very small number to prevent division by zero. Higher SNR indicates better image quality and less noise in that region, and vice versa.

Fig. 6 displays the SNR of RGB and thermal images during daytime and nighttime from the dataset. The fact that daytime RGB images have a higher LV despite having a higher SNR (i.e., less noise) evidently indicates that our LV metric is effective in capturing object complexity, not just capturing artifacts from image noise. In addition to this, we have the following conclusions:

- Both RGB and thermal images show more noise in the central regions during both daytime and nighttime, which aligns with our hypothesis.
- The SNR of thermal images is an order of magnitude higher than that of RGB images, reflecting a lower noise level and making them a good complement to RGB images. Thermal images maintain a high SNR under various lighting conditions, providing an advantage in suboptimal or complex lighting environments.
- The SNR of daytime images is higher than that of nighttime images, indicating that images collected during daytime have less noise. For thermal images, although they do not rely on visible light, daytime conditions allow objects to absorb more heat. This results in larger temperature differences between objects, which enhances the contrast in thermal images and ultimately leads to higher SNR. At nighttime, normally with smaller temperature variations for objects in environments, the signal strength captured by thermal sensors decreases, resulting in lower SNR.

In summary, spatial bias arises from the combined effects of target object complexity and image quality across different regions. Objects in central regions are typically farther away, smaller in size, and more prone to overlapping, which increases their complexity. Under limited image resolution, such distant and densely overlapped objects are more likely to be captured as noise, thus degrading image quality and ultimately leading to inferior segmentation performance.

C. Theoretical Interpretation

Our analysis in Sec. III-B reveals a close correlation between spatial segmentation performance degradation and two key factors: high object complexity and inferior image quality. Here we provide a qualitative explanation of our empirical findings using basic concepts from information theory. We use LV and SNR as empirical indicators and discuss them from an information-theoretic perspective, to explain why central regions are more difficult to segment and how this motivates our fusion strategy.

1) Entropy Analysis of Spatial Bias: Information entropy measures the uncertainty within a system. We adopt Shannon Entropy to reason about learning difficulty in a conceptual manner. For a discrete feature variable X with probability mass function $p(x)$, the entropy $\mathcal{H}(X)$ is defined as:

$$\mathcal{H}(X) = -\sum_{x \in X} p(x) \log p(x). \tag{5}$$

We conjecture that the inferior performance in central regions is partly due to the intrinsic difficulty of the task itself. The higher target complexity in these regions, reflected by larger LV, indicates that more diverse object classes and finer structures are densely packed there. This corresponds to a higher label entropy $\mathcal{H}(Y)$, where Y denotes the ground-truth classes, since the local class distribution in central regions is more complex than that in edge regions. In an ideal, noise-free and sufficiently high-resolution imaging setup, such rich structure and diversity could in principle provide more discriminative information and be beneficial for learning.

However, under the constraint of limited sensor resolution in practical autonomous driving settings, this intrinsic complexity can degrade into observation uncertainty. Many distant objects (e.g., small cars, pedestrians, riders) near the image center are projected onto only a few pixels. Their fine-grained details and boundaries cannot be fully resolved and are often aliased or blurred. As a result, high-frequency structures that should have been informative are partially destroyed and are perceived by the model as high-frequency noise. From an information-theoretic perspective, this degradation tends to increase the conditional entropy $\mathcal{H}(X|Y)$ of the observed features X: even for a fixed class Y, the observed features become more variable and ambiguous. In this sense, regions that are potentially the most informative in terms of $\mathcal{H}(Y)$ can, paradoxically, become the least reliable at the feature level due to these imaging constraints.

2) Modality Fusion and Relative Mutual Information: Effective RGB-T fusion should leverage the complementary information between the two modalities so that the fused representation retains as much task-relevant information as possible. Conceptually, for RGB and thermal feature spaces denoted by $\mathbf{R}$ and $\mathbf{T}$, we are interested in maximizing the mutual information between the fused features $f(\mathbf{R}, \mathbf{T})$ and the ground-truth labels Y, i.e., $I(f(\mathbf{R}, \mathbf{T}); Y)$, where f is the fusion function.

Our empirical analyses in Sec. III-B reveal an important asymmetry between the two modalities. In the central regions, the RGB modality shows both higher LV and lower SNR, which suggests that, under limited resolution, many fine structures are degraded into observation noise. In contrast, the thermal modality keeps a higher SNR and a relatively lower LV, meaning that, for the same underlying scene (i.e., the same label entropy $\mathcal{H}(Y)$), thermal features are less affected by noise and have a more stable relationship with Y. Although we do not explicitly compute mutual information in our experiments, these LV and SNR patterns qualitatively support the view that, in central regions, the thermal modality provides more reliable information about Y than the RGB modality, that is, $I(\mathbf{T}; Y)$ is relatively larger than $I(\mathbf{R}; Y)$.

Based on this observation, our Gaussian-guided Regional Balancing Masking (GRBM) can be seen as a spatially-aware information gate. Instead of letting unreliable RGB features in central regions dominate the fusion, GRBM applies a Gaussian-shaped prior centered on the image to probabilistically mask RGB features and keep more thermal information in these areas. This encourages the fusion function $f(\mathbf{R}, \mathbf{T})$ to rely more on the thermal modality where the SNR is higher,

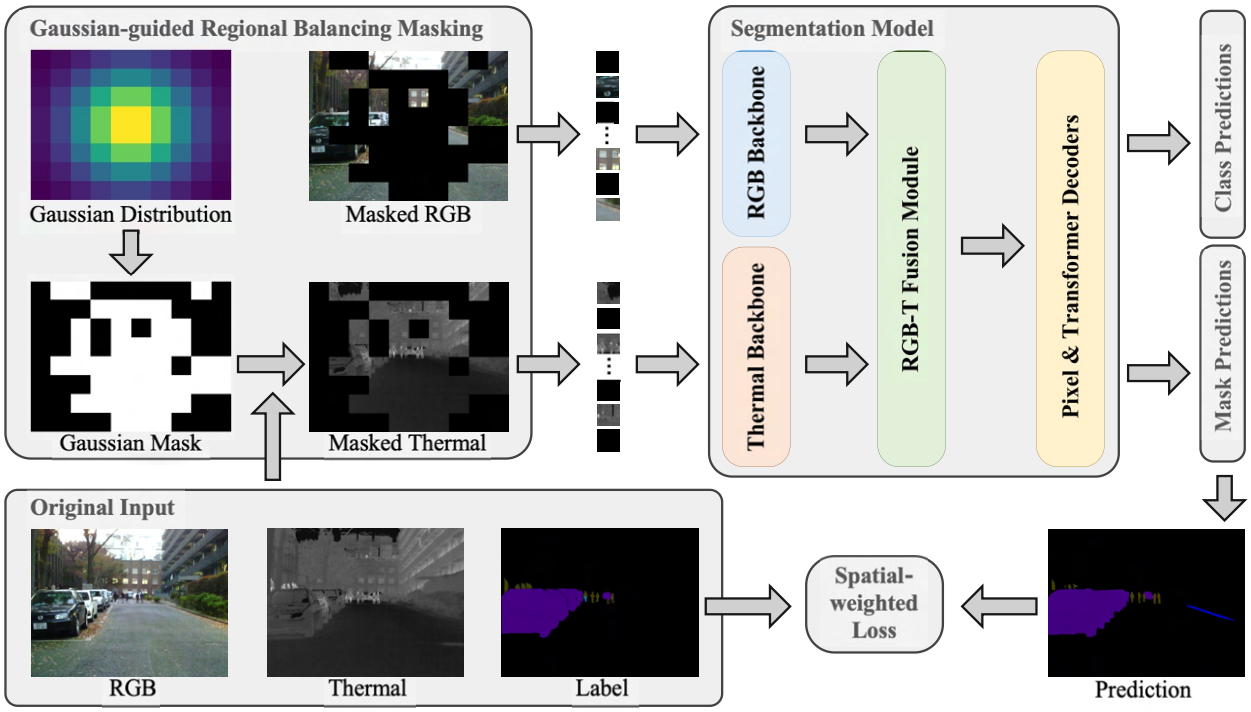


Fig. 7. The overall framework of our method to improve the RGB-T semantic segmentation. Complementary masks for RGB and thermal images are generated through the Gaussian-guided Regional Balancing Masking (GRBM) method. We use CRM [11] as the segmentation model. The spatial-weighted loss is calculated based on the model's predictions and the ground-truth labels.

especially in central regions. In this way, GRBM provides a simple heuristic to improve the usefulness of the fused features for prediction.

IV. The Proposed Method

A. Method Overview

We use our findings to improve RGB-T semantic segmentation. The overall framework of our method is illustrated in Fig. 7. Firstly, we generate complementary masks for RGB and thermal images using our proposed Gaussian-guided Regional Balancing Masking (GRBM) method, which is designed to better utilize thermal image information in central regions during feature fusion in the segmentation model. Then, the masked RGB and thermal images are fed into the segmentation model. The backbone, fusion module, and decoder of the model are borrowed from CRM [11]. Finally, we calculate our designed spatial-weighted loss between model predictions and the corresponding ground truth to improve the segmentation performance.

B. Gaussian-guided Regional Balancing Masking (GRBM)

From our previous analysis, we have shown that segmentation performance in central regions is relatively inferior. Our theoretical interpretation in Sec. III-C further posits that this is due to central regions being a high-entropy space, where the RGB modality is corrupted by noise while the thermal modality offers more reliable information. Based on this information-theory rationale, our Gaussian-guided Regional Balancing Masking (GRBM) method is designed to act as a spatially-aware information gate. We aim to prioritize thermal features more in the central regions and RGB features more in the edge regions. We generate a pair of complementary binary masks for RGB and thermal images based on the Gaussian distribution:

$$\boldsymbol{M}_{rgb} + \boldsymbol{M}_{th} = \mathbf{1}, \tag{6}$$

where $\boldsymbol{M}_{rgb}$ and $\boldsymbol{M}_{th}$ are respectively the masks for RGB and thermal images, $\mathbf{1}$ represents a matrix of all ones.

Given a pair of input RGB-T images $(\boldsymbol{I}_{rgb}, \boldsymbol{I}_{th}) \in \mathbb{R}^{H\times W\times 3}\times \mathbb{R}^{H\times W\times 1}$, we partition the images into $N_x \times N_y$ grids. N_x and N_y are calculated by:

$$N_x = \lfloor \frac{W_{img}}{\mathrm{PS}} \rfloor, \quad N_y = \lfloor \frac{H_{img}}{\mathrm{PS}} \rfloor, \tag{7}$$

where W_{img} and H_{img} are the width and height of the image, PS is the edge length of each partitioned patch. The symbol $\lfloor\cdot\rfloor$ denotes the floor function, which rounds the result down to the nearest integer.

Then, a Gaussian distribution representing the distance to the grid center is generated on the partitioned grid:

$$\mathcal{W}(i,j) = \exp\left(-\frac{(i-\frac{N_x}{2})^2 + (j-\frac{N_y}{2})^2}{2\sigma_p^2}\right), \tag{8}$$

where (i, j) represents the coordinates of the patches in the grid, and σ_p controls the extent of the attention area. $\mathcal{W}(i, j)$ corresponds to the Gaussian distribution shown in Fig. 7, representing the distance from each patch to the grid center. For instance, the closer the patch (i, j) to the grid center, the larger the value of $\mathcal{W}(i, j)$, and vice versa.

The choice of the Gaussian function provides a smooth, continuous model for the gradual transition of information quality from center to edges, aligning with our experimental

findings. In addition, under the principle of maximum entropy, the Gaussian distribution represents the least biased functional form for a unimodal, centrally-peaked prior, making it a principled default choice.

Then, we can generate complementary binary masks for RGB and thermal images based on the Gaussian distribution:

$$M_{rgb} = \text{Bernoulli}(\mathcal{W}(x, y)), \tag{9}$$

$$M_{th} = \mathbf{1} - M_{rgb}, \tag{10}$$

where Bernoulli$(\cdot)$ represents the Bernoulli distribution, which describes binary random events. If the input probability value $\mathcal{W}(x, y)$ is larger, the probability of generating a mask value of 1 at the corresponding position is higher (i.e., masking RGB features). In contrast, a smaller input probability is more likely to retain RGB features. So, the probability of generating a mask value of 1 in M_{rgb} is higher near central regions, while the probability of generating a mask value of 1 in M_{th} is higher near edges.

To enhance the model's robustness in understanding multi-modal features, we employ three different mask combination patterns:

$$\left\langle M_{rgb}, M_{th} \right\rangle_i = \begin{cases} \langle \mathbf{0}, M_{th} \rangle & \text{if } i = 0 \\ \left\langle M_{rgb}, \mathbf{0} \right\rangle & \text{if } i = 1 \\ \left\langle M_{rgb}, M_{th} \right\rangle & \text{if } i = 2 \end{cases}, \tag{11}$$

where $i \in \{0, 1, 2\}$ is a randomly sampled index. $\mathbf{0}$ represents an all-zero matrix, indicating that no mask is used. This stochastic selection of masking patterns is a critical component for regularization [11]. It ensures that the model does not learn to completely discard any single modality, and instead is forced to develop a robust understanding of features from both RGB and thermal inputs, preventing over-reliance on a single data stream.

Then, the mask fusion process of RGB and thermal images can be defined as follows:

$$I'_{rgb} = (\mathbf{1} - M_{rgb}) \odot I_{rgb} + M_{rgb} \odot \tau, \tag{12}$$

$$I'_{th} = (\mathbf{1} - M_{th}) \odot I_{th} + M_{th} \odot \tau, \tag{13}$$

where $\odot$ denotes element-wise multiplication. τ represents a learnable mask token, initialized via a truncated normal distribution with initial values close to zero but containing slight randomness. So, during the mask fusion process, the masked areas are replaced with τ, effectively masking out the lower SNR and more complex RGB features in central regions, while utilizing more thermal features. Following this, the masked RGB and thermal images are fed into the feature extraction and fusion module.

Note that our GRBM is different from feature fusion. Many methods, particularly those employing cross-modal attention (e.g., CMX [9] and CAINet [29]), aim to dynamically learn how to combine features after they have been extracted, but our GRBM operates at the input level. It functions as a spatially-aware inductive bias that injects prior knowledge specific to autonomous driving scenarios before the fusion stage. Our motivation is to provide a segmentation network with more reliable information, rather than relying on the network to learn features that may have already been degraded by noise.

C. Spatial-Weighted Loss

The central regions of RGB and thermal images contain more noise and exhibit greater complexity in target objects compared with the edge regions, primarily due to insufficient image quality.

If we highlight the central regions during loss calculation to alleviate spatial bias, the segmentation model would learn more noise, leading to degradation in the overall segmentation performance (details of this conclusion are presented in the experimental section). In contrast, we introduce a spatial weighting mechanism to dynamically adjust the loss calculation for the edge regions of images, which encourages the segmentation model to focus more on the less noisy and less complex edge regions during training. Specifically, we first obtain a distance matrix based on the Euclidean distance of each pixel from the image center:

$$d_{ij} = \sqrt{(y_i - y_c)^2 + (x_i - x_c)^2}, \tag{14}$$

where (x_c, y_c) represents the coordinates of the image center, and (i, j) represents the coordinates of a pixel. Then, the matrix $d_{ij} \in [0, d_{max}]$ is normalized and its range is linearly transformed to map values to the interval $[1, 1 + \alpha]$:

$$w_{ij} = 1 + \frac{\alpha d_{ij}}{d_{max}}, \tag{15}$$

where d_{max} indicates the maximum value in d_{ij}, and α is the scaling factor. Central pixels get lower weights (closer to 1) to ensure that the central regions are not overlooked during loss calculation, while edge pixels get higher weights (closer to $1 + \alpha$).

The weight matrix w_{ij} is incorporated into both the cross-entropy and dice losses to enhance pixel-level mask prediction during training. The weighted cross-entropy loss is calculated as:

$$\ell_{wce} = \frac{1}{W_{img} \times H_{img}} \sum_{i=1}^{W_{img}} \sum_{j=1}^{H_{img}} w_{i,j} \Big[t_{i,j} \log \left(\text{Sigmoid}(s_{i,j}) \right) + (1 - t_{i,j}) \log \left(1 - \text{Sigmoid}(s_{i,j}) \right) \Big], \tag{16}$$

where $t_{i,j}$ represents the target mask matrix, $s_{i,j}$ represents the prediction probability matrix, Sigmoid$(\cdot)$ denotes the Sigmoid function. The weighted dice loss is calculated as:

$$\ell_{wdice} = 1 - \frac{2 \sum_{i=1}^{W_{img}} \sum_{j=1}^{H_{img}} w_{i,j} \odot t_{i,j} \odot \text{Sigmoid}(s_{i,j})}{\sum_{i=1}^{W_{img}} \sum_{j=1}^{H_{img}} w_{i,j} \odot t_{i,j} + \sum_{i=1}^{W_{img}} \sum_{j=1}^{H_{img}} w_{i,j} \odot \text{Sigmoid}(s_{i,j})}, \tag{17}$$

where $\odot$ denotes element-wise multiplication.

By applying spatial-weighted loss functions, the model maintains the same level of attention on the central regions while focusing more on edge regions. From the theoretical view, this weighting strategy serves as an effective regularization technique. By reducing the influence of the high-noise, high-complexity central regions during loss calculation, it prevents the segmentation model from overfitting to unreliable

TABLE III
THE RESULTS (%) OF THE ABLATION STUDY ON STANDARD DEVIATION (σ_p) IN GAUSSIAN DISTRIBUTION. VAR. IS THE VARIANCE OF IOU ACROSS ALL REGIONS (σ^2_{IoU}). MIOU IS THE MEAN IOU CALCULATED PIXEL-WISELY ACROSS ALL THESE REGIONS (NOT A MEAN VALUE OVER ALL PER-REGION IoU_i). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

σ_p	IoU_1	IoU_2	IoU_3	IoU_4	IoU_5	mIoU	Var.
-	55.28	62.79	58.75	63.36	63.85	61.24	10.90
N_y	**57.88**	61.89	59.11	60.40	61.15	60.38	**2.06**
$N_y/2$	55.87	**62.88**	59.38	60.94	63.21	61.14	7.19
$N_y/4$	56.55	62.64	**60.14**	**64.48**	**64.81**	**61.79**	9.45
$N_y/8$	54.77	62.73	59.19	60.57	62.80	60.54	8.73
$N_y/16$	54.87	62.18	57.95	61.26	64.44	60.82	11.29
$N_y/32$	54.07	61.32	58.83	63.35	63.56	60.63	10.34

TABLE IV
THE RESULTS (%) OF THE ABLATION STUDY ON PATCH SIZE (PS) IN COMPLEMENTARY MASKING GENERATION. VAR. IS THE VARIANCE OF IOU ACROSS ALL REGIONS (σ^2_{IoU}). MIOU IS THE MEAN IOU CALCULATED PIXEL-WISELY ACROSS ALL THESE REGIONS (NOT A MEAN VALUE OVER ALL PER-REGION IoU_i). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

PS	IoU_1	IoU_2	IoU_3	IoU_4	IoU_5	mIoU	Var.
8	55.00	62.33	56.62	61.93	64.60	60.64	13.34
16	55.04	61.23	58.81	62.61	63.44	60.53	**9.18**
32	54.74	62.31	**60.14**	63.55	63.48	61.61	10.83
64	**56.55**	**62.64**	**60.14**	**64.48**	**64.81**	**61.79**	9.45

signals and encourages it to prioritize learning features from the cleaner data at the edges. This method is similar to curriculum learning strategies [64], where the model masters *easier* concepts (edge regions) to build a strong feature foundation before fully tackling the *harder* ones (central regions), and thus improve the overall robustness and accuracy for segmentation performance. The other parts of the loss function remain the same as in CRM [11].

D. Region-Based Evaluation Metric

To examine whether there is spatial bias in the segmentation results and measure the degree of such bias, we need to evaluate the segmentation performance in different regions. We adopt a ring-shaped partitioning strategy, as illustrated in Fig. 2 (c), which partitions an image into K regions of equal area $\mathcal{R} = \{R_1, R_2, ..., R_K\}$ from the center to edges. The IoU for each region is calculated as follows:

$$\text{IoU}_k = \frac{\|\text{TP}_k\|}{\|\text{TP}_k\| + \|\text{FP}_k\| + \|\text{FN}_k\|}, \tag{18}$$

where TP_k, FP_k and FN_k respectively represent the number of true positives, false positives, and false negatives in region R_k. By comparing the IoU results from different regions (IoU_k), we can determine whether there exists spatial bias. Moreover, we can measure the degree of spatial bias by calculating the variance of IoU across all regions in an image:

$$\sigma^2_{\text{IoU}} = \frac{1}{K}\sum_{k=1}^{K}\left(\text{IoU}_k - \mu_{\text{IoU}}\right)^2, \tag{19}$$

where $\mu_{\text{IoU}} = \frac{1}{K}\sum_{k=1}^{K}\text{IoU}_k$ represents the global mean IoU across all regions. The increasing variance of the IoU across all regions (σ^2_{IoU}) indicates more severe spatial bias in the segmentation results, and vice versa.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Datasets

In this study, we train and test our method using two publicly available RGB-T semantic segmentation datasets for autonomous driving applications.

1) MFNet Dataset: The dataset contains 820 pairs of daytime and 749 pairs of nighttime urban scene RGB-T images with hand-labeled annotations. The resolution of the RGB-T images is 640×480. The dataset includes 9 semantic classes: unlabeled background and 8 common object classes. We follow the dataset split strategy used by MFNet [7], using 1568 pairs of images (including flipped images) for training, 392 pairs for validation, and 393 pairs for testing.

2) KP Dataset: The dataset includes 95,000 video frames of resolution 640×512 , with 62,500 daytime images and 32,500 nighttime images. Kim et al. [65] annotate 503 pairs of daytime and 447 pairs of nighttime RGB-T images, with 19 semantic classes identical to those in the Cityscapes dataset [66]. We adopt the same dataset split method as CRM [11], using 499 annotated pairs for training, 140 pairs for validation, and 311 pairs for testing.

B. Implementation Details

We set the patch size PS = 64 in Eq. (7), the scaling factor $\alpha = 2$ in Eq. (15), and the standard deviation $\sigma_p = N_y/4$ in Eq. (8), which is varying with the height of the images in the training set. In Eq. (12) and Eq. (13), τ is a learnable truncated normal distribution, which is initialized with mean $\mu = 0$ and standard deviation $\sigma = 0.02$.

We use the Swin Transformer [67] as the backbone. Our method is implemented using PyTorch with Detectron2, and trained using two NVIDIA RTX 3090 GPU cards. We use the AdamW optimizer and the poly learning rate scheduler for training, with the initial learning rate set to 10^{-4}. The random color jitter, random horizontal flip, and random crop are applied to both RGB and thermal images as the data augmentation methods. Our proposed GRBM module is applied only during training and is disabled during inference. As a result, it does not modify the network architecture at test time and introduces no additional computational cost. The inference speed of our method is therefore identical to that of the baseline CRM. On a single NVIDIA RTX 3090 GPU, our method with a Swin-S backbone achieves 15.83 frame-per-second (FPS) on the MFNet dataset and 15.72 FPS on the KP dataset, respectively.

C. Ablation Study

1) Ablation on Gaussian-guided Regional Balancing Masking: To trade-off performance and model complexity, we use the Swin-S transformer [67] as the backbone for our ablation

TABLE V
THE RESULTS (%) OF THE ABLATION STUDY ON SCALING FACTOR (α) IN SPATIAL-WEIGHTED LOSS. VAR. IS THE VARIANCE OF IOU ACROSS ALL REGIONS (σ^2_{IoU}). MIOU IS THE MEAN IOU CALCULATED PIXEL-WISELY ACROSS ALL THESE REGIONS (NOT A MEAN VALUE OVER ALL PER-REGION IoU_i). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

α	IoU_1	IoU_2	IoU_3	IoU_4	IoU_5	mIoU	Var.
0	56.55	62.64	60.14	**64.48**	64.81	61.79	9.45
1	**56.57**	62.12	59.37	63.65	64.46	61.68	**8.45**
1.5	56.34	63.23	59.33	64.32	64.86	62.10	10.70
2	56.53	**63.87**	59.93	63.72	**65.09**	**62.14**	10.01
2.5	55.63	63.73	**60.55**	64.41	64.57	62.10	9.65
3	55.86	62.37	60.09	63.99	64.04	61.72	9.39

TABLE VI
THE RESULTS (%) OF DIFFERENT SPATIAL WEIGHTING STRATEGIES ON THE LOSS FUNCTION. VAR. IS THE VARIANCE OF IOU ACROSS ALL REGIONS (σ^2_{IoU}). MIOU IS THE MEAN IOU CALCULATED PIXEL-WISELY ACROSS ALL THESE REGIONS (NOT A MEAN VALUE OVER ALL PER-REGION IoU_i). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Focus	IoU_1	IoU_2	IoU_3	IoU_4	IoU_5	mIoU	Var.
-	**56.55**	62.64	**60.14**	**64.48**	64.81	61.79	9.45
Edge	56.53	**63.87**	59.93	63.72	**65.09**	**62.14**	10.01
Center	55.92	63.11	59.47	63.80	63.37	61.56	**9.19**

studies. We first validate the effectiveness of our GRBM method. The core component in the mask generation is the creation of a relative distance Gaussian distribution, which determines the probability of each pixel being selected as a mask. In Eq. (8), we modulate the width of the Gaussian distribution by adjusting its standard deviation σ_p, which in turn determines the extent of the attention area. With larger σ_p, the Gaussian distribution is wider, and the attention area covers a broader region. In contrast, with smaller σ_p, the Gaussian distribution is narrower, and the attention area is more concentrated. We set $\sigma_p = N_y/M_{div}$, where M_{div} is an integer, linking σ_p to the image height N_y. So, a smaller M_{div} results in a larger σ_p, leading to slower weight decay away from the center in the Gaussian distribution, while a larger M_{div} results in a smaller σ_p, leading to faster weight decay and a mask that is more concentrated in image central regions. We set $M_{div} \in \{1, 2, 4, 8, 16, 32\}$, and obtain the segmentation results presented in Tab. III. The first row displays the baseline method for comparison, which is CRM [11] with Swin-S as the backbone.

We find that when $\sigma_p = N_y/M_{div}$ with $M_{div} \leq 4$, the variance of the segmentation results is consistently lower than that of the baseline method. This validates that our GRBM method can effectively alleviate spatial bias in the segmentation results. When $\sigma_p = N_y$, the variance of the segmentation results is significantly lower than that of the baseline method, and the IoU_1 in the central region is also notably higher. However, the IoU_3, IoU_4, and IoU_5 in the edge regions are lower than those of the baseline method, resulting in a significantly lower mIoU for the whole image compared to the baseline method. We find that when $\sigma_p = N_y/32$, the mask is overly concentrated in the center of the image, the IoU_1 in the central region is actually lower than that of the baseline method, while the segmentation performance in the edge regions is well maintained. Consequently, the mIoU for the whole image is lower than that of the baseline method. When $\sigma_p = N_y/4$, the segmentation performance in both the central and edge regions is improved, and the variance of segmentation results across different regions is lower than that of the baseline method. In addition, the mIoU for the entire image exceeds that of the baseline method. So, we select $\sigma_p = N_y/4$ as our optimal model.

We use Eq. (7) to generate grids from images, different grid sizes can be obtained by varying the size of each patch. To evaluate the impact of patch size on model performance, we change the patch size ($\text{PS} \in \{8, 16, 32, 64\}$) while keeping the other parameters constant. The segmentation performance in different regions is shown in Tab. IV.

We find that when the patch size is set to its maximum value of 64, the segmentation performance in all regions reach the best results, and the variance of segmentation results across different regions is relatively small. When the patch size is too small ($\text{PS} = 8$), each patch contains very little image information, and the model can extract limited local information from each patch. It becomes challenging to establish effective global information. As a result, the segmentation performance of the model degrades, and the variance of segmentation results across different regions is large. So, we select $\text{PS} = 64$ as our optimal model.

2) Ablation on Spatial-Weighted Loss: In Eq. (15), the loss function increases its focus on areas far from the center of the image as the scaling factor α increases. To assess the impact of α on model performance, we compare the segmentation results of models with different α values in Tab. V. Here, $\alpha = 0$ represents the model without the spatial-weighted loss, serving as the baseline method.

We find that when α is set to 1.5, 2, and 2.5, the mIoU of the images exceeds that of the baseline method. However, their variance also increases compared with the baseline. This aligns with our expectations of regularization. By relatively down-weighting the noisy central regions, the model avoids overfitting to unreliable cues. While the performance in the center (Region 1) remains stable, the model learns more robust features from the higher-quality edge regions (Region 5), leading to an improvement in the overall mIoU. This confirms that the strategy effectively trades off local bias for global robustness.

Tab. VI compares the impacts of different spatial weighting strategies on the loss function. The first row represents the model without any spatial weighting strategy, while the second row corresponds to our baseline method, which focuses more on image edge regions during training (as shown in the fourth row of Tab. V). The third row adopts the opposite strategy that focuses more on image central regions during training.

From the segmentation results, we can see that the model that emphasizes image edge regions achieves an improvement in IoU_5 and higher mIoU at the cost of increased variance. On the other hand, the model that focuses more on image central regions not only shows worse segmentation performance on the image edge regions, but also presents lower segmentation accuracy (IoU_1) than the baseline. This further validates that

TABLE VII
THE COMPARATIVE RESULTS ON THE MFNET DATASET. WE SHOW ACC (%) AND IOU (%) FOR EACH CLASS, AS WELL AS MACC (%) AND MIOU (%) ACROSS ALL THE CLASSES. ALL THE RESULTS OF THE COMPARED METHODS ARE IMPORTED FROM THEIR ORIGINAL PAPERS. FOR SEGMIF, THE SYMBOL "-" MEANS THAT THE CORRESPONDING DATA ARE NOT AVAILABLE IN THE ORIGINAL PAPER. THE RESULTS DEMONSTRATE THE EFFECTIVENESS OF OUR METHOD, WITH THE TOP TWO RESULTS IN EACH COLUMN HIGHLIGHTED RESPECTIVELY IN **BOLD** AND UNDERLINE. THE PUBLICATION VENUE IS FOLLOWED BY THE PUBLICATION YEAR.

Method	Venue	Backbone	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		
RTFNet [8]	RAL'19	ResNet-152	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
SegMiF [32]	ICCV'23	MiT-B3	96.3	87.8	89.6	71.4	81.2	63.2	63.5	47.5	66.7	31.1	-	-	**85.3**	48.9	84.8	50.3	74.8	56.1
CENet [2]	RAL'23	ResNet	92.0	85.8	78.9	70.0	74.9	61.4	64.8	46.8	39.8	29.3	65.7	8.7	54.1	47.8	77.1	56.9	71.8	56.1
CACFNet [30]	TIV'23	ConvNeXt-B	95.9	89.2	**93.6**	69.5	82.0	63.3	74.0	46.6	49.0	32.4	45.8	7.9	69.8	54.9	82.1	58.3	76.7	57.8
EAEFNet [20]	RAL'23	ResNet-152	95.4	87.6	85.2	72.6	79.9	63.8	70.6	48.6	47.9	35.0	62.8	14.2	62.7	52.4	71.9	58.3	75.1	58.9
CMX [9]	TITS'23	MiT-B2	92.2	89.4	81.3	74.8	73.4	64.7	63.5	47.3	38.8	30.1	36.3	8.1	53.3	52.4	67.7	59.4	67.3	58.2
CMNeXt [33]	CVPR'23	MiT-B4	94.4	90.2	83.9	74.2	77.3	63.8	55.7	45.4	47.5	38.1	32.1	13.4	55.8	51.8	63.8	58.6	67.8	59.3
CAINet [29]	TMM'24	MobileNet-V2	93.0	88.5	74.6	66.3	**85.2**	**68.7**	65.9	**55.4**	34.7	31.5	65.6	9.0	55.6	48.9	**85.0**	60.7	73.2	58.6
EGFNet [31]	TITS'24	ConvNeXt	**96.5**	89.8	92.1	71.6	84.8	63.9	**76.1**	46.7	44.6	31.3	38.7	6.7	71.1	52.0	78.1	57.4	75.6	57.5
		Swin-T	94.6	90.0	85.1	73.1	75.2	63.7	73.0	47.9	51.3	40.7	64.4	9.9	60.0	54.4	72.7	54.2	75.0	59.1
CRM [11]	ICRA'24	Swin-S	95.1	90.6	88.5	**75.5**	78.5	67.2	69.3	48.3	52.3	43.4	62.6	11.8	60.9	56.8	77.2	59.3	76.0	61.2
		Swin-B	95.2	**90.8**	85.6	74.8	78.1	66.6	74.9	45.9	52.7	43.5	70.7	12.2	65.1	**57.4**	80.8	**62.1**	78.0	61.3
		Swin-T	95.2	89.8	83.8	73.9	74.2	63.8	73.1	46.8	55.2	42.1	59.7	10.8	61.6	53.8	71.0	55.6	74.8	59.4
Ours		Swin-S	95.2	90.3	85.9	74.9	78.0	66.9	68.1	48.3	60.1	47.6	53.1	15.8	62.1	56.2	80.8	60.9	75.8	62.1
		Swin-B	95.0	90.2	85.1	74.0	83.1	64.2	67.9	47.9	**71.1**	**55.1**	**71.1**	**42.2**	58.0	52.7	81.8	57.3	**79.1**	**64.6**

TABLE VIII
THE COMPARATIVE RESULTS ON THE KP DATASET. WE SHOW IOU (%) FOR EACH CLASS, AS WELL AS MIOU (%) ACROSS ALL THE CLASSES. ALL THE RESULTS OF THE COMPARED METHODS ARE IMPORTED FROM THEIR ORIGINAL PAPERS. THE SYMBOL "-" MEANS NOT APPLICABLE. THE RESULTS DEMONSTRATE THE EFFECTIVENESS OF OUR METHOD, WITH THE TOP TWO RESULTS IN EACH COLUMN HIGHLIGHTED RESPECTIVELY IN **BOLD** AND UNDERLINE. THE PUBLICATION VENUE IS FOLLOWED BY THE PUBLICATION YEAR.

Method	Venue	Backbone	Road	Sidewalk	Building	Wall	Fence	Pole	Traffic light	Traffic sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	mIoU
MFNet [7]	IROS'17	-	93.5	23.6	75.1	0.0	0.1	9.1	0.0	0.0	69.3	0.2	90.4	24.0	0.0	69.6	0.3	0.3	0.0	0.0	0.6	24.0
RTFNet [8]	RAL'19	ResNet-152	94.6	39.4	86.6	0.0	0.6	0.0	0.0	0.0	81.7	3.7	92.8	58.4	0.0	87.7	0.0	0.0	0.0	0.0	0.5	28.7
CMX [9]	TITS'23	MiT-B4	97.7	53.8	90.2	0.0	47.1	46.2	10.9	45.1	87.2	34.3	93.5	74.5	0.0	91.6	0.0	59.7	0.0	46.1	0.2	46.2
		Swin-T	98.8	56.4	89.0	0.0	62.3	54.1	31.2	31.2	84.3	23.2	94.4	83.6	0.0	94.7	0.0	77.7	0.0	51.4	40.7	51.2
CRM [11]	ICRA'24	Swin-S	98.8	60.7	92.1	0.0	60.4	55.1	31.1	53.2	89.1	27.7	95.0	81.4	17.7	95.2	**1.1**	83.3	0.0	49.9	42.3	54.4
		Swin-B	**99.0**	61.8	91.8	0.0	58.7	50.6	39.1	**55.4**	89.2	23.2	94.3	**85.2**	2.9	95.3	0.0	80.5	0.0	**66.2**	**54.6**	55.1
		Swin-T	98.7	55.0	91.3	0.0	63.8	53.5	29.4	50.9	87.9	22.8	94.8	82.2	0.0	95.0	0.1	83.2	0.0	48.0	38.1	52.3
Ours		Swin-S	98.9	**62.6**	91.6	0.0	57.2	51.4	39.4	52.3	88.3	24.3	94.8	84.5	20.6	**95.6**	0.0	**84.9**	0.0	49.3	47.7	54.9
		Swin-B	98.8	61.0	**92.6**	0.0	**66.8**	**58.7**	**42.0**	50.3	**90.1**	**36.1**	**95.3**	85.0	**39.2**	95.3	0.0	77.4	0.0	52.6	38.8	**56.8**

the spatial bias in the segmentation results is caused by higher complexity of the target objects in the image central regions. With inferior image quality, there is a large amount of noise in the image central regions. Overemphasis on central regions leads the model to learn from noisier signals, which would affect overall performance. So, to trade-off the overall segmentation accuracy and spatial bias, we choose the model with $\alpha = 2$ from Tab. V as our best model.

D. Comparative Study

In this section, we compare our method with previous RGB-T semantic segmentation networks on two RGB-T semantic segmentation datasets from autonomous driving scenarios.

1) Evaluation on MFNet Dataset: We select Swin-T, Swin-S, and Swin-B [67] as the backbones for our semantic segmentation network and compare our network with RTFNet [8], SegMiF [32], CENet [2], CACFNet [30], EAEFNet [20], CMX [9], CMNeXt [33], CAINet [29], EGFNet [31], and CRM [11]. The results are shown in Tab. VII.

Note that several state-of-the-art methods, such as CMX [9] and CAINet [29], employ sophisticated cross-modal attention mechanisms to dynamically learn feature fusion. By employing a simpler max-operation fusion compared to these complex modules, our method achieves significantly better segmentation performance for some rare classes (e.g., `Car Stop` and `Guardrail`), as well as a higher mIoU. Notably, with the stronger Swin-B backbone, our method achieves a new state-of-the-art mIoU of 64.6%, outperforming the baseline CRM (61.3%) by a significant margin of 3.3%. Specifically, our method with Swin-B achieves superior performance on challenging classes such as `Guardrail`, improving the IoU from 12.2% to 42.2%. This improvement indicates that our method effectively preserves details that are often lost by the other methods, validating the effectiveness of our method.

2) Evaluation on KP Dataset: We select Swin-T, Swin-S, and Swin-B [67] as the backbones for our semantic segmentation network, and compare them with MFNet [7], RTFNet [8], CMX [9], and CRM [11]. As shown in Tab. VIII, our method achieves higher mIoUs than CRM when using the same backbones. Specifically, our method with Swin-B achieves significant advantages in most classes, especially in improving the IoU of the challenging `Rider` class by 21.5%.

3) Qualitative Demonstrations: Fig. 8 and Fig. 9 qualitatively compare the segmentation results of our method and CRM [11] on the MFNet [7] and KP [65] datasets, respectively. The white boxes in Fig. 8 and Fig. 9 highlight the superior performance of our method over CRM in the image central regions.

For the results on the MFNet dataset, the first row shows that only our method with Swin-S successfully detects the

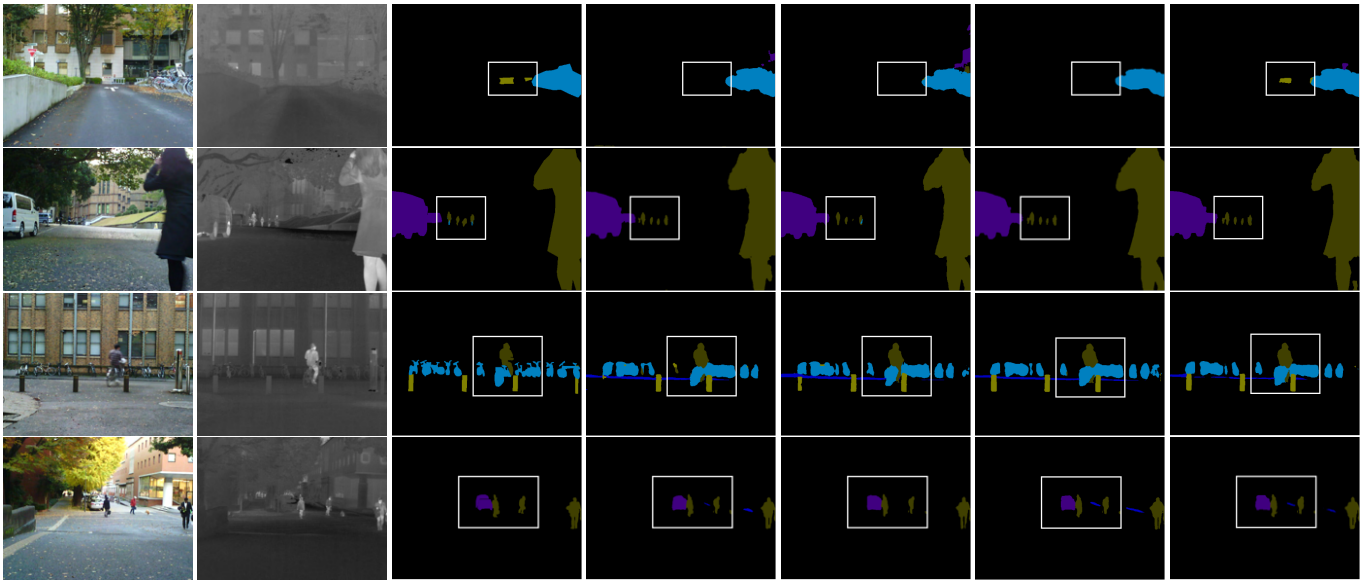


Fig. 8. Sample qualitative demonstrations for RGB-T semantic segmentation on MFNet [7] dataset. From left to right: RGB images, thermal images, ground-truth labels, results of CRM with the Swin-T backbone, results of our method with the Swin-T backbone, results of CRM with the Swin-S backbone, results of our method with the Swin-S backbone.

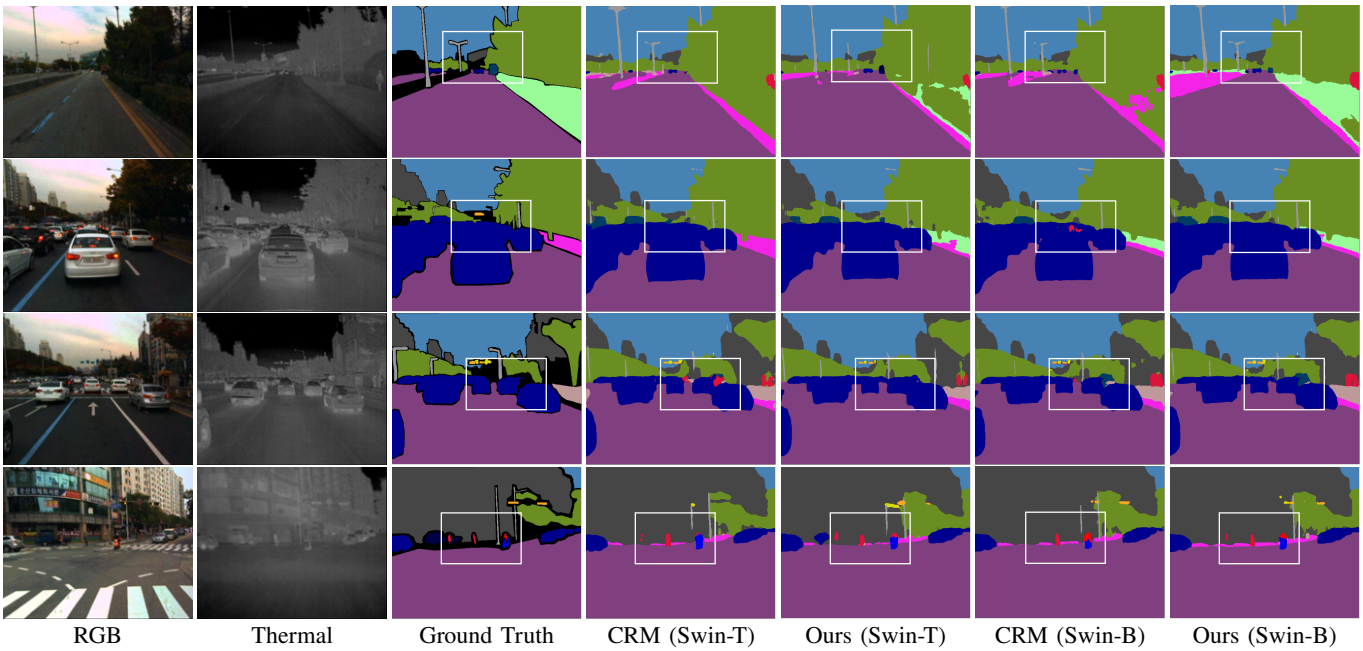


Fig. 9. Sample qualitative demonstrations for RGB-T semantic segmentation on the KP [65] dataset. From left to right: RGB images, thermal images, ground-truth labels, results of CRM with the Swin-T backbone, results of our method with the Swin-T backbone, results of CRM with the Swin-B backbone, results of our method with the Swin-B backbone.

`Car Stop` class in the center of the image, while the second row shows that only our method with Swin-T successfully detects the `Bicycle` class in the center of the image. In the third row, CRM with Swin-T mistakenly identifies the `Bike` class in the central area as the `Car Stop` class. In contrast, our method avoids this issue and achieves more accurate segmentation of the leg of the `Person` class. In the fourth row, both CRM with Swin-T and Swin-S produce false positives by detecting the background as the `Curve` class, whereas our method demonstrates better performance in addressing this issue. These successful segmentation cases highlight the improvement of our method on segmenting the central regions.

For the results on the KP dataset, the first row illustrates that CRM with Swin-T and Swin-B fails to accurately distinguish between the `Car` and `Bus` classes in image central regions. Both models mistakenly identify the `Bus` class as the `Car` class. In contrast, our method with Swin-B successfully differentiates the `Bus` class from the `Car` class. The second and third rows show that CRM incorrectly detects the `Person` class, while our method avoids false positives. In the fourth row, our method achieves better segmentation results for the

`Person` class on the left side of the central regions compared to the CRM. Therefore, based on the visualized semantic segmentation results, we can conclude that our method effectively mitigates spatial bias in the segmentation results, enhances the segmentation performance of challenging classes in the central regions, and reduces false positives.

Despite the effectiveness of our method, it still has limitations. As indicated in Tab. VIII, our performance for the `Motorcycle` class on the KP dataset is lower than the baseline CRM with the same backbone. This is also illustrated in the fourth row of Fig. 9. For very small and complex objects, such as the motorcycle highlighted in the white boxes, both our method and CRM struggle to generate an accurate segmentation map. This indicates that when an object's core features are limited due to small scale or occlusion, the primary challenge shifts from mitigating spatial bias to a more fundamental problem of feature scarcity. This remains a common and challenging open question for the broader field of semantic segmentation.

VI. Conclusions and Future Work

In this paper, we identified the spatial bias problem in RGB-T semantic segmentation for autonomous driving. We discussed the relationship between spatial bias and class imbalance, highlighting their differences, and conducted a detailed analysis on the reasons for spatial bias. According to our experiments, spatial bias is related to the complexity of target objects and the varying image quality across different regions. To alleviate spatial bias, we proposed the GRBM method, which better leverages thermal images in image central regions. This is because thermal images in the central regions typically exhibit better image quality and lower complexity. In addition, we designed a spatial-weighted loss to enhance the model's attention to areas farther from image central regions with fewer noise, further boosting its performance. The experimental results demonstrate that our method achieves competitive performance compared with state-of-the-arts on both the MFNet and KP datasets.

Note that our method is designed for images captured by perspective cameras on vehicles, so it may not be suitable for images from other domains, such as satellite and remote sensing, or medical imaging, where spatial characteristics are totally different from those in autonomous driving. However, we believe that the proposed analytical framework for spatial bias would still be generalizable, and could be adapted to study the spatial characteristics from other domains. So, in addition to the performance improvement brought by our method, another key contribution of this work is the proposed analytical framework for spatial bias.

Despite the success of our method, there still exist limitations. For example, the use of real-world data does not allow us to isolate variables for analysis. So, future study could explore the use of simulation software or advanced generative models to create synthetic datasets. This would allow for rigorous analysis and experimental studies. For instance, object complexity in central regions could be varied while keeping other factors (e.g., background and noise levels) constant.

For future work, we can use generative models to enhance image quality for existing datasets, with a particular focus on image central regions. The enhancement could mitigate the impacts caused by the inferior image quality in central regions, thus improving the segmentation performance. In addition, designing a dynamic gating mechanism that can adapt to unusual environmental conditions (e.g., extreme heat causing thermal artifacts on road surfaces) would be a promising improvement for future work. Moreover, using large vision-language models (VLMs) with specially designed prompts to guide the generation of images across various weather conditions and seasons would also be a potential direction. This could enhance the diversity of training data and improve the generalizability for segmentation performance.

References

[1] H. Li, H. K. Chu, and Y. Sun, "Improving rgb-thermal semantic scene understanding with synthetic data augmentation for autonomous driving," *IEEE Robot. Autom. Lett.*, pp. 1–8, 2025.

[2] Z. Feng, Y. Guo, and Y. Sun, "Cekd: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images," *IEEE Robot. Autom. Lett.*, vol. 8, no. 4, pp. 2205–2212, 2023.

[3] Z. Feng, Y. Guo, D. Navarro-Alarcon, Y. Lyu, and Y. Sun, "Inconseg: Residual-guided fusion with inconsistent multi-modal data for negative and positive road obstacles segmentation," *IEEE Robot. Autom. Lett.*, vol. 8, no. 8, pp. 4871–4878, 2023.

[4] M. R. U. Saputra, C. X. Lu, P. P. B. de Gusmao, B. Wang, A. Markham, and N. Trigoni, "Graph-based thermal–inertial slam with probabilistic neural networks," *IEEE Trans. Robot.*, vol. 38, no. 3, pp. 1875–1893, 2022.

[5] 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 Trans. Multimedia*, vol. 24, pp. 2526–2538, 2021.

[6] 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.*, vol. 18, no. 3, pp. 1000–1011, 2021.

[7] 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/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2017, pp. 5108–5115.

[8] 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.

[9] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, 2023.

[10] J. Huang, J. Li, N. Jia, Y. Sun, C. Liu, Q. Chen, and R. Fan, "Roadformer+: Delivering rgb-x scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion," *IEEE Trans. Intell. Veh.*, 2024.

[11] U. Shin, K. Lee, I. S. Kweon, and J. Oh, "Complementary random masking for rgb-thermal semantic segmentation," in *Proc. IEEE Int. Conf. Robot. Autom.* IEEE, 2024, pp. 11 110–11 117.

[12] C. Shen, S. Yu, B. I. Epureanu, and T. Ersal, "An efficient global trajectory planner for highly dynamical nonholonomic autonomous vehicles on 3-d terrains," *IEEE Trans. Robot.*, vol. 40, pp. 1309–1326, 2024.

[13] D. Milojevic, G. Zardini, M. Elser, A. Censi, and E. Frazzoli, "Codei: Resource-efficient task-driven co-design of perception and decision making for mobile robots applied to autonomous vehicles," *IEEE Trans. Robot.*, pp. 1–20, 2025.

[14] Z. Zheng, Y. Chen, Q. Hou, X. Li, P. Wang, and M.-M. Cheng, "Zone evaluation: Revealing spatial bias in object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 8636–8651, 2024.

[15] S. Qiu, X. Cheng, H. Lu, H. Zhang, R. Wan, X. Xue, and J. Pu, "Subclassified loss: Rethinking data imbalance from subclass perspective for semantic segmentation," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 1547–1558, 2024.

[16] J. Tian, N. C. Mithun, Z. Seymour, H.-P. Chiu, and Z. Kira, "Striking the right balance: Recall loss for semantic segmentation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2022, pp. 5063–5069.

[17] A. Zhang, C. Eranki, C. Zhang, J.-H. Park, R. Hong, P. Kalyani, L. Kalyanaraman, A. Gamare, A. Bagad, M. Esteva, and J. Biswas, "Toward robust robot 3-d perception in urban environments: The ut campus object dataset," *IEEE Trans. Robot.*, vol. 40, pp. 3322–3340, 2024.

[18] X. Li, Z. Jie, W. Wang, C. Liu, J. Yang, X. Shen, Z. Lin, Q. Chen, S. Yan, and J. Feng, "Foveanet: Perspective-aware urban scene parsing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 784–792.

[19] C. Yang, Z. Huang, and N. Wang, "Querydet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13 668–13 677.

[20] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, and T. L. Lam, "Explicit attention-enhanced fusion for rgb-thermal perception tasks," *IEEE Robot. Autom. Lett.*, vol. 8, no. 7, pp. 4060–4067, 2023.

[21] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images," *IEEE Trans. Image Process.*, vol. 32, pp. 1329–1340, 2023.

[22] W. Zhou, F. Sun, Q. Jiang, R. Cong, and J.-N. Hwang, "Wavenet: Wavelet network with knowledge distillation for rgb-t salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 3027–3039, 2023.

[23] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1224–1235, 2022.

[24] W. Zhou, X. Yang, W. Yan, and Q. Jiang, "Hybrid knowledge distillation for rgb-t crowd density estimation in smart surveillance systems," *IEEE Internet Things J.*, vol. 12, no. 7, pp. 9276–9289, 2025.

[25] W. Zhou, Y. Wang, and X. Qian, "Knowledge distillation and contrastive learning for detecting visible-infrared transmission lines using separated stagger registration network," *IEEE Trans. Circuits Syst. I, Reg. Papers*, pp. 1–13, 2024.

[26] J. Li, P. Yun, Y. Xu, Y. Zhang, M. Sun, Q. Chen, I. Alexander, and R. Fan, "Hapnet: Toward superior rgb-thermal scene parsing via hybrid, asymmetric, and progressive heterogeneous feature fusion," *Biomimetic Intelligence and Robotics*, p. 100309, 2026.

[27] H. Li, H. K. Chu, and Y. Sun, "Temporal consistency for rgb-thermal data-based semantic scene understanding," *IEEE Robot. Autom. Lett.*, vol. 9, no. 11, pp. 9757–9764, 2024.

[28] H. Li and Y. Sun, "Igfnet: Illumination-guided fusion network for semantic scene understanding using rgb-thermal images," in *Proc. IEEE Int. Conf. Robot. Biomimetics*. IEEE, 2023, pp. 1–6.

[29] Y. Lv, Z. Liu, and G. Li, "Context-aware interaction network for rgb-t semantic segmentation," *IEEE Trans. Multimedia*, vol. 26, pp. 6348–6360, 2024.

[30] W. Zhou, S. Dong, M. Fang, and L. Yu, "Cacfnet: Cross-modal attention cascaded fusion network for rgb-t urban scene parsing," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 1919–1929, 2024.

[31] S. Dong, W. Zhou, C. Xu, and W. Yan, "Egfnet: Edge-aware guidance fusion network for rgb–thermal urban scene parsing," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 657–669, 2024.

[32] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 8115–8124.

[33] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1136–1147.

[34] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "Ccafnet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in rgb-d images," *IEEE Trans. Multimedia*, vol. 24, pp. 2192–2204, 2022.

[35] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "Irfr-net: Interactive recursive feature-reshaping network for detecting salient objects in rgb-d images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 3, pp. 4132–4144, 2025.

[36] Z. Tu, W. Zhou, X. Qian, and W. Yan, "Hybrid knowledge distillation network for rgb-d co-salient object detection," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 55, no. 4, pp. 2695–2706, 2025.

[37] Z. Wan, P. Zhang, Y. Wang, S. Yong, S. Stepputtis, K. Sycara, and Y. Xie, "Sigma: Siamese mamba network for multi-modal semantic segmentation," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 1734–1744.

[38] D. Qashqai, E. Mousavian, S. B. Shokouhi, and S. Mirzakuchaki, "Csfnet: A cosine similarity fusion network for real-time rgb-x semantic segmentation of driving scenes," *arXiv preprint arXiv:2407.01328*, 2024.

[39] S. Dong, Y. Feng, Q. Yang, Y. Huang, D. Liu, and H. Fan, "Efficient multimodal semantic segmentation via dual-prompt learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2024, pp. 14 196–14 203.

[40] B. Li, D. Zhang, Z. Zhao, J. Gao, and X. Li, "Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation," in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 1308–1317.

[41] B. Chen, C. Gong, and J. Yang, "Importance-aware semantic segmentation for autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 137–148, 2019.

[42] K. Xiang, K. Wang, and K. Yang, "Importance-aware semantic segmentation with efficient pyramidal context network for navigational assistant systems," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2019, pp. 3412–3418.

[43] X. Liu, Y. Lu, X. Liu, S. Bai, S. Li, and J. You, "Wasserstein loss with alternative reinforcement learning for severity-aware semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 587–596, 2022.

[44] S. Rota Bulo, G. Neuhold, and P. Kontschieder, "Loss max-pooling for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, July 2017.

[45] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020.

[47] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 499–513, 2020.

[48] S. A. Taghanaki, Y. Zheng, S. Kevin Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh, "Combo loss: Handling input and output imbalance in multi-organ segmentation," *Computerized Med. Imag. Graph.*, vol. 75, pp. 24–33, 2019.

[49] X. He, J. Liu, W. Wang, and H. Lu, "An efficient sampling-based attention network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 2850–2863, 2022.

[50] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2021, pp. 2918–2928.

[51] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proc. IEEE Int. Conf. Comput. Vis.*, October 2021, pp. 1833–1844.

[52] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2022, pp. 5728–5739.

[53] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, 2016.

[54] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2016.

[55] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 971–982, 2011.

[56] T.-J. Liu and K.-H. Liu, "No-reference image quality assessment by wide-perceptual-domain scorer ensemble method," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1138–1151, 2018.

[57] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[58] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[59] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1408–1412, 2017.

[60] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, 2014.

[61] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, 2018.

[62] J. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in *Proc. 15th Int. Conf. Pattern Recognit.*, vol. 3, 2000, pp. 314–317 vol.3.
[63] S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recognit.*, vol. 46, no. 5, pp. 1415–1432, 2013.
[64] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, ser. ICML '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 41–48.
[65] 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. Lett.*, vol. 6, no. 4, pp. 6497–6504, 2021.
[66] 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.*, June 2016.
[67] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, October 2021, pp. 10 012–10 022.

Haotian Li received the bachelor's degree in opto-electronic information science and engineering from the Tianjin University, Tianjin, China, in 2019, the master's degree in optical engineering from the Tianjin University, Tianjin, China, in 2022, and the Ph.D. degree in mechanical engineering from The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, in 2026.

His research interests include semantic segmentation, computer vision, autonomous driving, and deep learning, etc.

Henry K. Chu (Member, IEEE) received the bachelor's degree in mechanical engineering (mechatronics option) from the University of Waterloo, Waterloo, ON, Canada, in 2005, and the MASc. and Ph.D. degrees in mechanical and industrial engineering from the University of Toronto, Toronto, ON, Canada, in 2007 and 2011, respectively.

He was a Postdoctoral Fellow with the University of Toronto and the City University of Hong Kong, Hong Kong. He is currently an Associate Professor with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong.

His research interests include soft continuum robots, manipulation, vision-based control, artificial intelligence and mechatronic systems.

Yuxiang Sun (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, in 2017.

He is currently an Assistant Professor with the Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong. His research interests include robotics and embodied AI, autonomous driving, robotic perception and control, autonomous navigation, vision-language action, vision-language navigation, quadruped and humanoid robots, semantic scene understanding, localization and mapping, etc.

Prof. Sun serves as an Associate Editor for IEEE Transactions on Intelligent Transportation Systems, IEEE Transactions on Automation Science and Engineering, IEEE Transactions on Intelligent Vehicles, IEEE Robotics and Automation Letters, IEEE International Conference on Robotics and Automation, and IEEE/RSJ International Conference on Intelligent Robots and Systems.