

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

A STUDY ON RGB-THERMAL SEMANTIC
SEGMENTATION FOR AUTONOMOUS DRIVING

LI HAOTIAN

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University

Department of Mechanical Engineering

A Study on RGB-Thermal Semantic Segmentation for
Autonomous Driving

Li Haotian

A thesis submitted in partial fulfilment of the
requirements for the degree of Doctor of Philosophy

November 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

_____ Li Haotian (Name of student)

Abstract

Semantic scene understanding is a fundamental capability of autonomous driving systems, providing crucial perception information for downstream tasks such as localization, navigation, and decision-making. In complex lighting scenarios such as nighttime and strong glare, semantic segmentation methods relying solely on RGB images exhibit performance degradation. Thermal images, being less affected by changes in ambient lighting, can supplement RGB images. Therefore, RGB-Thermal (RGB-T) semantic segmentation methods have received widespread attention in recent years. Although current methods have made significant progress in segmentation capability, several key issues remain unresolved, such as: lack of interpretability of multi-modal fusion strategies, lack of temporal consistency in segmentation results, limited training data, and uneven spatial distribution of segmentation performance. These issues limit the application potential of RGB-T semantic segmentation in autonomous driving.

To address these challenges, the main contributions of this study are as follows:

First, addressing the insufficient interpretability of multi-modal fusion, this dissertation analyzes the inherent characteristics of RGB and thermal images under different lighting conditions and proposes an Illumination-Guided Fusion Network (IGFNet). This network generates a weight mask for each pixel through an

Illumination Estimation Module (IEM), quantifying the importance of RGB information at that location. Subsequently, the Illumination-Guided Cross-Modal Rectification Module (IGCM-RM) adaptively rectifies and fuses RGB and thermal features, thereby more effectively utilizing multi-modal information.

Second, to alleviate the problem of inconsistent segmentation results between consecutive frames, this dissertation proposes a temporal consistency enhancement framework. Existing methods mostly focus on single-frame segmentation performance, neglecting temporal consistency, which may lead to hesitation in autonomous driving decisions. To alleviate this, this dissertation designs a Virtual View Image Generation (VVIG) module, which uses a virtual view transformation (VVT) matrix to synthesize the next frame image, and proposes two consistency loss functions to constrain the segmentation results of consecutive frames. Simultaneously, Consistent Accuracy (CA) is introduced as an evaluation metric, balancing segmentation accuracy and temporal consistency. Experiments demonstrate that this method improves temporal consistency while maintaining high segmentation accuracy.

Third, this dissertation proposes SyntheticSeg, a synthetic data augmentation method, to alleviate the problems of scarce training data and class imbalance. First, a large-scale RGB-T synthetic dataset is constructed using a layout-to-image generative model for joint training with real data. Furthermore, a sampling mechanism considering the scarcity of semantic layout and segmentation difficulty is designed to selectively sample from synthetic data, thereby alleviating class imbalance. Experiments show that SyntheticSeg improves performance by 2.2% compared to baseline methods, with particularly significant results on rare and difficult-to-segment classes such as guardrail and car stop.

Fourth, this dissertation is the first to systematically study the spatial performance imbalance problem in RGB-T semantic segmentation. The study found that in autonomous driving scenarios, the segmentation performance of the image central regions is often weaker than that of the edge regions. The finding challenges the conventional wisdom that "more training samples lead to better performance" and points out that target complexity and image quality are the main causes of spatial bias. Based on this, a Gaussian-guided Regional Balancing Masking (GRBM) method is proposed, which enhances the model's utilization of the thermal features of the central region through masking; simultaneously, a spatially weighted loss function is designed to further improve model performance. Experimental results show that this method effectively alleviates spatial bias and improves overall segmentation accuracy.

Publications arising from the thesis

- [1] **Haotian Li**, and Yuxiang Sun, “IGFNet: Illumination-Guided Fusion Network for Semantic Scene Understanding using RGB-Thermal Images,” in *2023 IEEE International Conference on Robotics and Biomimetics (RO-BIO)*, 2023, pp. 1-6.
- [2] **Haotian Li**, Henry K. Chu, and Yuxiang Sun, “Temporal Consistency for RGB-Thermal Data-Based Semantic Scene Understanding,” *IEEE Robotics and Automation Letters (RA-L)*, vol. 9, no. 11, pp. 9757–9764, 2024.
- [3] **Haotian Li**, Henry K. Chu, and Yuxiang Sun, “Improving RGB-Thermal Semantic Scene Understanding with Synthetic Data Augmentation for Autonomous Driving,” *IEEE Robotics and Automation Letters (RA-L)*, vol. 10, no. 5, pp. 4452-4459, 2025.
- [4] **Haotian Li**, Henry K. Chu, and Yuxiang Sun, “Towards Balanced Spatial Segmentation in Autonomous Driving: Analysis and Improvement of RGB-T Semantic Segmentation,” *IEEE Transactions on Robotics (T-RO)*. (under review)

Acknowledgements

As I reach the end of my PhD journey, I am filled with deep gratitude. I would like to express my sincere thanks to everyone who has supported and accompanied me along the way. Every conversation, piece of advice, and shared experience has been a valuable part of this meaningful chapter of my life.

I am especially grateful to my supervisors, Professor Yuxiang Sun and Professor CHU Kar Hang, Henry, for their guidance, patience, and encouragement. Their insights have shaped the way I approach research and have taught me the importance of integrity and perseverance. I am also thankful for the research opportunities they provided, which not only enriched my academic experience but also offered the financial support that made my studies possible.

My heartfelt thanks go to my colleagues and friends in the laboratory. Their kindness and willingness to help created a friendly and collaborative atmosphere. Working and learning together with them made these past years truly rewarding, and I will always cherish the memories we shared.

I am deeply indebted to my parents, my girlfriend, and my family for their endless love and understanding. Their faith in me has been the foundation of my strength, and their encouragement has carried me through every challenge. They have celebrated my small victories, comforted me in difficult times, and reminded

me of what truly matters. Without their unwavering support, it would have been impossible for me to stay focused and move forward with confidence. I am forever grateful for their patience, sacrifices, and belief in my dreams.

Finally, I would like to thank myself for the resilience and commitment that have guided me throughout this long journey. There were moments of doubt, frustration, and uncertainty, but I chose to keep going, to trust the process, and to continue growing. I am grateful for the perseverance that kept me moving forward and for the courage to embrace both success and failure as part of learning. Looking back, I feel proud—not only of the work accomplished, but also of the person I have become through this experience.

Contents

1	Introduction	2
1.1	Background and Motivation	2
1.2	Challenges in RGB-T Semantic Segmentation	3
1.2.1	Challenge I: Lack of Interpretability in Multi-Modal Fusion	4
1.2.2	Challenge II: Insufficient Temporal Consistency	4
1.2.3	Challenge III: Limited Training Data and Class Imbalance	6
1.2.4	Challenge IV: Spatial Bias in Segmentation Performance .	7
1.3	Proposed Framework and Main Contributions	8
1.4	Dissertation Organization	11
2	Literature Review	13
2.1	RGB-Thermal Semantic Segmentation	13
2.1.1	CNN-based Methods	13
2.1.2	Transformer-based Methods	14
2.1.3	Fusion Strategies	16
2.2	Temporal Consistency in Semantic Segmentation	17
2.2.1	Optical Flow-based Methods	17
2.2.2	Consistency Evaluation Metrics	18

2.3	Data Augmentation and Generative Models	19
2.3.1	Synthetic Data Generation	19
2.3.2	Layout-to-Image Generation	19
2.4	Addressing Imbalance in Semantic Segmentation	20
2.4.1	Class Imbalance	20
2.4.2	Spatial Imbalance and Bias	22
2.4.3	Loss Function Designs	22
3	Illumination-Guided Fusion Network	23
3.1	Motivation	23
3.2	The Proposed Method	24
3.2.1	Network Overview	24
3.2.2	Illumination Estimation Module	26
3.2.3	Illumination-Guided-Cross-Modal Rectification Module	28
3.3	Experimental Results and Discussions	30
3.3.1	Dataset	30
3.3.2	Implementation Details	30
3.3.3	Ablation Study	31
3.3.4	Comparative Study	32
3.4	Conclusions	34
4	Temporal Consistency Framework	35
4.1	Motivation	35
4.2	The Proposed Method	36
4.2.1	The Framework Overview	36
4.2.2	Virtual View Image Generation Module	37

4.2.3	The Loss Functions	39
4.2.4	The Evaluation Metric	40
4.3	Experimental Results and Discussions	42
4.3.1	Datasets	42
4.3.2	Training Details	42
4.3.3	Ablation Study	43
4.3.3.1	Ablation on the VVIG Module	43
4.3.3.2	Ablation on Evaluation Metric	46
4.3.3.3	Ablation on Coefficients of Loss Functions	48
4.3.4	Comparative Experiments	49
4.3.5	Qualitative Demonstrations	51
4.4	Conclusions and Future Work	53
5	Synthetic Data Augmentation	55
5.1	Motivation	55
5.2	The Proposed Method	57
5.2.1	The Overall Framework	57
5.2.2	Sampling Mechanism	59
5.2.3	Data Distribution of Synthetic Dataset	61
5.3	Experimental Results and Discussions	63
5.3.1	Implementation Details	63
5.3.2	Ablation Study	64
5.3.2.1	Ablation on Training Data	64
5.3.2.2	Ablation on Sampling Mechanism	65
5.3.3	Comparative Study	68

5.3.3.1	The Quantitative Results	68
5.3.3.2	The Qualitative Demonstrations	70
5.3.3.3	Suboptimal Case Analysis	71
5.4	Conclusions and Future Work	72
6	Spatial Balancing	74
6.1	Motivation	74
6.2	Spatial Bias in Segmentation	76
6.2.1	Differences from Class Imbalance	76
6.2.2	Reasons Causing Spatial Bias	77
6.2.2.1	Sample Size	79
6.2.2.2	Class Distribution	81
6.2.2.3	Target Object Complexity	83
6.2.2.4	Image Quality	86
6.2.3	Theoretical Interpretation	88
6.2.3.1	Spatial Bias as a Problem of High Local Entropy	88
6.2.3.2	Modality Fusion and Mutual Information	90
6.3	The Proposed Method	91
6.3.1	Method Overview	91
6.3.2	Gaussian-guided Regional Balancing Masking (GRBM) .	92
6.3.3	Spatial-Weighted Loss	96
6.3.4	Region-Based Evaluation Metric	98
6.4	Experimental Results and Discussions	99
6.4.1	Datasets	99
6.4.1.1	MFNet Dataset	99

6.4.1.2	KP Dataset	100
6.4.2	Implementation Details	100
6.4.3	Ablation Study	101
6.4.3.1	Ablation on Gaussian-guided Regional Balanc- ing Masking	101
6.4.3.2	Ablation on Spatial-Weighted Loss	103
6.4.4	Comparative Study	105
6.4.4.1	Evaluation on MFNet Dataset	106
6.4.4.2	Evaluation on KP Dataset	106
6.4.4.3	Qualitative Demonstrations	108
6.5	Conclusions and Future Work	110
7	Conclusions, Limitations, and Future Work	112
7.1	Conclusions	112
7.2	Limitations	113
7.3	Future Work	115
	References	117
	Appendix	128

List of Figures

1.1	Inconsistent segmentation results across consecutive frames. The RGB-T images in the first row are sampled from the MFNet (Ha et al. 2017) dataset. The RGB-T images in the second row are the synthesized images at the next moment. The thermal images are visualized with the <i>jet</i> color map. We use the recent network CMX (J. Zhang, Huayao Liu, et al. 2023) to obtain the segmentation results. The blue and purple colors in the segmentation maps represent bike and car, respectively.	5
1.2	The pixel ratio of each class in the training set of MFNet dataset (Ha et al. 2017) and the IoU for each class of CMX (J. Zhang, Huayao Liu, et al. 2023) and CRM (Shin et al. 2024). We choose CMX using MiT-B2 and MiT-B4 as backbone, and CRM using Swin-S and Swin-B as backbone. This figure shows the correlation between the class pixel ratio in the training set and the IoU for that class.	6

1.3	(a) Visualized heatmap of the pixel-level IoU distribution, based on segmentation results from the test images of the MFNet dataset. $\text{IoU}_{\text{center}}$ and IoU_{edge} are respectively the IoUs at the central and edge regions of the image. mIoU is the mean IoU over the whole image; (b) Visualized heatmap of the Sample Occupancy Rate (SOR), calculated from the training set of the MFNet dataset. SOR indicates that how frequently pixels are occupied by a non-background label, higher SOR indicates more training samples at a pixel location. $\text{SOR}_{\text{center}}$ and SOR_{edge} are respectively the SOR at the central and edge regions of the image. mSOR is the mean SOR over the whole image.	7
1.4	Overview of the Proposed Framework for RGB-T Semantic Segmentation.	11
3.1	The overview of IGFNet for RGB-Thermal semantic segmentation. The inputs are an RGB image and a thermal image. IGFNet consists of a four-stage RGB encoder, a four-stage thermal encoder, a Illumination Estimation Module (IEM), a three-stage Illumination-Guided-Cross-Modal Rectification Module (IGCM-RM), a Cross-Modal Feature Rectification Module (CM-FRM), a four-stage Feature Fusion Module (FFM) and a decoder. The encoder and decoder are borrowed from Segformer (Xie et al. 2021). The CM-FRM and FFM are borrowed from CMX (J. Zhang, Huayao Liu, et al. 2023). The thermal image is colored with the <i>jet</i> color map.	25

3.2	a) The pipeline of our proposed IEM. The input to IEM is an RGB image, from which we derive a grey image by calculating the mean illumination. We then obtain the weight mask by applying a Gaussian filter to this grey image. b) The graph illustrates the application of the Gaussian filter, mapping the grey image to the weight mask. c) The corresponding thermal image shows complementary information.	26
3.3	The structure of our proposed IGCM-RM. The inputs to this module are the feature maps of the RGB and thermal images from the encoder, as well as the weight mask generated by the IEM. After rectification guided by the weight mask, the rectified feature maps are divided into two streams. One stream serves as input to the subsequent stage encoder, while the other is input to the FFM for feature fusion. The part in green background is borrowed from CMX(J. Zhang, Huayao Liu, et al. 2023).	28
3.4	Sample qualitative demonstrations. The rows from top to bottom are RGB images, thermal images, ground truth, CMX results, and IGFNet results.	33

- 4.1 The overview of our temporal-consistent framework for RGB-T semantic segmentation. Img_1 is sampled from MFNet. Img_2 is generated using our virtual view image generation (VVIG) module. The encoders and decoders are borrowed from CMX (J. Zhang, Huayao Liu, et al. 2023). The warped prediction $\hat{y}_{12}$ and ground truth y_2 are also generated by the VVIG module. The *jet* color map is used here to visualize the thermal images. 37
- 4.2 The pipeline of our VVIG module. α, γ, β are respectively the randomly generated Euler angles about the $\hat{Z}, \hat{Y}, \hat{X}$ axes. The VVT matrix (i.e., M_{VVT}) is generated from the Euler angles and the intrinsic matrix. The blue dashed box on the transformed image represents the shape of the original image after the M_{VVT} transformation. The red dashed box on the cropped image shows the largest inner rectangle within the valid range for cropping. The virtual image is obtained from the cropped image by interpolation. 38
- 4.3 (a) Intersection over union (IoU): $y_2 = A + B$, represents the ground truth of Img_2 ; $\hat{y}_2 = B + C$, represents the prediction of Img_2 . (b) Temporal consistency (TC): $\hat{y}_{12} = A + B$, represents the prediction of Img_1 mapped to Img_2 by M_{VVT} ; $\hat{y}_2 = B + C$, represents the prediction of Img_2 . (c) Consistent accuracy (CA): $\hat{y}_{12} = A + B + D + E$, represents the prediction of Img_1 mapped to Img_2 by M_{VVT} ; $\hat{y}_2 = B + C + E + F$, represents the prediction of Img_2 ; $y_2 = D + E + F + G$, represents the ground truth of Img_2 . 41

- 4.4 The ablation study results (%) of temporal consistency (TC) and consistent accuracy (CA). The horizontal axis shows each class in the MFNet dataset (Ha et al. 2017). Variants (G), (H) and (I) are the three variants of our method. The figure illustrates the superiority of variant (H) in terms of segmentation consistency. The figure is best viewed in color. 47
- 4.5 Qualitative demonstration for RGB-T semantic segmentation on the MFNet (Ha et al. 2017) dataset. The rows from top to bottom are RGB images, thermal images, ground truth, CMX results, Ours-dice results, Ours-con results, and Ours-con-acc results. All of the variants use MiT-B2 as backbones. The 1st and 2nd columns are the 1st set of images, the 3rd and 4th columns are the 2nd set of images, the 5th and 6th columns are the 3rd set of images, and the 7th and 8th columns are the 4th set of images. Among them, the first two sets and the last two sets are the images of nighttime and daytime, respectively. 52
- 5.1 The overall framework of our SyntheticSeg. (a) illustrates the training process of RGB-T generative model; (b) illustrates the pipeline for joint training of segmentation models using both synthetic and real RGB-T datasets. 57

5.2	The pixel ratio of each class in the training set of MFNet dataset (Ha et al. 2017) and the IoU for each class of CMX (J. Zhang, Huayao Liu, et al. 2023) and CRM (Shin et al. 2024). We choose CMX using MiT-B2 and MiT-B4 as backbone, and CRM using Swin-S and Swin-B as backbone. This figure shows the correlation between the class pixel ratio in the training set and the IoU for that class.	59
5.3	The normalized pixel ratio of each class in the synthetic dataset with different R_{max}	62
5.4	Qualitative comparison for semantic segmentation of RGB-T images on MFNet (Ha et al. 2017) dataset. The rows from top to bottom are RGB images, thermal images, ground truth of semantic layouts, results obtained by training CMX (J. Zhang, Huayao Liu, et al. 2023) with MiT-B2 using real data, and results obtained by training CMX (J. Zhang, Huayao Liu, et al. 2023) with MiT-B2 using real and sampled synthetic data. The first four columns and the last four columns are the samples of nighttime and daytime, respectively.	70
5.5	Visualization of the segmentation suboptimal case. (a) and (b) are real RGB-T images from MFNet dataset; (c) is the corresponding semantic layout; (d) and (e) are synthetic RGB-T images generated from (c); (f) is the prediction of segmentation model.	71

- 6.1 Visualization of three partitioning strategies and their corresponding performance variations in different regions. CRM-B (Shin et al. 2024) and RTFNet-152 (Sun, Zuo, and M. Liu 2019) 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. 78
- 6.2 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. 81
- 6.3 Visualization of nine ring-shaped regions with equal areas. Regions are numbered from 1 to 9. Region 1 is the innermost region and region 9 to the outermost region. 82

6.4	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.	85
6.5	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.	87
6.6	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 (Shin et al. 2024) as the segmentation model. The spatial-weighted loss is calculated based on the model's predictions and the ground truth labels.	92

- 6.7 Qualitative demonstration for RGB-T semantic segmentation on MFNet (Ha et al. 2017) dataset. From left to right: (1) RGB images, (2) thermal images, (3) ground truth labels, (4) results of CRM with Swin-T backbone, (5) results of our method with Swin-T backbone, (6) results of CRM with Swin-S backbone, (7) results of our method with Swin-S backbone. 107
- 6.8 Qualitative demonstration for RGB-T semantic segmentation on the KP (Y.-H. Kim et al. 2021) dataset. From left to right: (1) RGB images, (2) thermal images, (3) ground truth labels, (4) results of CRM with Swin-T backbone, (5) results of our method with Swin-T backbone, (6) results of CRM with Swin-B backbone, (7) results of our method with Swin-B backbone. 108

List of Tables

3.1	The results of the ablation study. '✓' means the IEM and IGCM-RM are used behind the n-th stage of the encoder. '—' means the CM-FRM (J. Zhang, Huayao Liu, et al. 2023) is used behind the n-th stage of the encoder. The best results are highlighted in bold font.	31
3.2	The comparative per-class results on the MFNet dataset (Ha et al. 2017). To evaluate the methods, we utilized the IoU of each class and the mean IoU of all classes. The results demonstrate the superiority of our IGFNet, with the top two results in each column are highlighted in red and blue.	32
4.1	Results (%) of the ablation study on the VVIG module with different losses. Img_1 and Img_2 indicate the losses applied to the current frame and next frame, respectively. The best results are highlighted in bold font. We use the average results generated from the same 3 sets of random Euler angles as the final results of TC (%) and CA (%).	43

4.2	Results (%) of the ablation study on the components of the VVIG module. Rot. and Tra. indicate the rotation and translation of the VVIG module, respectively. The best results are highlighted in bold font. We use the average results generated from the same 3 sets of random Euler angles as the final results of TC (%) and CA (%).	46
4.3	Results (%) of the ablation study on the coefficients of the loss functions. α and β indicate the importance of Dice loss and consistency loss, respectively. The best results are highlighted in bold font. We use the average results generated from the same 3 sets of random Euler angles as the final results of TC (%) and CA (%).	48
4.4	The comparative per-class results on the MFNet dataset. We use Acc (%) and IoU (%) for each class and the mAcc (%) and mIoU (%) for all the classes. The results demonstrate the superiority of our method, with the top two results in each column highlighted in bold and <u>underline</u> . The publication venue is followed by the publication year. Since CMX with MiT-B4 backbone has not released the Acc results based on its pre-trained weights, we use “-” to indicate the data absence.	50
4.5	The comparative per-class results based on RTFNet (Sun, Zuo, and M. Liu 2019). We use Acc (%) and IoU (%) for each class and the mAcc (%) and mIoU (%) for all the classes. We use ResNet-50 and ResNet-152 as backbone to compare RTFNet and our method. The results demonstrate the scalability of our method, with the best results of each backbone are highlighted in bold .	51

5.1	The results (%) of the ablation study on the training data. $1\times$ and $2\times$ indicate using the real or synthetic training data once or twice for training, respectively. The best results are highlighted in bold .	64
5.2	The results (%) of the ablation study on the maximum sampling number R_{max} . The best results are highlighted in bold .	65
5.3	The results (%) of the ablation study on the sampling mechanism. P_c is the pixel ratio of each class. μ_c is the mean loss of each class. A $\checkmark$ under P_c or μ_c indicates whether P_c or μ_c is included in the sampling mechanism for synthetic data. 'OverSamp' indicates whether the real training set is oversampled to ensure the training size of synthetic and real data are the same. The best results are highlighted in bold .	66
5.4	The per-class results (%) of models using different method to alleviate class imbalance. The best results are highlighted in bold .	68
5.5	The per-class results (%) of models using different training sets. We used RTFNet (Sun, Zuo, and M. Liu 2019), CMX (J. Zhang, Huayao Liu, et al. 2023), and CRM (Shin et al. 2024) with different backbones as the RGB-T segmentation models. Δ represents the mIoU improvement when training with both synthetic and real data compared to using only real data. The best results of each backbone are highlighted in bold .	69

6.1	The IoU (%) of different classes across various regions. Region 1 represents the innermost region, while region 9 represents the outermost region. CRM-B (Shin et al. 2024) is used as the segmentation network. For each class, the two lowest IoU values in different regions are denoted in bold and <u>underline</u> , respectively. .	83
6.2	Results (%) of the ablation study on standard deviation (σ_p) in Gaussian distribution. Var. is the variance of IoU across all regions (σ_{IoU}^2). mIoU is the IoU of the whole image. The best results are highlighted in bold	101
6.3	Results (%) of the ablation study on patch size (PS) in complementary masking generation. Var. is the variance of IoU across all regions (σ_{IoU}^2). mIoU is the IoU of the whole image. The best results are highlighted in bold	103
6.4	Results (%) of the ablation study on scaling factor (α) in spatial-weighted loss. Var. is the variance of IoU across all regions (σ_{IoU}^2). mIoU is the IoU of the whole image. The best results are highlighted in bold	104
6.5	Results (%) of different spatial weighting strategies on the loss function. Var. is the variance of IoU across all regions (σ_{IoU}^2). mIoU is the IoU of the whole image. The best results are highlighted in bold	104

- 6.6 The comparative per-class results (%) on the MFNet dataset. We report the Acc (%) and IoU (%) for each class and the mAcc (%) and mIoU (%) for all the classes. The results demonstrate the superiority of our method, with the top two results in each column highlighted in **bold** and underline. The publication venue is followed by the publication year. 105
- 6.7 The comparative per-class IoU (%) on the KP dataset. We report the IoU (%) for each class and the mIoU (%) for all the classes. The results demonstrate the superiority of our method, with the top two results in each column highlighted in **bold** and underline. The publication venue is followed by the publication year. 107

Chapter 1

Introduction

1.1 Background and Motivation

Scene understanding based on semantic segmentation is a key capability for autonomous driving, providing fundamental perception for downstream tasks such as localization (W. Ma, S. Huang, and Sun 2024; H. Xu, Huaping Liu, S. Huang, et al. 2024; H. Xu, Huaping Liu, Meng, et al. 2023), navigation (Y. Feng and Sun 2024; Teng et al. 2023; Y. Feng, Hua, and Sun 2023), and decision-making (Y. Feng, Hua, and Sun 2023). Although advances in deep learning have driven the development of semantic segmentation techniques based on RGB images (Ronneberger, Fischer, and Brox 2015; Badrinarayanan, Kendall, and Cipolla 2017; L.-C. Chen, Y. Zhu, et al. 2018), they still face significant challenges in practical applications. Due to the inherent limitations of visible light cameras, existing models can only capture rich textures under good lighting conditions, while their performance degrades significantly under complex lighting conditions such as nighttime or glare (Ha et al. 2017; Sun, Zuo, and M. Liu 2019), leading to scene understanding fail-

ure. These complex scenes are very common in real-world driving, and the failure of the perception system can pose serious safety risks. Therefore, achieving robust and accurate segmentation in various environments has become increasingly urgent.

Thermal imaging is a sensing method independent of visible light. It can effectively capture the thermal radiation information of objects, thus providing crucial supplementary information under complex lighting conditions that are difficult for RGB images to handle (Q. Zhang et al. 2021). Given the high similarity in spatial structure between thermal images and RGB images, RGB-Thermal (RGB-T) semantic segmentation has developed rapidly in recent years. Early research was mainly based on convolutional neural networks (CNNs), such as MFNet (Ha et al. 2017) and RTFNet (Sun, Zuo, and M. Liu 2019), which achieved acceptable segmentation performance through multi-modal fusion strategies. Subsequently, Transformer-based methods (Xie et al. 2021; J. Zhang, Huayao Liu, et al. 2023; Shin et al. 2024) established long-range contextual dependencies between pixels by treating image pixels as sequences, driving progress in this field. Among them, CMX (J. Zhang, Huayao Liu, et al. 2023), based on Segformer (Xie et al. 2021), achieved leading performance on the RGB-T semantic segmentation task. Nevertheless, RGB-T semantic segmentation still faces some key challenges that have not yet been resolved in the practical application of autonomous driving.

1.2 Challenges in RGB-T Semantic Segmentation

Despite significant progress in RGB-T semantic segmentation methods in recent years, they still face challenges in autonomous driving applications, including fea-

ture fusion design, temporal consistency, insufficient training data, and spatial performance imbalance.

1.2.1 Challenge I: Lack of Interpretability in Multi-Modal Fusion

Existing fusion methods typically ignore the differences in physical properties between RGB and thermal images under different lighting conditions. These methods often treat the network as a black box, implicitly learning the associations between modalities. However, RGB images provide rich texture details that thermal images lack, and thermal images are more robust to changes in lighting. In scenes with overexposure during the day (e.g., the sky), underexposure at night (e.g., low-light areas), or glare (e.g., from headlights), RGB information is easily degraded or lost (Song et al. 2023). In these situations, thermal images can provide more stable complementary features. Ignoring these physical characteristics in existing methods leads to a lack of interpretability in the fusion mechanism, making it difficult to fully exploit the potential of the two modalities.

1.2.2 Challenge II: Insufficient Temporal Consistency

Most current RGB-T semantic segmentation methods primarily optimize segmentation accuracy for single-frame images, neglecting the consistency of segmentation results across consecutive frames (Ha et al. 2017; Sun, Zuo, and M. Liu 2019; J. Zhang, Huayao Liu, et al. 2023; Shin et al. 2024). As shown in Fig. 1.1, existing methods may produce inconsistent segmentation outputs in adjacent frames, such as incorrectly identifying a vehicle as a bicycle, leading to a jump in semantic la-

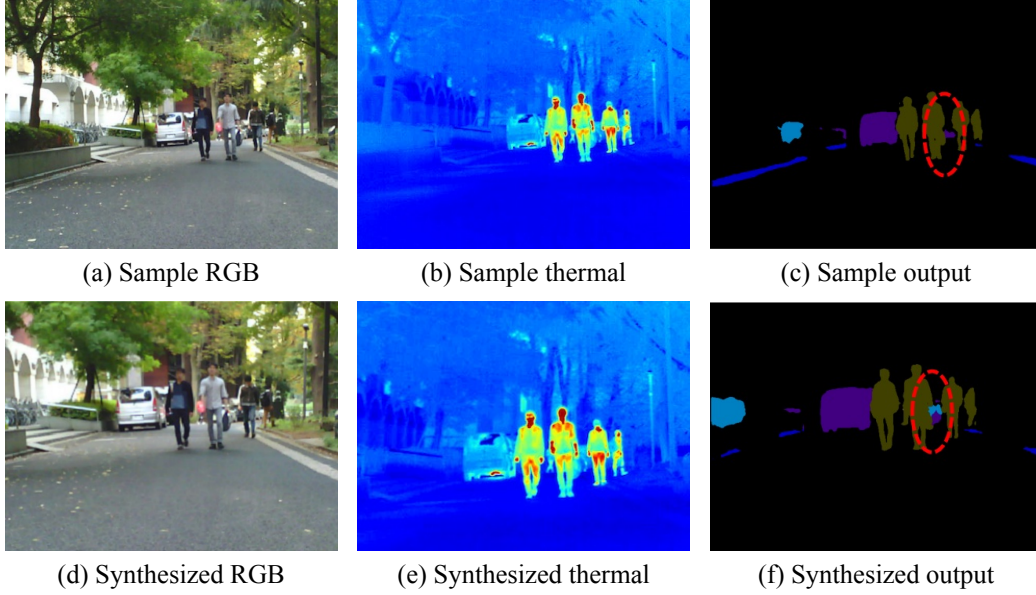


Figure 1.1: Inconsistent segmentation results across consecutive frames. The RGB-T images in the first row are sampled from the MFNet (Ha et al. 2017) dataset. The RGB-T images in the second row are the synthesized images at the next moment. The thermal images are visualized with the *jet* color map. We use the recent network CMX (J. Zhang, Huayao Liu, et al. 2023) to obtain the segmentation results. The blue and purple colors in the segmentation maps represent bike and car, respectively.

bels over time. This inconsistency can further affect the decision-making stability of autonomous driving systems, causing hesitation or misjudgment. To improve temporal consistency, some studies have attempted to introduce optical flow methods (Cheng et al. 2017; Nilsson and Sminchisescu 2018); however, under complex lighting conditions, optical flow estimation itself is prone to errors, thus limiting the applicability of such methods in real-world scenarios.

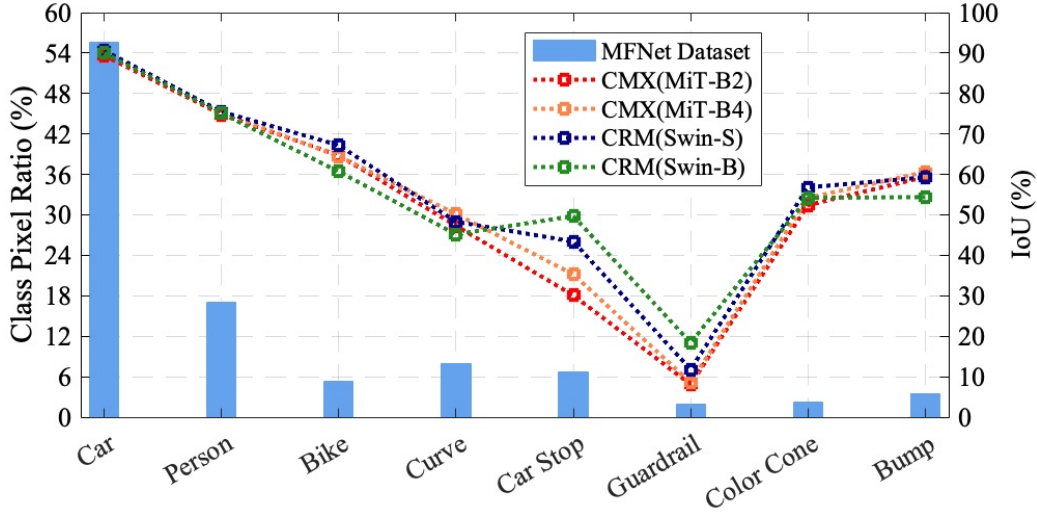


Figure 1.2: The pixel ratio of each class in the training set of MFNet dataset (Ha et al. 2017) and the IoU for each class of CMX (J. Zhang, Huayao Liu, et al. 2023) and CRM (Shin et al. 2024). We choose CMX using MiT-B2 and MiT-B4 as backbone, and CRM using Swin-S and Swin-B as backbone. This figure shows the correlation between the class pixel ratio in the training set and the IoU for that class.

1.2.3 Challenge III: Limited Training Data and Class Imbalance

In supervised RGB-T semantic segmentation tasks, the high cost of data annotation limits the size of existing datasets. For example, the largest available dataset, MFNet (Ha et al. 2017), contains only 1568 pairs of RGB-T images in its training set (including 784 flipped pairs). With limited data, simply increasing model complexity is insufficient to consistently improve segmentation performance. Furthermore, as shown in Fig. 1.2, existing datasets suffer from severe class imbalance. Analysis shows that the pixel ratio of each class is generally positively correlated with its segmentation performance: for example, the Car has the highest pixel ratio and the highest IoU, while the Guardrail has the lowest pixel ratio and the lowest IoU. However, there are exceptions: the Car Stop, with a higher pixel ratio, has

a lower IoU than the Bump. This indicates that segmentation performance is not only affected by the class pixel ratio but also closely related to the segmentation difficulty.

1.2.4 Challenge IV: Spatial Bias in Segmentation Performance

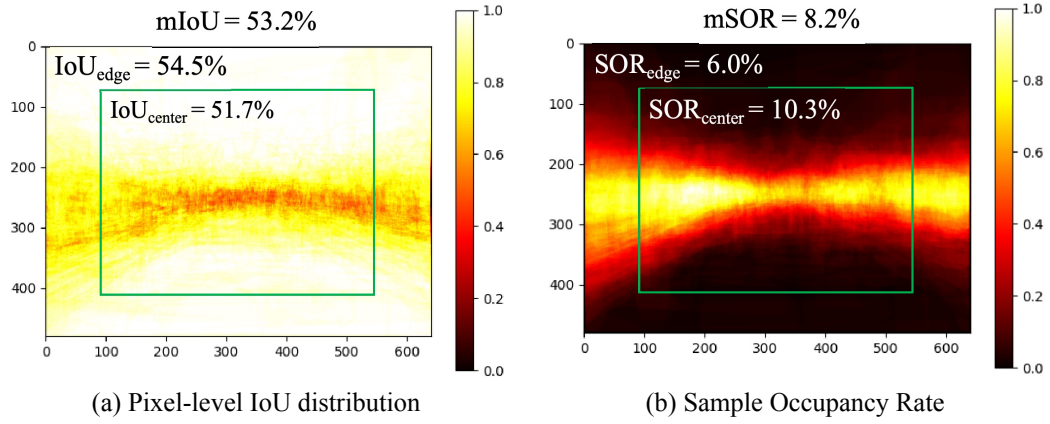


Figure 1.3: (a) Visualized heatmap of the pixel-level IoU distribution, based on segmentation results from the test images of the MFNet dataset. IoU_{center} and IoU_{edge} are respectively the IoUs at the central and edge regions of the image. $mIoU$ is the mean IoU over the whole image; (b) Visualized heatmap of the Sample Occupancy Rate (SOR), calculated from the training set of the MFNet dataset. SOR indicates that 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 the central and edge regions of the image. $mSOR$ is the mean SOR over the whole image.

While existing RGB-T semantic segmentation models perform well in overall accuracy, their segmentation performance varies across different image regions, exhibiting a "spatial bias". As shown in Fig. 1.3(a), the model typically performs worse in image central region than in edge region. It is noteworthy that the central region often contains targets crucial to autonomous driving safety, such as vehicles and pedestrians. Inaccurate segmentation would directly impact the reliability of

downstream tasks.

To explore the causes of spatial bias, we further analyzed the relationship between training data distribution and performance. Traditionally, class imbalance is attributed to an uneven distribution of training samples across different classes (Qiu et al. 2024; Tian et al. 2022). To verify whether spatial bias stems from a similar cause, we calculated the Sample Occupancy Rate (SOR) of target objects in the MFNet dataset, i.e., the probability of a target appearing at each pixel location in the training set. The results are shown in Fig. 1.3(b). Comparing Fig. 1.3(a) and (b), we find that regions with higher SOR (i.e., more training samples) do not show better segmentation performance. This phenomenon challenges the traditional assumption that "the more training samples, the better the segmentation performance," indicating that the causes of spatial bias are more complex.

1.3 Proposed Framework and Main Contributions

To systematically address four key challenges in RGB-T semantic segmentation, this dissertation proposes a comprehensive framework comprising four core components. Its main contributions are summarized below:

1. **Illumination-Guided Fusion Network (IGFNet):** To address the lack of interpretability in multi-modal fusion, we propose the Illumination-Guided Fusion Network (IGFNet). The core of this network is an Illumination Estimation Module (IEM), which generates weight masks based on pixel illumination values, explicitly modeling the importance of RGB information at each pixel location. Subsequently, the Illumination-Guided Cross-Modal Rectification Module (IGCM-RM) module adaptively corrects and fuses

RGB and thermal features based on these masks. This fusion mechanism based on physical illumination characteristics replaces the traditional "black box" fusion strategy. Experimental results on the MFNet dataset show that IGFNet achieves a 59.0% mIoU, laying an interpretable single-frame fusion foundation for understanding multi-modal complementarity.

2. **Temporal Consistency Framework:** To improve the consistency of segmentation results between consecutive frames, we propose a temporal consistency framework. This framework includes a Virtual View Image Generation (VVIG) module to generate images at the next moment, along with consistency loss functions to enhance temporal consistency of segmentation results. Furthermore, we introduce a balanced evaluation metric, Consistent Accuracy (CA), that measures both segmentation accuracy and consistency. Experiments demonstrate that the framework achieves a 59.9% mIoU, effectively improving temporal consistency and exhibiting good scalability across different network architectures.
3. **Synthetic Data Augmentation Method (SyntheticSeg):** To alleviate the scarcity of real-world RGB-T training data and the class imbalance problem, we propose a data-driven method, SyntheticSeg. This method utilizes a layout-to-image generative model to construct a large-scale, high-fidelity synthetic RGB-T dataset containing 31,360 image pairs. In addition, we design a novel sampling mechanism that simultaneously considers the scarcity and segmentation difficulty of each semantic layout, thereby intelligently sampling from the synthetic dataset. Experimental results show that SyntheticSeg achieves an mIoU of 60.4%, a 2.2% improvement over the base-

line method, particularly in rare and difficult-to-segment classes such as guardrail and car stop.

4. **Spatial Balancing Method:** We are the first to systematically study the spatial bias problem in RGB-T segmentation, finding that its root cause is not the number of samples, but rather the object complexity and image quality. Specifically, RGB images exhibit high target complexity and low signal-to-noise ratio (SNR) in central regions, while thermal images show low target complexity and high SNR ratio in the same regions. Based on this, we propose the Gaussian Guided Region Balance Masking (GRBM) method, which uses masks to rely more on reliable thermal features in image central regions. Furthermore, we introduce a spatial-weighted loss function to adjust the learning weights for different regions. Experiments verify that this method effectively alleviates spatial bias and improves overall segmentation performance.

In summary, these four contributions together build a comprehensive framework for advancing RGB-T semantic segmentation, and Fig. 1.4 provides an overview of them. Illumination-Guided Fusion (IGFNet) can also be applied to Temporal Consistency Enhancement to strengthen feature fusion for single-frame data. Meanwhile, Temporal Consistency Enhancement can improve segmentation consistency both for single-frame semantic segmentation and for spatial balancing segmentation. Synthetic data augmentation not only enhances the capability of single-frame RGB-T semantic segmentation but can also be combined with spatial balancing segmentation to further boost segmentation performance. Moreover, the analysis from spatial balancing segmentation can offer insights for synthetic

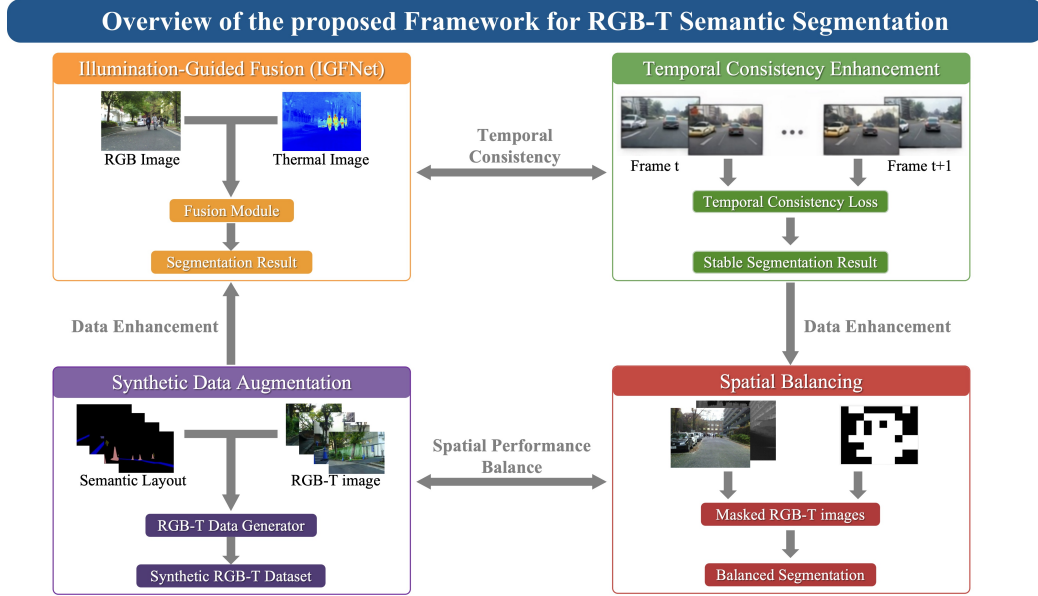


Figure 1.4: Overview of the Proposed Framework for RGB-T Semantic Segmentation.

data augmentation. If high-quality training data, particularly in the central regions of the image, can be generated, there is a potential to further improve the model's segmentation ability.

1.4 Dissertation Organization

The remainder of this dissertation is organized as follows:

Chapter 2 reviews relevant research in the RGB-T semantic segmentation, including CNN-based and transformer-based methods and their fusion strategies, temporal consistency methods in semantic segmentation, data augmentation and data generation techniques, and solutions to class imbalance and spatial bias problems.

Chapter 3 details the proposed Illumination-Guided Fusion Network (IGFNet), including its overall framework, the core Illumination Estimation Module (IEM),

and the Illumination-Guided Cross-Modal Rectification Module (IGCM-RM). Furthermore, this chapter presents extensive ablation experiments and comparative experiments on the MFNet dataset.

Chapter 4 proposes a temporal consistency framework, including the Virtual View Image Generation (VVIG) module, the consistency loss functions, and the Consistent Accuracy (CA) metric. Experimental results show that this method significantly improves temporal consistency performance while maintaining high segmentation accuracy.

Chapter 5 introduces the SyntheticSeg method, including the generation process of the synthetic dataset and a novel sampling mechanism that considers class scarcity and segmentation difficulty. Extensive experiments have validated the effectiveness of this proposed method.

Chapter 6 first systematically analyzes the causes of spatial bias, then introduce the Gaussian-guided Regional Balancing Masking (GRBM) method and the spatial-weighted loss function, and experimentally verify their effectiveness in mitigating spatial bias.

Chapter 7 summarizes the main contributions of this paper, discusses the limitations of current research, and provides an outlook on future research directions.

Chapter 2

Literature Review

This chapter reviews important literature in four related research areas, including: RGB-T semantic segmentation methods, temporal consistency enhancement approaches, data augmentation strategies, and solutions to imbalance problems.

2.1 RGB-Thermal Semantic Segmentation

In recent years, RGB-T semantic segmentation has attracted widespread attention due to its strong perceptual robustness under complex lighting conditions. Current methods can be divided into two categories: CNN-based and transformer-based, and various fusion strategies have emerged to effectively integrate complementary information from RGB and thermal images.

2.1.1 CNN-based Methods

Current research on RGB-T semantic segmentation primarily focuses on developing delicate models to efficiently fuse multi-modal data, thereby enhancing seg-

mentation performance. MFNet (Ha et al. 2017) and RTFNet (Sun, Zuo, and M. Liu 2019) both utilize dual-encoder architecture to fuse the feature maps. FuseSeg (Sun, Zuo, Yun, et al. 2021) continues the previous structure, and obtains better segmentation results through a two-stage fusion strategy. AFNet (J. Xu, K. Lu, and H. Wang 2021) is proposed to utilize attention to guide the fusion of RGB and thermal features. To minimize the differences between RGB and thermal modalities, ABMDRNet (Q. Zhang et al. 2021) utilizes a bi-directional image-to-image translation-based method. GMNet (W. Zhou, Jinfu Liu, et al. 2021) divides the feature extraction into three levels: junior, intermediate, and senior.

Several studies have explored the possibility of using other methods to enhance RGB-T segmentation performance. CAINet (Lv, Zhi Liu, and G. Li 2024) leverages auxiliary tasks and global context to enhance the complementary reasoning and detailed aggregation of multi-modal features. EGFNet (Dong et al. 2024) uses prior edge maps and multi-modal fusion modules to enhance the feature maps. Feng et al. (Z. Feng, Guo, and Sun 2023) also utilized edge information to enhance the segmentation performance of the thermal-only task.

2.1.2 Transformer-based Methods

Vaswani et al. (Vaswani et al. 2017) proposed an attention mechanism that only relies on self-attention to establish long-range dependencies among the input. DANet (J. Fu et al. 2019) utilizes self-attention to selectively build connections between local features and global dependencies to enhance the segmentation ability. CCNet (Zilong Huang et al. 2019) proposes criss-cross attention module to obtain dense and global contextual information. Dosovitskiy et al. (Dosovitskiy et al. 2020)

and Touvron et al. (Touvron et al. 2021) proposed Vision Transformer (ViT) and a teacher-student strategy based on attention, respectively. Furthermore, SETR (S. Zheng et al. 2021) divides the input image into a series of image blocks and uses a pure transformer structure to extract the global contextual features of the input image at different stages. Similarly, Segmenter (Strudel et al. 2021) utilizes a purely transformer-based approach to achieve semantic segmentation.

Segformer (Xie et al. 2021) proposes a hierarchical transformer structure for extracting and fusing multi-resolution features, and employs a lightweight decoder composed of a multilayer perceptron (MLP) for semantic label prediction. CMX (J. Zhang, Huayao Liu, et al. 2023) extends Segformer to multi-modal scenarios, effectively improving the performance of RGB-T semantic segmentation by aligning and enhancing visible light and thermal infrared features through a cross-modal feature correction module (CM-FRM). To further alleviate the model's over-reliance on a single modality, Shin et al. (Shin et al. 2024) proposed a complementary random masking strategy, combined with self-distillation loss, to enhance the synergistic effect of multi-modal features. Recently, Zhang et al. (J. Zhang, R. Liu, et al. 2023) further explored the feasibility of introducing other modal information into multi-modal semantic segmentation, and proposed a general segmentation framework applicable to RGB-X scenarios, expanding the application scope of modality fusion. Besides, Zhou et al. (H. Zhou et al. 2025) proposed a novel deformation-resilient multigranularity learning (DML) method that achieves robust feature alignment and multimodal complementarity without pixel-level registration. Zhou et al. (X. Zhou et al. 2025) proposed an adaptive gated fusion network (AGFNet) for RGB-T semantic segmentation, which effectively explores multimodal complementarity via a cross-modal adaptive gated-attention fusion

module, enhances spatial details through edge-guided mechanism.

2.1.3 Fusion Strategies

Various strategies (H. Li and Sun 2023; Lv, Zhi Liu, and G. Li 2024; W. Zhou, Dong, et al. 2023; J. Huang et al. 2024) have been proposed to utilize the complementary features of RGB and thermal images effectively. CACFNet (W. Zhou, Dong, et al. 2023) utilizes cross-modal attention and cascaded fusion to enhance the segmentation performance. MMSMCNet (W. Zhou, H. Zhang, et al. 2023) uses modal memory fusion and morphological multi-scale assistance to enhance cross-modal features. CAInet (Lv, Zhi Liu, and G. Li 2024) effectively captures and leverages the complementary relationships between RGB and thermal features through context-aware interactions. RoadFormer+ (J. Huang et al. 2024) extracts heterogeneous features from various modalities and merges the features from receptive fields of different scales. SegMiF (Jinyuan Liu et al. 2023) explicitly generates a fused RGB-T image before semantic segmentation, typically by learning a mapping that integrates RGB and thermal signals at the pixel level. While pre-fusion methods offer better interpretability and are beneficial for multi-task learning, they typically require building a complete image reconstruction process (such as an encoder-decoder architecture) before segmentation, which introduces significant computational overhead.

Existing methods primarily improve segmentation capabilities by constructing more advanced models; however, their fusion strategies are often considered a black box, resulting in a lack of interpretability. To overcome this limitation, our proposed IGFNet starts with the key physical factor, illumination conditions,

re-examining the feature essence of RGB-T images, and thus proposes an interpretable fusion mechanism guided by illumination.

2.2 Temporal Consistency in Semantic Segmentation

Temporal consistency of environmental perception is crucial for the reliable operation of autonomous driving system, and inconsistencies in prediction directly affect the stability of decision-making. However, existing RGB-T semantic segmentation methods generally focus on improving the segmentation accuracy of single-frame images, while neglecting the temporal consistency of prediction results between consecutive frames.

2.2.1 Optical Flow-based Methods

To ensure temporal consistency for semantic segmentation, Cheng et al. (Cheng et al. 2017) proposed a bi-directional framework to obtain the foreground segmentation and optical flow at the same time. The optical flow is used as the complementary information for the segmentation task. Nilsson et al. (Nilsson and Sminchisescu 2018) applied optical flow for video segmentation. They enhanced segmentation accuracy and consistency by utilizing the unlabeled RGB images in the dataset and propagating labels through optical flow. Zhang et al. (Y. Zhang et al. 2022) proposed AuxAdapt to improve the temporal consistency of RGB segmentation networks without using optical flow, by learning from the own decisions of the network and a small auxiliary network. However, in real-world applications, especially under challenging lighting conditions, it is difficult to compute accurate optical flow. Our proposed temporal consistency framework does not rely on op-

tical flow calculations. It can synthesize the image of the next moment by using the Virtual View Image Generation (VVIG) module. This feature makes it more applicable in complex and variable lighting environments.

2.2.2 Consistency Evaluation Metrics

The widely used intersection-over-union (IoU) is the standard evaluation metric for measuring segmentation accuracy on single frames. To evaluate segmentation consistency, Liu et al. (Y. Liu et al. 2020) used optical flow to warp the segmentation map of the current frame to align with the previous frame, and computed IoU between the warped segmentation map and the original segmentation map of the previous frame. They define this as the temporal consistency (TC) score. Zhang et al. (Y. Zhang et al. 2022) extended the TC metric to evaluate the segmentation consistency. Park et al. (Park et al. 2022) proposed a perceptual-consistency-based metric, which calculates the temporal consistency by comparing the average cosine similarity of the feature maps of consecutive frames.

However, existing consistency evaluation metrics only measure the consistency of segmentation results, regardless of the segmentation correctness. These metrics may fail to evaluate the performance when the segmentation results of consecutive frames produce the same errors for the same object. Our proposed Consistent Accuracy (CA) metric addresses this limitation by taking into account the ground truth of the segmentation result. The metric evaluates both the accuracy and the consistency of a segmentation network.

2.3 Data Augmentation and Generative Models

The scarcity of high-quality labeled data is a key bottleneck in training high-performance RGB-T segmentation models. In recent years, with the rapid development of generative models, augmenting the training set with synthetic data has become an important method to solving this problem.

2.3.1 Synthetic Data Generation

The introduction of Diffusion series networks (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021) has enabled the generation of high-quality images through a diffusion process that gradually transitions from noises to clear images. Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) establish a novel link between denoising score matching and Langevin dynamics. Dhariwal et al. (Dhariwal and Nichol 2021) demonstrated that diffusion models can surpass Generative Adversarial Networks (GANs) in image synthesis quality and diversity.

2.3.2 Layout-to-Image Generation

Building on the foundation of diffusion models, a series of works (W. Wang et al. 2022; Xue et al. 2023) propose layout-to-image generative models to generate RGB images based on semantic layouts. Wang et al. (W. Wang et al. 2022) proposed a framework that processes semantic layouts and noisy images differently, enhancing the quality and diversity of generated images. Xue et al. (Xue et al. 2023) explored image generation with unseen semantics using pre-trained text-to-image diffusion models.

However, these methods are designed for RGB images and cannot be directly applied to RGB-T images. The proposed SyntheticSeg method first employs a layout-to-image generation model to generate paired RGB and thermal images, constructing a large-scale synthetic RGB-T dataset to alleviate the data scarcity problem. To further improve model performance, we design a novel sampling mechanism that balances class scarcity and segmentation difficulty, thereby enhancing the learning of difficult samples in long-tailed distributions and effectively mitigating the class imbalance problem.

2.4 Addressing Imbalance in Semantic Segmentation

Class imbalance and spatial imbalance are two major challenges in semantic segmentation, both of which can lead to a decline in the segmentation performance.

2.4.1 Class Imbalance

The class imbalance problem (Tian et al. 2022; Qiu et al. 2024) 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.

To address class imbalance in object detection, Saez et al. (Sáez et al. 2015) enhanced SMOTE with an ensemble-based noise filter, addressing noisy and borderline examples in imbalanced datasets. Choi et al. (H. K. Choi, J. Choi, and H. J. Kim 2022) introduced a token-level data augmentation method for transformers

that is efficient and guided by attention. Venkataramanan et al. (Venkataramanan et al. 2022) presented a data augmentation technique that interpolates aligned features, combining the geometry of one image with the texture of another to enhance representation learning. However, Mixup series methods can produce unrealistic sample combinations, potentially leading the model to learn incorrect information.

In addition, specific loss functions (Lin et al. 2017; Cui et al. 2019) shift focus away from simple samples and emphasize the classification of more challenging ones. Importance-aware loss (B. Chen, Gong, and Yang 2019) prioritizes the accurate segmentation of critical objects by assigning different weights to classes based on their importance for driving safety. The Loss Max-Pooling (LMP) method (Rota Bulo, Neuhold, and Kotschieder 2017) adaptively re-weights pixel contributions based on observed losses to handle uneven training data distributions. Tian et al. (Tian et al. 2022) introduced Recall Loss, a new loss function to balance precision and recall in semantic segmentation. Qiu et al. (Qiu et al. 2024) proposed Subclassified Loss, which addresses class imbalance by focusing on subclasses.

Unlike methods that rely on optimizing loss functions, the SyntheticSeg method proposed in this dissertation addresses the issue at the data level to alleviate class imbalance. This method expands the dataset by generating synthetic data and strategically samples based on class scarcity and segmentation difficulty, thereby directly increasing the number and diversity of difficult samples in long-tailed distributions from the data source.

2.4.2 Spatial Imbalance and Bias

Although Zheng et al. (Z. Zheng et al. 2024) revealed a spatial bias in object detection tasks where model performance degrades in image edge regions, we observed the opposite phenomenon in RGB-T semantic segmentation for autonomous driving scenarios: the performance bottleneck significantly appears in image central regions. This unique spatial bias problem has not yet been systematically explored and resolved in current research.

2.4.3 Loss Function Designs

Several loss functions have been proposed to alleviate imbalance in semantic segmentation. The Hausdorff distance loss (Karimi and Salcudean 2020) optimizes boundary quality by applying a distance transformation to the ground truth and the predicted result. The Dice loss (Milletari, Navab, and Ahmadi 2016) alleviates class imbalance by maximizing the overlap between the prediction and the ground truth. The Combo Loss (Taghanaki et al. 2019) further integrates Dice and cross-entropy to address the challenge of input-output imbalance.

Unlike the above methods, the spatial-weighted loss function proposed in this dissertation introduces a spatial prior, dynamically assigning loss weights based on the distance between pixels and the image center. This mechanism encourages the model to maintain attention to key central regions while focusing more on learning edge regions with more stable features and lower noise.

Chapter 3

Illumination-Guided Fusion Network

3.1 Motivation

As discussed in Chapter 1, RGB-T semantic segmentation exhibits stronger robustness under complex lighting conditions. Thermal images can serve as an effective supplement to RGB images in low-light conditions, providing reliable information. However, as mentioned in Section 1.2, most existing fusion methods treat the network as a black box, implicitly learning intermodal correlations only through the design of fusion modules, while ignoring the inherent characteristics between RGB images and thermal images.

This chapter proposes an interpretable multi-modal fusion mechanism that uses illumination as a physical cue to explicitly model the importance of each region in RGB images, thereby guiding the fusion process. The basic principle is that in well-exposed areas, the RGB image is rich in texture information and should

be given higher weight in the fusion process; while in underexposed areas, RGB information degrades, and thermal features should be relied upon more. This interpretable fusion strategy effectively utilizes complementary information to improve the accuracy of semantic segmentation. Specifically, we propose an Illumination-Guided Fusion Network (IGFNet), whose core consists of an Illumination Estimation Module (IEM) and an Illumination-Guided Cross-Modal Rectification Module (IGCM-RM). The IEM maps the grayscale value of each pixel in the RGB image to a weight factor using a Gaussian filter, reflecting the importance of that pixel in the fusion process and providing guidance for subsequent fusion. The IGCM-RM uses this weight mask to recalibrate the multi-modal feature map at multiple stages, thereby achieving pixel-level adaptive fusion based on illumination conditions.

Chapter Organization. The remainder of this chapter is organized as follows. Section 3.2 provides an overview of the overall architecture and core components of IGFNet. Section 3.3 presents the experimental results, including ablation studies and performance comparisons on the MFNet dataset. Finally, Section 3.4 summarizes the main contributions of this chapter.

3.2 The Proposed Method

3.2.1 Network Overview

As illustrated in Fig. 3.1., we adopt the feature fusion structure comprising two encoders and one decoder. The two encoders are responsible for extracting features from the RGB and thermal images, respectively. In poorly exposed environ-

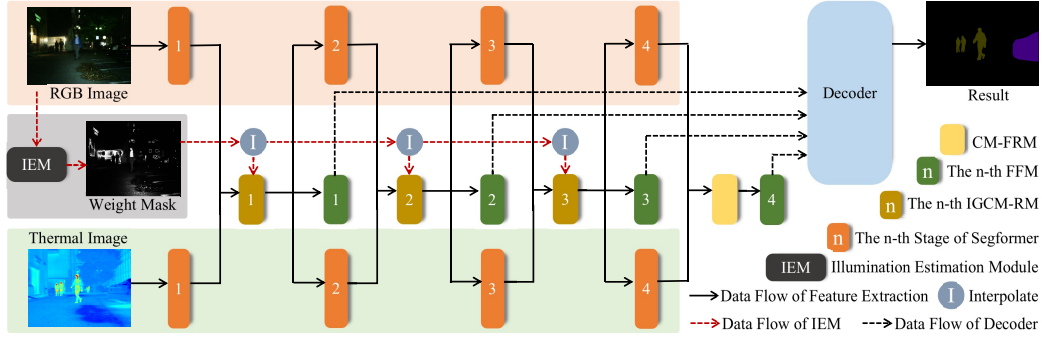


Figure 3.1: The overview of IGFNet for RGB-Thermal semantic segmentation. The inputs are an RGB image and a thermal image. IGFNet consists of a four-stage RGB encoder, a four-stage thermal encoder, a Illumination Estimation Module (IEM), a three-stage Illumination-Guided-Cross-Modal Rectification Module (IGCM-RM), a Cross-Modal Feature Rectification Module (CM-FRM), a four-stage Feature Fusion Module (FFM) and a decoder. The encoder and decoder are borrowed from Segformer (Xie et al. 2021). The CM-FRM and FFM are borrowed from CMX (J. Zhang, Huayao Liu, et al. 2023). The thermal image is colored with the *jet* color map.

ments, thermal images can provide complementary information to RGB images. By combining the information from both modalities, it is possible to achieve mutual correction and improve segmentation accuracy. Specifically, in extreme environments such as glare areas caused by vehicle headlights or areas with low ambient light at night, RGB images may lose texture information. In these cases, thermal images can provide complementary features to better understand the scene.

To exploit this property, we proposed the Illumination Estimation Module (IEM) to guide multi-modal fusion by generating an illumination-based weight mask. Subsequently, we used interpolation modules to hierarchical downsample the weight mask, and the results were applied to the following feature fusion modules. Then, we proposed the Illumination-Guided-Cross-Modal Rectification Module (IGCM-RM) to utilize the weight mask in guiding the rectification of both RGB and thermal features. In this manner, the rectified features are divided into

two streams, one stream is used as the input of the next stage encoder, while the other stream is used as input to the Feature Fusion Module (FFM). Since the weight mask essentially represents spatial attention, and the spatial information in high-level features decreases as the network depth increases, we only apply the IGCM-RM to the first three stages of the encoder. Then, we apply Cross-Modal Feature Rectification Module (CM-FRM) to the forth stage of the encoder.

3.2.2 Illumination Estimation Module

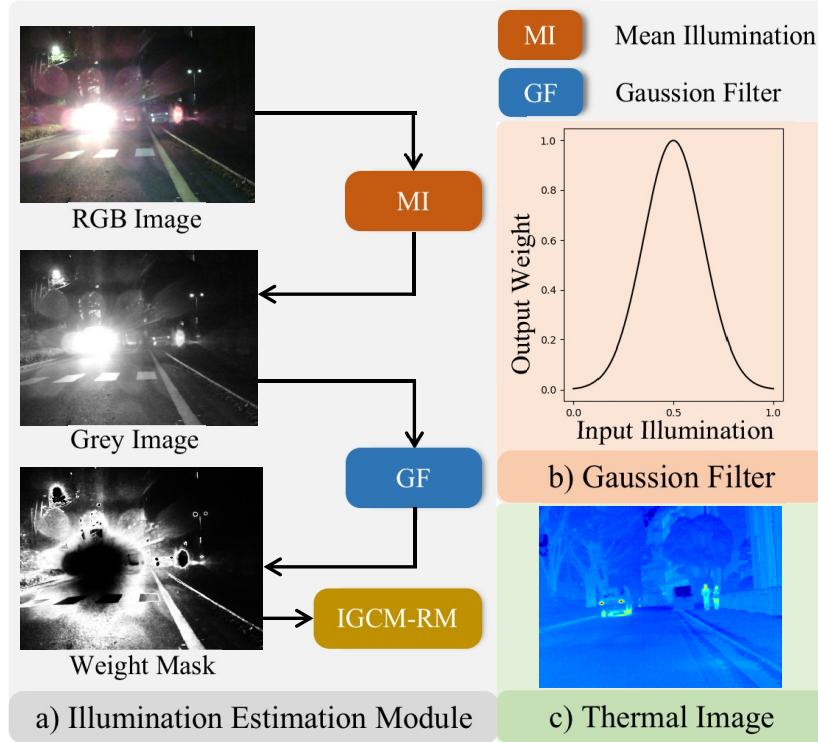


Figure 3.2: a) The pipeline of our proposed IEM. The input to IEM is an RGB image, from which we derive a grey image by calculating the mean illumination. We then obtain the weight mask by applying a Gaussian filter to this grey image. b) The graph illustrates the application of the Gaussian filter, mapping the grey image to the weight mask. c) The corresponding thermal image shows complementary information.

As previously discussed, RGB and thermal images provide distinct information under varying illumination conditions. To leverage this, we have developed the IEM to generate a weight mask indicating the importance of each RGB pixel based on illumination.

The pipeline of the IEM is shown in the part (a) of Fig. 3.2. The first step in our process involves calculating the average grayscale value of the three channels in the RGB image. This produces a grey image that reflects the illuminance values at different positions. Then, we propose a Gaussian filter to map the grayscale values of individual pixels to weights representing the importance of RGB information. The Gaussian function of this filter is described as follows:

$$f(x) = \exp\left(\frac{-(x - \mu)^2}{2\sigma^2}\right) \quad (1)$$

where $\mu = 0.5$ and $\sigma = 0.15$. As shown in the part (b) of Fig. 3.2, μ determines the midpoint position of the Gaussian function. We have chosen a value of $\mu = 0.5$, as this yields the best exposure and most abundant texture information in the RGB image. At this point, the corresponding output weight is 1, indicating that the RGB information is the most important. For instance, the RGB image in Fig. 3.2 shows the glare-affected area caused by vehicle headlights, and we cannot find any useful texture in this area. In contrast, the corresponding thermal image provides complementary information. The weight mask we have obtained maps the glare-affected area to 0 and the well-exposed areas to 1 according to the Gaussian distribution. This guides the network to primarily utilize thermal features in this area and RGB features in well-exposed areas during the feature fusion stage. Additionally, we can observe that the edges of the glare-affected area are assigned values close to 1

in the weight mask. This indicates that RGB information will be primarily utilized in feature fusion, facilitating accurate segmentation of object edges.

3.2.3 Illumination-Guided-Cross-Modal Rectification Module

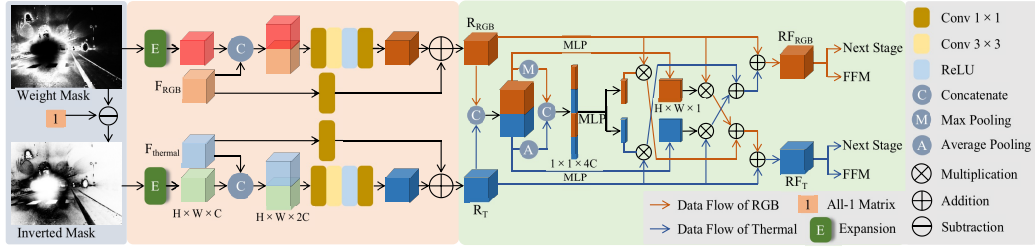


Figure 3.3: The structure of our proposed IGCM-RM. The inputs to this module are the feature maps of the RGB and thermal images from the encoder, as well as the weight mask generated by the IEM. After rectification guided by the weight mask, the rectified feature maps are divided into two streams. One stream serves as input to the subsequent stage encoder, while the other is input to the FFM for feature fusion. The part in green background is borrowed from CMX(J. Zhang, Huayao Liu, et al. 2023).

After obtaining the weight mask from the IEM, we proposed the IGCM-RM to recalibrate the features of RGB and thermal images, guided by the weight mask. The structure of the IGCM-RM is shown in Fig. 3.3. First, we obtain the inverted mask by subtracting the weight mask from a all-ones matrix. This inverted mask represents the importance of various positions within the thermal feature map during fusion.

For better illustration, we will use the stream of the weight mask $Mask \in \mathbb{R}^{H \times W \times 1}$ and the RGB feature map $F_{RGB} \in \mathbb{R}^{H \times W \times C}$ as an example. To ensure that the dimensions of the weight mask correspond with the feature maps, we expand the weight mask along the channel dimension to get $Mask \in \mathbb{R}^{H \times W \times C}$. Instead of performing an element-wise multiplication of $Mask \in \mathbb{R}^{H \times W \times C}$ and

$F_{RGB} \in \mathbb{R}^{H \times W \times C}$, we first concatenate them. Then, the resulting output is fed into a block, which consists of two 1×1 convolution layers, a 3×3 convolution layer, and a ReLU layer. The purpose of this block is to adaptively recalibrate the values of the feature map F_{RGB} , guided by the weight mask $Mask$. The output of this stream undergoes an element-wise addition with the output of another stream. This second stream originates from F_{RGB} and passes through a 1×1 convolution layer, with the goal of providing the original information of the RGB features. The outputs of this stage are $R_{RGB} \in \mathbb{R}^{H \times W \times C}$ and $R_T \in \mathbb{R}^{H \times W \times C}$.

The rectified RGB feature, $R_{RGB} \in \mathbb{R}^{H \times W \times C}$, which is embedded with illumination information, is then concatenated with the corresponding thermal feature, $R_T \in \mathbb{R}^{H \times W \times C}$. Subsequently, max pooling and average pooling are applied to the output features, $R_{RGB-T} \in \mathbb{R}^{H \times W \times 2C}$, respectively. Then, the concatenated output vectors, $Y \in \mathbb{R}^{4C}$, is fed into the MLP layer. The output of this MLP layer contain two channel weights, $C_{RGB} \in \mathbb{R}^C$ and $C_T \in \mathbb{R}^C$. Similarly, two spatial weights, $S_{RGB} \in \mathbb{R}^{H \times W \times 1}$ and $S_T \in \mathbb{R}^{H \times W \times 1}$, can be generated by another MLP layer. The output of IGCM-RM is formulated as:

$$RF_{RGB} = R_{RGB} + \lambda_1 C_{RGB} \otimes R_{RGB} + \lambda_2 S_{RGB} \odot R_{RGB}$$

$$RF_T = R_T + \lambda_1 C_T \otimes R_T + \lambda_2 S_T \odot R_T \quad (2)$$

where $\otimes$ denotes channel-wise multiplication, $\odot$ denotes spatial-wise multiplication, λ_1 and λ_2 are two hyper-parameters. Inspired by (J. Zhang, Huayao Liu, et al. 2023), we set $\lambda_1 = \lambda_2 = 0.5$. Finally, the rectified features, RF_{RGB} and RF_T , will be sent to the next stage encoder and the FFM for feature fusion.

3.3 Experimental Results and Discussions

3.3.1 Dataset

For our experiments, we utilized the available MFNet dataset released in (Ha et al. 2017). This dataset contains 9 classes (i.e., unlabelled background, car, person, bike, curve, car stop, guardrail, color cone, and bump), with manually annotated labels for semantic segmentation. The dataset contains 2,390 pairs of RGB and thermal images, derived from the original dataset consisting of 1,596 pairs. We employed the same split scheme as described in (Ha et al. 2017) to train our network, with 1568 pairs used for training, 392 for validation, and 393 for testing.

3.3.2 Implementation Details

Our IGFNet is implemented using PyTorch and trained and tested on a PC equipped with an NVIDIA RTX 3090 (24GB RAM). For the encoder, we utilize the Mix Transformer encoder (MiT-B2) pretrained on ImageNet (Russakovsky et al. 2015) as the backbone, while the decoder employs an MLP, both of which were proposed in SegFormer (Xie et al. 2021). We train the network using the AdamW optimizer (Kingma and Ba 2014), with a weight decay set to 0.01. The initial learning rate is set to $6e^{-5}$, and a poly learning rate schedule is employed. The momentum and the decay strategy are set to 0.9. The batch size is set to 4 during training and we use cross-entropy as the loss function. We use Precision (Pre), Accuracy (Acc), F1, Intersection over Union (IoU), Floating Point Operations (FLOPs) and Parameters (Params) to evaluate our network.

3.3.3 Ablation Study

Table 3.1: The results of the ablation study. '✓' means the IEM and IGCM-RM are used behind the n-th stage of the encoder. '—' means the CM-FRM (J. Zhang, Huayao Liu, et al. 2023) is used behind the n-th stage of the encoder. The best results are highlighted in bold font.

No.	Stage				mPre	mAcc	mF1	mIoU	FLOPs(G)	Params(M)
	1st	2nd	3rd	4th						
(A)	—	—	—	—	75.93	67.32	69.85	58.19	66.87	66.57
(B)	✓	—	—	—	73.25	72.67	70.39	58.46	67.54	66.60
(C)	—	✓	—	—	71.65	72.01	69.97	58.40	67.52	66.68
(D)	—	—	✓	—	73.88	71.36	69.96	58.24	67.86	67.29
(E)	—	—	—	✓	72.30	72.07	69.73	58.18	67.50	68.42
(F)	✓	✓	—	—	74.50	72.04	70.77	58.89	68.19	66.71
(G)	✓	✓	✓	—	73.92	72.91	71.04	59.01	69.18	67.44
(H)	✓	✓	✓	✓	73.03	72.18	70.21	58.47	69.82	69.29

We conducted an ablation study to evaluate the benefits of IEM and IGCM-RM, and subsequently selected the optimal structure for our IGFNet. In this section, we use CMX (J. Zhang, Huayao Liu, et al. 2023) based on MiT-B2 as the baseline. In the first part of our study, we positioned the two modules (IEM and IGCM-RM) behind the encoder at different stages. For instance, in No.B network, the two modules were placed behind the first stage encoder. Subsequently, the rectified RGB and thermal features were input into the FFM and the second stage encoder. Table 3.1 shows that, except for No.E, the segmentation performance (i.e., mF1 and mIoU) of all networks (i.e., No.B, No.C, and No.D) with IEM and IGCM-RM added is better than the baseline. Additionally, our results indicate that the network's performance gradually decreases as the stage at which the two

modules are inserted is delayed. This can be attributed to the fact that the weight mask obtained through IEM essentially represents spatial attention. As the stage of the encoder progresses, the spatial information contained within the high-level features learned by the network diminishes. Hence, we designed several combinations based on this characteristic in the second stage. For instance, network No.F placed IEM and IGCM-RM behind the first and second stages at the same time. As shown in Table 3.1, the network No.G obtains the best performance. In contrast, the performance of network No.H was found to be inferior to that of the network No.G. This observation is in alignment with our conclusion before. According to FLOPs and Params, it can be seen that adding IEM and IGCM-RM does not bring much computational complexity to the framework. Based on the results of our ablation study, we choose network No.G as the optimal structure of our IGFNet.

3.3.4 Comparative Study

We compared our proposed IGFNet with the state-of-the-art methods, including MFNet (Ha et al. 2017), RTFNet (Sun, Zuo, and M. Liu 2019), AFNet (J. Xu, K. Lu, and H. Wang 2021), ABMDRNet (Q. Zhang et al. 2021), FEANet (F. Deng et al. 2021), and CMX(MiT-B2) (J. Zhang, Huayao Liu, et al. 2023).

Table 3.2: The comparative per-class results on the MFNet dataset (Ha et al. 2017). To evaluate the methods, we utilized the IoU of each class and the mean IoU of all classes. The results demonstrate the superiority of our IGFNet, with the top two results in each column are highlighted in red and blue.

Method	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		
MFNet (Ha et al. 2017)	77.2	65.9	67.0	58.9	53.9	42.9	36.2	29.9	19.1	9.9	0.1	0.0	30.3	25.2	30.0	27.7	45.1	39.7
RTFNet (Sun, Zuo, and M. Liu 2019)	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
AFNet (J. Xu, K. Lu, and H. Wang 2021)	91.2	86.0	76.3	67.4	72.8	62.0	49.8	43.0	35.3	28.9	24.5	4.6	50.1	44.9	61.0	56.6	62.2	54.6
ABMDRNet (Q. Zhang et al. 2021)	94.3	84.8	90.0	69.6	75.7	60.3	64.0	45.1	44.1	33.1	31.0	5.1	61.7	47.4	66.2	50.0	69.5	54.8
FEANet (F. Deng et al. 2021)	93.3	87.8	82.7	71.1	76.7	61.1	65.5	46.5	26.6	22.1	70.8	6.6	66.6	55.3	77.3	48.9	73.2	55.3
CMX(MiT-B2) (J. Zhang, Huayao Liu, et al. 2023)	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
IGFNet(ours)	93.2	88.0	83.4	74.0	71.8	62.7	67.6	48.2	45.4	36.0	68.5	14.2	58.8	52.4	68.3	57.5	72.9	59.0

1) *Quantitative Results:* As shown in Table 3.2, we can find that the mIoU of our proposed IGFNet is higher than other methods. Specifically, our IGFNet

demonstrates a significant advantage over other methods in accurately segmenting the most challenging categories of car stop and guardrail. Additionally, it achieves the highest mIoU for the segmentation of curve. This shows that IGFNet significantly improves the segmentation performance of small and difficult-to-recognize objects by using illumination-guided feature fusion. Though FEANet achieves the highest Acc in the segmentation of guardrail and bump, its corresponding IoU do not show a corresponding performance. In contrast, our proposed IGFNet consistently maintains a highly competitive performance of both Acc and IoU across the segmentation of all classes.

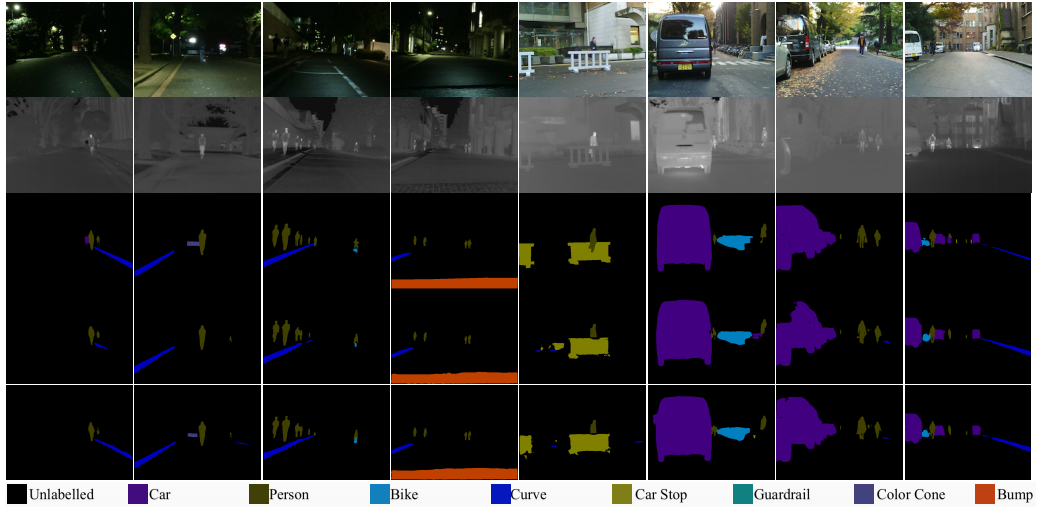


Figure 3.4: Sample qualitative demonstrations. The rows from top to bottom are RGB images, thermal images, ground truth, CMX results, and IGFNet results.

2) *Qualitative Results:* To intuitively demonstrate the effectiveness of our IGFNet, we selected four groups of daytime and nighttime images as examples, respectively. We compared them against the two best-performing methods, i.e., CMX(MiT-B2) and IGFNet, listed in Table 3.2. As shown in Fig. 3.4., the segmentation performance of our IGFNet on curve and car stop is significantly su-

perior to that of CMX. This is consistent with the quantitative results in the first part of Section 3.3.4, which demonstrate that the strategy of utilizing illumination to guide feature map fusion in IGFNet can effectively enhance segmentation accuracy, especially for objects that are difficult to distinguish.

3.4 Conclusions

We proposed here a novel semantic segmentation network IGFNet, which introduces an interpretable RGB-Thermal fusion network and utilizes illumination to guide the fusion of multi-modal features. The IEM employs a Gaussian filter to generate a weight mask, indicating the significance of each RGB pixel based on the illumination. Then, the generated weight mask is utilized as a cue within the IGCM-RM to rectify the multi-modal features, enabling their more effective fusion. The experimental results show that our proposed IGFNet achieves better performance than the state-of-the-art methods. However, there are still some limitations in our IGFNet. For instance, we found that the semantic segmentation results of consecutive frames lack consistency, which can lead to hesitation or misjudgment in autonomous driving decisions. Therefore, we would like to solve this issue in the future work.

Chapter 4

Temporal Consistency Framework

4.1 Motivation

In Chapter 3, we proposed IGFNet, which effectively improves the segmentation accuracy of single-frame images through an interpretable fusion strategy. However, its design does not consider the consistency of segmentation results between consecutive frames. In autonomous driving scenarios, temporal consistency is crucial. Inter-frame jitter in segmentation results can cause disorder in the decision-making system and even lead to serious consequences. For example, if a target correctly identified as a "pedestrian" in a previous frame is misidentified as a "car" in a subsequent frame, it will directly cause the downstream planning module to generate contradictory instructions. As mentioned in Section 1.2, existing RGB-T segmentation methods generally focus on improving single-frame accuracy but neglect in-depth exploration of temporal consistency.

To address the aforementioned issues, this chapter focuses on how to maintain high single-frame accuracy while ensuring temporal consistency in RGB-T se-

mantic segmentation. Specifically, we designed a Virtual View Image Generation (VVIG) module that uses a virtual view transformation matrix to simulate camera motion, and then synthesizes the virtual view of the next frame from the current frame. Secondly, we propose two consistency loss functions to impose temporal constraints on the segmentation results of consecutive frames. Furthermore, to comprehensively evaluate segmentation performance, we introduce a new metric which simultaneously measures segmentation accuracy and temporal consistency, thus overcoming the limitation of existing metrics that only focus on single-frame performance.

Chapter Organization. The remainder of this chapter is structured as follows: Section 4.2 provides an overview of the proposed temporal consistency framework, covering the VVIG module, the consistency loss functions, and the new evaluation metric. Section 4.3 systematically verifies the effectiveness of the method through ablation experiments and comparative experiments. Finally, Section 4.4 summarizes the entire chapter.

4.2 The Proposed Method

4.2.1 The Framework Overview

To study segmentation consistency, we need datasets with temporal-sequential images. But the existing RGB-T dataset, such as the MFNet dataset (Ha et al. 2017), includes only discrete image frames. So, we propose the virtual view image generation (VVIG) module to synthesize the sequential frames based on the existing dataset. The overview of our proposed temporal-consistent framework is shown in

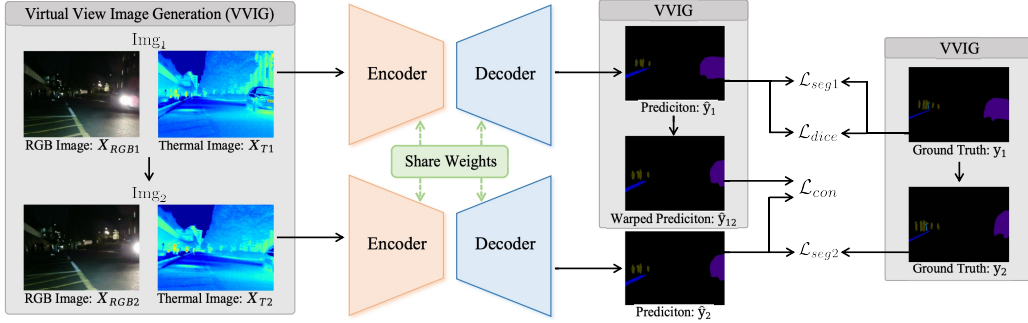


Figure 4.1: The overview of our temporal-consistent framework for RGB-T semantic segmentation. Img_1 is sampled from MFNet. Img_2 is generated using our virtual view image generation (VVIG) module. The encoders and decoders are borrowed from CMX (J. Zhang, Huayao Liu, et al. 2023). The warped prediction $\hat{y}_{12}$ and ground truth y_2 are also generated by the VVIG module. The *jet* color map is used here to visualize the thermal images.

Fig. 4.1. Img_1 , which represents the current frame, is the real image sampled from the MFNet (Ha et al. 2017). We use the VVIG module to synthesize Img_2 , which is the frame at the next moment. We choose CMX (J. Zhang, Huayao Liu, et al. 2023) as the segmentation network. It utilizes a unified fusion approach with a Cross-Modal Feature Rectification Module (CM-FRM) to calibrate bi-modal features. Additionally, a Feature Fusion Module (FFM) for long-range context exchange, achieving state-of-the-art performance across various RGB-X modalities (J. Zhang, Huayao Liu, et al. 2023). In this work, the encoders and decoders for the two frames are the same. They share the same weights. The CMX network can be replaced with other RGB-T semantic segmentation networks.

4.2.2 Virtual View Image Generation Module

As aforementioned, we propose the VVIG module to synthesize the image at the next moment. Fig. 4.2 shows the pipeline. To mimic the camera’s viewpoint at the next moment, we need to simulate both its rotation and translation. For rotation,

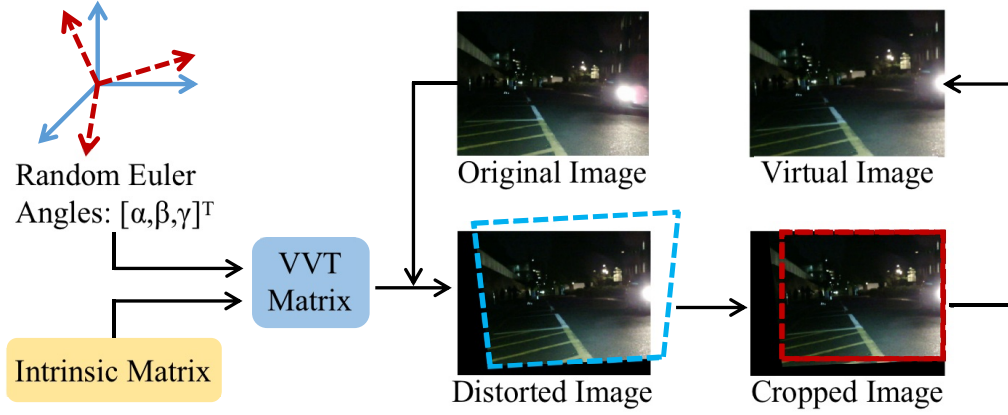


Figure 4.2: The pipeline of our VVIG module. α , γ , β are respectively the randomly generated Euler angles about the $\hat{Z}$, $\hat{Y}$, $\hat{X}$ axes. The VVT matrix (i.e., M_{VVT}) is generated from the Euler angles and the intrinsic matrix. The blue dashed box on the transformed image represents the shape of the original image after the M_{VVT} transformation. The red dashed box on the cropped image shows the largest inner rectangle within the valid range for cropping. The virtual image is obtained from the cropped image by interpolation.

we randomly generate the Euler angles α , γ , and β about the $\hat{Z}$, $\hat{Y}$, and $\hat{X}$ axes in the image coordinate system. The ranges of the angles are: $\alpha \in [-5^\circ, 5^\circ]$, $\gamma \in [-10^\circ, 10^\circ]$, $\beta \in [-10^\circ, 10^\circ]$. We use the Euler angles and the camera intrinsic matrix to calculate the virtual view transformation (VVT) matrix, which is then used to transform the original image. M_{VVT} is found by:

$$M_{\text{VVT}} = K R_z(\alpha) R_y(\beta) R_x(\gamma) K^{-1}, \quad (4.1)$$

where K represents the camera intrinsic matrix. The VVT matrix becomes equivalent to a homography matrix when the camera motion involves only rotations, without any translations. Detailed derivation and descriptions are presented in the Appendix.

After simulating rotation with the VVT matrix and obtaining the distorted im-

age, we discard the parts that appear beyond the field-of-view of the image, indicated by the white areas in the blue dashed box. The remaining area within the blue dashed box is then cropped using the red dashed box to its largest inner rectangle, defining the effective pixel area of the virtual image. Finally, the cropped image is resized to match the original image’s resolution by using interpolation to produce the virtual image. These cropping and resizing operations effectively simulate the forward movement of the camera. To maximize the protection of edge information in objects, we utilize bilinear interpolation for RGB and thermal images, and nearest-neighbor difference for ground truth during warping and resizing operations. As shown in Fig. 4.1, we feed an RGB image, a thermal image, a ground-truth image, and the prediction of Img_1 into the VVIG module. The module then generates the corresponding Img_2 .

4.2.3 The Loss Functions

As shown in Fig. 4.1, the framework conducts semantic segmentation on Img_1 and Img_2 separately. We adopt the cross-entropy loss (i.e., $\mathcal{L}_{seg1}$ and $\mathcal{L}_{seg2}$) for the semantic segmentation task. In addition, we apply the Dice loss (Milletari, Navab, and Ahmadi 2016) (i.e., $\mathcal{L}_{dice}$) to further improve the segmentation accuracy.

We propose two consistency loss functions (i.e., $\mathcal{L}_{con}$ and $\mathcal{L}_{con-acc}$) to improve the segmentation consistency across consecutive frames. Specifically, $\mathcal{L}_{con}$ transforms the segmentation map of Img_1 to Img_2 through M_{VVT} , leading to $\hat{y}_{12}$, which is then compared with the segmentation map of Img_2 , $\hat{y}_2$. The greater the similarity between $\hat{y}_{12}$ and $\hat{y}_2$, the better the segmentation consistency.

First, we have to transform the segmentation map of the current frame to the

position of next frame by $\hat{y}_{12} = M_{\text{VVT}}\hat{y}_1$, where M_{VVT} is used as the true value. Then, $\mathcal{L}_{\text{con}}$ is calculated to measure the inconsistency between the consecutive frames:

$$\mathcal{L}_{\text{con}} = 1 - \frac{1}{C} \sum_{c=1}^C \frac{2 \sum_{h=1, w=1}^{H \times W} \hat{y}_{12}^{(c, h, w)} \cdot \hat{y}_2^{(c, h, w)}}{\sum_{h=1, w=1}^{H \times W} (\hat{y}_{12}^{(c, h, w)} + \hat{y}_2^{(c, h, w)} + \sigma)}, \quad (4.2)$$

where h , w and c denote the row, column and class indices of predictions, respectively. $H \times W$ denotes the number of pixels and C denotes the number of classes. σ is a very small positive number that prevents the denominator from being zero. In this paper, we set $\sigma = 1 \times 10^{-7}$.

Considering that $\mathcal{L}_{\text{con}}$ only constrains the consistency between the segmentation maps of consecutive frames, if the segmentation maps make the same incorrect prediction for the same pixel, the loss function cannot constrain them effectively. So, we propose $\mathcal{L}_{\text{con-acc}}$ based on $\mathcal{L}_{\text{con}}$. This loss function incorporates the ground truth of the segmentation map of Img_2 , y_2 , and measures the consistency among $\hat{y}_{12}$, $\hat{y}_2$, and y_2 :

$$\mathcal{L}_{\text{con-acc}} = 1 - \frac{1}{C} \sum_{c=1}^C \frac{2 \sum_{h=1, w=1}^{H \times W} \hat{y}_{12}^{(c, h, w)} \cdot \hat{y}_2^{(c, h, w)} \cdot y_2^{(c, h, w)}}{\sum_{h=1, w=1}^{H \times W} (\hat{y}_{12}^{(c, h, w)} + \hat{y}_2^{(c, h, w)} + y_2^{(c, h, w)} + \sigma)}, \quad (4.3)$$

where y_2 is used as a mask to improve the segmentation accuracy.

4.2.4 The Evaluation Metric

Intersection-over-Union (IoU) is a widely-used metric that evaluates segmentation accuracy for a single frame. Temporal consistency (TC) (Y. Liu et al. 2020) is proposed to evaluate the consistency of RGB segmentation across consecutive frames.

While TC can capture the segmentation consistency for consecutive frames, it fails to account for scenarios where consecutive frames make the same incorrect prediction for the same target.

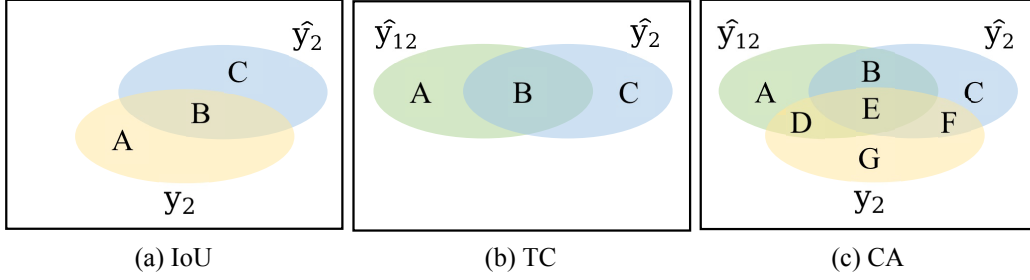


Figure 4.3: (a) Intersection over union (IoU): $y_2 = A + B$, represents the ground truth of Img_2 ; $\hat{y}_2 = B + C$, represents the prediction of Img_2 . (b) Temporal consistency (TC): $\hat{y}_{12} = A + B$, represents the prediction of Img_1 mapped to Img_2 by M_{VVT} ; $\hat{y}_2 = B + C$, represents the prediction of Img_2 . (c) Consistent accuracy (CA): $\hat{y}_{12} = A + B + D + E$, represents the prediction of Img_1 mapped to Img_2 by M_{VVT} ; $\hat{y}_2 = B + C + E + F$, represents the prediction of Img_2 ; $y_2 = D + E + F + G$, represents the ground truth of Img_2 .

To address this issue, we propose consistent accuracy (CA), which evaluates segmentation results in terms of both consistency and accuracy. Fig. 4.3 shows the schematic diagrams of IoU, TC and CA. According to the definition of IoU, it is calculated as $\text{IoU} = \frac{B}{A+B+C} \times 100\%$, indicating the similarity between the prediction and the ground truth. Liu et al. (Y. Liu et al. 2020) proposed TC to measure the segmentation consistency of RGB images. As shown in Fig. 4.3(b), TC is defined as $\text{TC} = \frac{B}{A+B+C} \times 100\%$, which measures the similarity between the current frame, $\hat{y}_2$, and the warped previous frame, $\hat{y}_{12}$, by optical flow. As aforementioned, it is challenging to calculate optical flow under unsatisfactory lighting conditions. So, we warp the current frame using M_{VVT} and utilize the ground truth y_2 as a supervision signal to enhance the segmentation accuracy. As shown in Fig. 4.3(c), CA is defined as $\text{CA} = \frac{E}{A+B+C+D+E+F+G} \times 100\%$. $\text{CA} = 100\%$

only if the predictions of consecutive frames are consistent and match the ground-truth labels. Otherwise, CA is decreased by both inconsistency and inaccuracy.

4.3 Experimental Results and Discussions

4.3.1 Datasets

We used the MFNet dataset (Ha et al. 2017) for our experiments. The dataset consists of 2,390 pairs of RGB-T images. It has 9 classes (i.e., unlabelled background, car, person, bike, curve, car stop, guardrail, color cone, and bump). We used the same split scheme as (Ha et al. 2017) to train our network, that is, 1568 pairs for training, 392 pairs for validation, and 393 pairs for testing.

4.3.2 Training Details

We implement our proposed method in PyTorch and train the networks with an NVIDIA RTX 3090 (24GB RAM) graphics card. We employ the CMX (J. Zhang, Huayao Liu, et al. 2023) network, utilizing the same dual-stream encoder to extract features from RGB and thermal modalities, and a decoder that integrates these features. Specifically, we adopt the four-stage Mix Transformer (MiT) encoder, pre-trained on ImageNet (J. Deng et al. 2009). This hierarchically structured Transformer encoder can avoid interpolation of positional codes and hence obtain multi-scale features. We choose MiT-B2 as the backbone to trade-off the performance and computational expenses. The decoder is a multi-layer perceptron (MLP) with an embedding dimension of 512, as proposed in SegFormer (Xie et al. 2021). We train the network with the AdamW optimizer, using a weight decay rate

of 1×10^{-3} . The initial learning rate is 6×10^{-5} , and we use the poly learning rate schedule (L.-C. Chen, Papandreou, et al. 2017). We use mean precision (mPre), mean accuracy (mAcc), mean F1 (mF1), mean intersection over union (mIoU), TC and CA for the quantitative evaluations.

4.3.3 Ablation Study

4.3.3.1 Ablation on the VVIG Module

Table 4.1: Results (%) of the ablation study on the VVIG module with different losses. Img_1 and Img_2 indicate the losses applied to the current frame and next frame, respectively. The best results are highlighted in bold font. We use the average results generated from the same 3 sets of random Euler angles as the final results of TC (%) and CA (%).

No.	Losses		mPre	mAcc	mF1	mIoU	TC	CA
	Img_1	Img_2						
(A)	$\mathcal{L}_{seg1}$	—	75.93	67.32	69.85	58.19	34.58	22.29
(B)	$\mathcal{L}_{seg1}$	$\mathcal{L}_{seg2}$	75.41	70.37	70.06	58.76	34.52	22.34
(C)	$\mathcal{L}_{seg1}$	$\mathcal{L}_{seg2} + \mathcal{L}_{dice2}$	74.91	73.86	71.21	59.47	34.69	22.74
(D)	$\mathcal{L}_{seg1}$	$\mathcal{L}_{seg2} + \mathcal{L}_{con}$	74.66	70.07	70.57	58.84	35.86	22.89
(E)	$\mathcal{L}_{seg1}$	$\mathcal{L}_{seg2} + \mathcal{L}_{con-acc}$	74.03	69.74	70.15	59.02	34.58	22.61
(F)	$\mathcal{L}_{seg1} + \mathcal{L}_{dice1}$	$\mathcal{L}_{seg2}$	72.90	74.53	72.47	60.33	34.30	22.70
(G)	$\mathcal{L}_{seg1} + \mathcal{L}_{dice1}$	$\mathcal{L}_{seg2} + \mathcal{L}_{dice2}$	76.93	71.04	72.99	60.78	32.35	22.65
(H)	$\mathcal{L}_{seg1} + \mathcal{L}_{dice1}$	$\mathcal{L}_{seg2} + \mathcal{L}_{con}$	74.01	74.17	71.46	59.90	36.40	23.29
(I)	$\mathcal{L}_{seg1} + \mathcal{L}_{dice1}$	$\mathcal{L}_{seg2} + \mathcal{L}_{con-acc}$	74.49	72.26	71.75	60.01	34.41	22.77

To demonstrate the effectiveness of the VVIG module, we use CMX (J. Zhang, Huayao Liu, et al. 2023) based on MiT-B2 as the baseline, denoted as variant (A) in Tab. 4.1. Variant (A) only uses the cross-entropy loss, which is the same as (J. Zhang, Huayao Liu, et al. 2023). As shown in Tab. 4.1, variants (B) to (I)

all employ the VVIG module to generate the virtual view images to improve the segmentation consistency. To improve the segmentation accuracy, we use the Dice loss $\mathcal{L}_{dice}$ and two cross-entropy losses, $\mathcal{L}_{seg1}$ and $\mathcal{L}_{seg2}$ for the segmentation of Img_1 and Img_2 , respectively. According to the definitions given in Eq. (4.2) and Eq. (4.3), we use $\mathcal{L}_{con}$ and $\mathcal{L}_{con-acc}$ to represent our proposed consistency losses for the segmentation of Img_2 .

From Tab. 4.1, despite achieving around 60% mIoU, all of the variants have a low TC consistency less than 40%. Our proposed consistency evaluation metric CA is even lower than 25%. This shows that existing methods, despite having good accuracy for single frames, suffer from poor consistency. From the results of variants (B) to (I), we find that the VVIG module can enhance the segmentation accuracy and consistency, resulting in higher mF1, mIoU, TC and CA than those of variant (A). The results of variants (C) and (F) show that the Dice loss can greatly enhance the segmentation accuracy, achieving much higher mF1 and mIoU. However, despite the significant improvement in mIoU for variant (F) over variant (A), the TC decreases. The reason may be that TC only measures segmentation consistency between consecutive frames without considering segmentation accuracy. On the other hand, our proposed CA follows the same trend as mIoU, since it incorporates the ground truth of the segmentation map as a supervision signal, and the CA reaches 100% only when the segmentation results of consecutive frames are perfectly accurate and consistent.

By comparing the results of variants (C), (D) and (E), we find that $\mathcal{L}_{dice}$ greatly enhances the segmentation accuracy, $\mathcal{L}_{con}$ greatly enhances the segmentation consistency, and $\mathcal{L}_{con-acc}$ achieves a balance between segmentation accuracy and consistency. From variant (F), we use $\mathcal{L}_{dice}$, $\mathcal{L}_{con}$ and $\mathcal{L}_{con-acc}$ on Img_2 , resulting

in variants (G), (H) and (I), respectively. Although variant (G) has the highest mIoU, its TC and CA are the lowest. Notably, TC of variant (G) is even lower than that of variant (A), indicating that $\mathcal{L}_{dice}$ reduces segmentation consistency despite its improvement in segmentation accuracy. While the mIoU of variant (H) is slightly lower than that of variant (G), it shows a significant improvement over that of the baseline variant (A). Importantly, variant (H) achieves the highest TC and CA, demonstrating that $\mathcal{L}_{con}$ can substantially enhance segmentation consistency while improving segmentation accuracy. Compared to variant (H), variant (I) achieves higher segmentation accuracy but lower segmentation consistency. Variant (I) represents a compromise between variants (G) and (H), as it incorporates the ground truth of the segmentation map as a constraint in $\mathcal{L}_{con-acc}$. Our experimental results and analyses indicate that, in semantic segmentation task, there is an inherent trade-off between segmentation accuracy and consistency. The above experiments demonstrate, emphasizing more on the segmentation accuracy would hinder consistency to some degree. Segmentation consistency is essential in practical applications. So, in this work, we adopt segmentation consistency as an overall evaluation metric, by which variant (H) is the best method.

To better analyze the impact of simulating rotation and translation in the VVIG module, an ablation study is conducted, the results of which are displayed in Tab. 4.2. Variant (A) represents the results without the VVIG module, variant (B) represents the results with the VVIG module simulating only rotation, variant (C) represents the results with the VVIG module simulating only translation, and variant (D) represents the results using the VVIG module that simulates both rotation and translation. The results indicate that simulating rotation significantly enhances mIoU, thereby improving segmentation accuracy. Meanwhile, simulating transla-

Table 4.2: Results (%) of the ablation study on the components of the VVIG module. Rot. and Tra. indicate the rotation and translation of the VVIG module, respectively. The best results are highlighted in bold font. We use the average results generated from the same 3 sets of random Euler angles as the final results of TC (%) and CA (%).

No.	VVIG		mPre	mAcc	mF1	mIoU	TC	CA
	Rot.	Tra.						
(A)			75.93	67.32	69.85	58.19	34.58	22.29
(B)	✓		73.47	74.87	71.55	59.75	34.58	22.61
(C)		✓	70.93	74.50	70.74	58.89	35.47	22.71
(D)	✓	✓	74.01	74.17	71.46	59.90	36.40	23.29

tion notably improves TC and CA, thereby enhancing segmentation consistency. Simultaneously simulating both rotation and translation leads to great improvements in both segmentation accuracy and consistency, thereby demonstrating the importance of simulating both rotation and translation in the VVIG module.

4.3.3.2 Ablation on Evaluation Metric

To further analyze the segmentation consistency performance of our proposed variants (G), (H), and (I), we conduct a comparative analysis in terms of the TC and CA across various classes within the MFNet dataset (Ha et al. 2017), as depicted in Fig. 4.4. TC measures the temporal consistency of the semantic segmentation results between consecutive frames. As shown in Fig. 4.4, variant (H) achieves the highest TC for most classes, except for person, where variant (I) performs better. It is noteworthy that variant (H) shows a significant improvement in TC for the car stop and guardrail compared to other variants, indicating that it has a greater effect on the segmentation consistency of challenging objects.

Contrary to the TC metric, the CA metric employs the ground truth of seg-

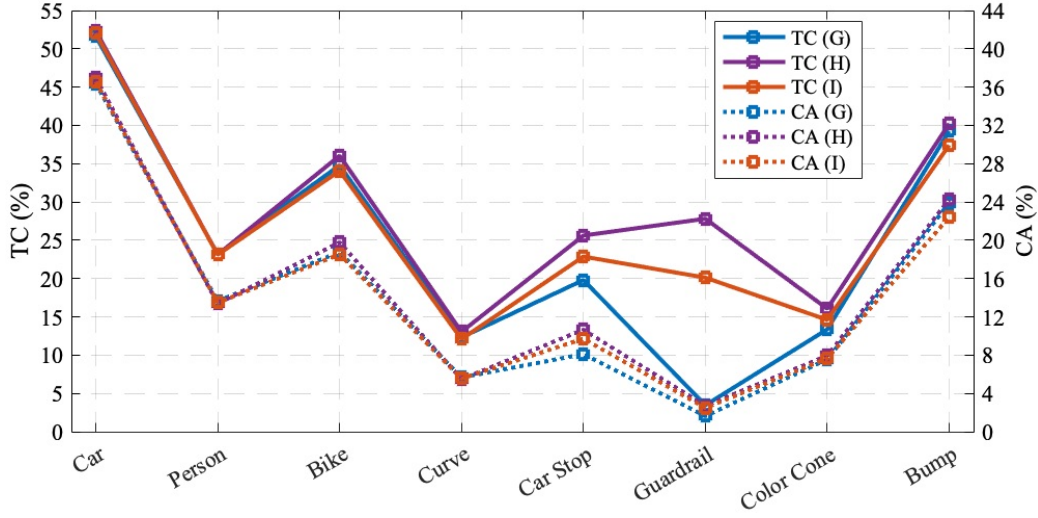


Figure 4.4: The ablation study results (%) of temporal consistency (TC) and consistent accuracy (CA). The horizontal axis shows each class in the MFNet dataset (Ha et al. 2017). Variants (G), (H) and (I) are the three variants of our method. The figure illustrates the superiority of variant (H) in terms of segmentation consistency. The figure is best viewed in color.

mentation map as the supervision signal. This approach ensures that, in scenarios where segmentation results across consecutive frames are consistent but inaccurate, CA does not yield relatively high values as observed with TC. This distinction is exemplified in the results for the car stop and guardrail depicted in Fig. 4.4. Despite the equal ranking of the three variants in TC and CA for both car stop and guardrail, the gap between the three in TC is significantly larger than that in CA. This implies that evaluating solely the consistency of segmentation results introduces considerable uncertainty, particularly when the segmentation results of consecutive frames exhibit the same error for a specific object. So, by leveraging the truth value of segmentation map, CA can more accurately reflect the segmentation performance. Variant (H) achieves the highest CA in almost all classes, indicating that its segmentation results are both accurate and consistent.

4.3.3.3 Ablation on Coefficients of Loss Functions

From the results in Tab. 4.1, we select variant (H) as the best method. It uses the cross-entropy loss and Dice loss on Img_1 , and cross-entropy loss and our proposed consistency loss on Img_2 . Among them, the Dice loss focuses on improving the segmentation accuracy of the current frame, while the consistency loss focuses on improving segmentation consistency. The total loss is calculated as:

$$\mathcal{L}_{total} = \mathcal{L}_{seg1} + \alpha \mathcal{L}_{dice} + \mathcal{L}_{seg2} + \beta \mathcal{L}_{con}, \quad (4.4)$$

where α and β denote the importance of Dice loss and consistency loss, respectively. We set α and β to 0.5, 1.0, and 2.0 in turn, to investigate their influences on the segmentation results.

Table 4.3: Results (%) of the ablation study on the coefficients of the loss functions. α and β indicate the importance of Dice loss and consistency loss, respectively. The best results are highlighted in bold font. We use the average results generated from the same 3 sets of random Euler angles as the final results of TC (%) and CA (%).

No.	Coefficients		mPre	mAcc	mF1	mIoU	TC	CA
	α	β						
(A)	0.5	0.5	73.70	73.07	71.27	59.56	35.02	22.87
(B)	0.5	1.0	73.20	71.26	70.09	58.63	34.94	22.53
(C)	0.5	2.0	71.26	74.81	71.47	59.12	36.36	23.06
(D)	1.0	0.5	71.37	77.53	71.78	59.89	35.22	22.76
(E)	1.0	1.0	74.01	74.17	71.46	59.90	36.40	23.29
(F)	1.0	2.0	71.44	74.51	70.76	58.74	36.31	23.10
(G)	2.0	0.5	72.99	72.54	70.47	58.85	34.75	22.49
(H)	2.0	1.0	73.58	73.64	71.48	59.85	35.70	23.01
(I)	2.0	2.0	70.23	73.00	69.90	58.20	36.16	22.84

As shown in Tab. 4.3, variant (E) obtains the highest mIoU, TC, and CA, in-

dicating that the method achieves the best segmentation accuracy and consistency when $\alpha = 1.0$ and $\beta = 1.0$. Comparing variants (G) and (I), we find that variant (I) has much higher TC and CA but much lower mIoU than variant (G). This means that variant (I) achieves better consistency but worse accuracy, indicating that a higher β can enhance segmentation consistency while suppressing accuracy. Meanwhile, variant (G) achieves higher accuracy but at the cost of consistency, indicating that overemphasizing segmentation accuracy may hinder consistency, which is consistent with the conclusion from the previous results. The comparison between variants (A) and (C), and variants (D) and (F) leads to the same conclusion. This also demonstrates that balancing segmentation accuracy with consistency is a challenging problem and worth further investigating. In summary, when $\alpha = 1$ and $\beta = 1$, our method achieves the highest mIoU, TC, and CA at the same time, which indicates that our method and our loss function effectively balance segmentation accuracy and consistency.

4.3.4 Comparative Experiments

We adopt CMX with MiT-B2 as the segmentation network for our framework. Based on different loss function strategies, we select variants (G), (H) and (I) in Tab. 4.1 for comparison, naming them Ours-dice, Ours-con and Ours-con-acc. They all use MiT-B2 as the backbone. We compare our method with RTFNet (Sun, Zuo, and M. Liu 2019), ABMDRNet (Q. Zhang et al. 2021), CENet (Z. Feng, Guo, and Sun 2023), CACFNet (W. Zhou, Dong, et al. 2023), MMSMCNet (W. Zhou, H. Zhang, et al. 2023), EAEFNet (Liang et al. 2023), CMX (J. Zhang, Huayao Liu, et al. 2023), CMNeXt (J. Zhang, R. Liu, et al. 2023), CAINet (Lv, Zhi Liu, and G. Li

2024), and EGFNet (Dong et al. 2024). CMX utilizes both MiT-B2 and MiT-B4 as backbones, while CMNeXt uses MiT-B4 as its backbone. Since CMX with MiT-B2 has not reported its per-class Acc, and CMNeXt with MiT-B4 has not reported its per-class Acc and IoU, we get the missing results by testing the networks with their pre-trained weights. However, CMX with MiT-B4 does not provide any pre-trained weights, so we could not compare its Acc with the other methods.

Table 4.4: The comparative per-class results on the MFNet dataset. We use Acc (%) and IoU (%) for each class and the mAcc (%) and mIoU (%) for all the classes. The results demonstrate the superiority of our method, with the top two results in each column highlighted in **bold** and underline. The publication venue is followed by the publication year. Since CMX with MiT-B4 backbone has not released the Acc results based on its pre-trained weights, we use “-” to indicate the data absence.

Method	Backbone	Venue	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 (Sun, Zuo, and M. Liu 2019)	ResNet-152	RAL'19	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
ABMDRNet (Q. Zhang et al. 2021)	ResNet-50	CVPR'21	94.3	84.8	90.0	69.6	75.7	60.3	64.0	45.1	44.1	33.1	31.0	5.1	61.7	47.4	66.2	50.0	69.5	54.8
CENet (Z. Feng, Guo, and Sun 2023)	ResNet	RAL'23	92.0	85.8	78.9	70.0	74.9	61.4	64.8	46.8	39.8	29.3	<u>65.7</u>	8.7	54.1	47.8	77.1	56.9	71.8	56.1
CACFNet (W. Zhou, Dong, et al. 2023)	ConvNeXt-B	TIV'23	95.9	89.2	93.6	69.5	82.0	63.3	74.0	46.6	49.0	32.4	45.8	7.9	<u>69.8</u>	54.9	<u>82.1</u>	58.3	76.7	57.8
MMSMCNet (W. Zhou, H. Zhang, et al. 2023)	MiT-B3	TCSVT'23	<u>96.2</u>	89.2	<u>93.2</u>	69.1	83.4	63.5	<u>74.4</u>	46.4	56.6	41.9	26.9	8.8	70.2	48.8	77.5	57.6	75.2	58.1
EAEFNet (Liang et al. 2023)	ResNet-152	RAL'23	95.4	87.6	85.2	72.6	79.9	63.8	70.6	48.6	47.9	35.0	62.8	<u>14.2</u>	62.7	52.4	71.9	58.3	75.1	58.9
CMX (J. Zhang, Huayao Liu, et al. 2023)	MiT-B2	TITS'23	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
CMX (J. Zhang, Huayao Liu, et al. 2023)	MiT-B4	TITS'23	-	<u>90.1</u>	-	<u>75.2</u>	-	64.5	-	<u>50.2</u>	-	35.3	-	8.5	-	54.2	-	60.6	-	59.7
CMNeXt (J. Zhang, R. Liu, et al. 2023)	MiT-B4	CVPR'23	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 (Lv, Zhi Liu, and G. Li 2024)	MobileNet-V2	TMM'24	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 (Dong et al. 2024)	ConvNeXt	TITS'24	96.5	89.8	92.1	71.6	<u>84.8</u>	63.9	76.1	46.7	44.6	31.3	38.7	6.7	71.1	52.0	78.1	57.4	<u>75.6</u>	57.5
Ours-dice	MiT-B2		93.1	88.8	89.2	74.5	76.4	64.5	69.3	47.8	<u>52.0</u>	36.6	25.6	22.0	64.7	53.7	70.0	<u>60.9</u>	71.0	60.8
Ours-con	MiT-B2		93.7	89.1	88.5	74.7	79.4	<u>65.6</u>	70.6	46.3	44.2	37.6	62.2	11.1	62.7	54.5	67.1	62.0	74.2	59.9
Ours-con-acc	MiT-B2		94.0	88.9	86.7	75.4	76.6	63.4	67.0	47.0	46.7	<u>39.0</u>	49.3	12.6	61.3	55.7	69.6	59.9	72.3	<u>60.0</u>

As shown in Tab. 4.4, our proposed methods achieve the highest mIoU among the state-of-the-art methods. Specifically, compared with CMX using the same MiT-B2 backbone, our methods achieve a 1.7% to 2.6% higher mIoU. Compared with CMNeXt that uses the larger backbone MiT-B4, our methods also increase mIoU by 0.2% to 1.1%. Moreover, the mAcc of Ours-con with MiT-B2 is 6.9% higher than that of CMX with MiT-B2. Despite achieving the highest mAcc, CACFNet (W. Zhou, Dong, et al. 2023) suffers from a low mIoU, which implies a high rate of false positives. Therefore, we can conclude that our method not only improves the segmentation consistency but also achieve the best overall accuracy compared to other methods.

To demonstrate the scalability of our method to other RGB-T semantic segmentation networks, we choose the well-known CNN-based RTFNet (Sun, Zuo,

Table 4.5: The comparative per-class results based on RTFNet (Sun, Zuo, and M. Liu 2019). We use Acc (%) and IoU (%) for each class and the mAcc (%) and mIoU (%) for all the classes. We use ResNet-50 and ResNet-152 as backbone to compare RTFNet and our method. The results demonstrate the scalability of our method, with the best results of each backbone are highlighted in **bold**.

Method	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 (Sun, Zuo, and M. Liu 2019)	ResNet-50	91.3	86.3	78.2	67.8	71.5	58.2	59.8	43.7	32.1	24.3	13.4	3.6	40.4	26.0	73.5	57.2	62.2	51.7
Ours-con		91.3	87.5	85.8	71.1	73.9	60.6	61.7	43.8	35.0	22.5	43.3	3.8	45.4	38.4	70.0	54.5	67.3	53.3
RTFNet (Sun, Zuo, and M. Liu 2019)	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
Ours-con		91.5	86.5	83.3	71.3	74.5	61.7	67.5	47.0	43.4	28.4	14.1	2.4	45.7	41.2	73.2	50.5	65.8	54.1

and M. Liu 2019) as our segmentation network and keep the parameter quantities unchanged. Tab. 4.5 compares the segmentation results of the original RTFNet and Ours-con based on different backbones. With ResNet-50 as the backbone, Ours-con outperforms RTFNet in segmentation accuracy (Acc and IoU) for all classes except car stop and bump, and increases mAcc and mIoU by 5.1% and 1.6%, respectively. With ResNet-152 as the backbone, Ours-con significantly increases the Acc of guardrail from 0.0% to 14.1%, and the IoU of color cone from 29.1% to 41.2%, indicating that our method can significantly enhance the segmentation accuracy for challenging objects. Ours-con with ResNet-152 also achieves better segmentation accuracy that increases mAcc and mIoU by 2.7% and 0.9%, respectively.

4.3.5 Qualitative Demonstrations

Fig. 4.5 qualitatively demonstrates sample segmentation results. In each set of images, the left column is the original current frame from the MFNet (Ha et al. 2017) dataset, and the right column is the synthesized frame generated by our method. The segmentation consistency can be seen by comparing the consecutive frames. We can see that our three methods generally outperform CMX with MiT-B2 in terms of segmentation accuracy, especially for small objects such as car stops (refer to the 3rd and 4th columns). We can also find that Ours-con achieves

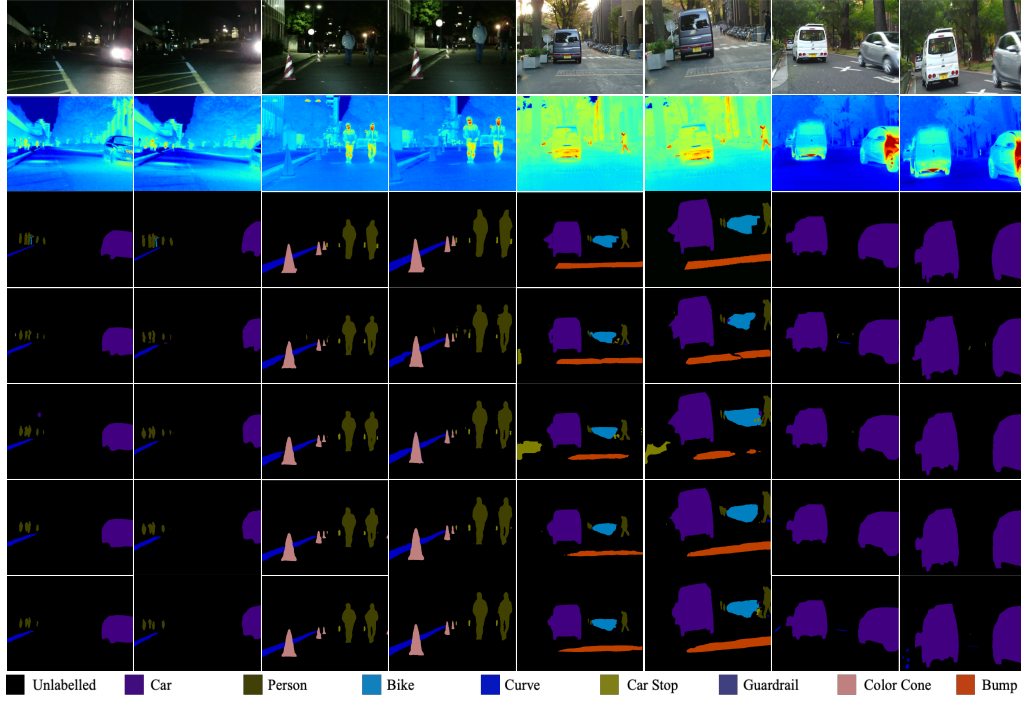


Figure 4.5: Qualitative demonstration for RGB-T semantic segmentation on the MFNet (Ha et al. 2017) dataset. The rows from top to bottom are RGB images, thermal images, ground truth, CMX results, Ours-dice results, Ours-con results, and Ours-con-acc results. All of the variants use MiT-B2 as backbones. The 1st and 2nd columns are the 1st set of images, the 3rd and 4th columns are the 2nd set of images, the 5th and 6th columns are the 3rd set of images, and the 7th and 8th columns are the 4th set of images. Among them, the first two sets and the last two sets are the images of nighttime and daytime, respectively.

the most consistent results. Moreover, the results of all three of our methods are more consistent than those of CMX. Specifically, the segmentation results of CMX show significant inconsistencies for all sets of images. Similarly, the segmentation results of Ours-dice for 1st, 2nd, and 3rd sets are also inconsistent. The same is observed for Ours-con-acc in the 2nd, 3rd, and 4th sets. Although Ours-con presents the lowest mIoU among our methods, it shows much better consistency than the other methods. This result is consistent with the findings from the ablation

experiments presented in Tab. 4.1.

Ours-con, which uses the proposed consistency loss, achieves the best segmentation accuracy and has much better consistency than the other methods. So, in real applications, Ours-con could be preferable to Ours-dice, which sacrifices consistency for accuracy.

4.4 Conclusions and Future Work

This letter presents the viewpoint that consistency should be valued in RGB-T semantic segmentation in addition to accuracy. To this end, we proposed a temporal-consistent framework for consistent and accurate RGB-T semantic segmentation. The proposed framework includes a VVIG module, which can synthesize a virtual frame at the next moment. Moreover, the proposed consistency loss improves segmentation consistency without compromising accuracy. The proposed metric is able to evaluate segmentation results in terms of both accuracy and consistency. The experimental results show that our method outperforms the state-of-the-art networks.

Despite these contributions, our framework focuses primarily on short-term consecutive frames and does not model complex long-term temporal dynamics such as object occlusion or multi-object interaction. Additionally, the VVIG module is designed for pairwise image generation rather than continuous video streams, which limits its direct applicability to real-time autonomous driving scenarios.

In future work, we will address these limitations by exploring novel architectures for long-sequence modeling to capture spatio-temporal dependencies across multiple frames. Furthermore, we aim to extend the framework to video-level con-

sistency by developing end-to-end models capable of generating and segmenting continuous video streams in real time.

Chapter 5

Synthetic Data Augmentation

5.1 Motivation

While the multi-modal fusion interpretability (Chapter 3) and temporal consistency (Chapter 4) discussed in previous chapters have effectively improved segmentation accuracy, these methods all rely on sufficient training data. As discussed in Section 1.2, RGB-T semantic segmentation is facing a severe data shortage. Specifically, the largest publicly available dataset, MFNet, contains only 1568 pairs of training images. This limited data size, coupled with severe class imbalance, jointly constrains the performance of existing models.

Specifically, the data shortage manifests in two ways: First, the absolute number of training samples is insufficient to support the training of deep neural networks, easily leading to overfitting and causing the model to memorize samples rather than learn generalizable features; second, there is an imbalance in the number of samples among different classes. Figure 1.2 shows that the class pixel ratio is highly correlated with segmentation performance. For example, the segmen-

tation accuracy of rare classes such as Guardrail and Car Stop is significantly lower than that of common classes such as Car. It is worth noting that although the pixel ratio of the Car Stop class is higher than that of Bump, its IoU is actually lower, indicating that segmentation difficulty also plays a crucial role.

This chapter aims to alleviate the problems of data scarcity and class imbalance in RGB-T semantic segmentation through synthetic data augmentation. We propose the SyntheticSeg method, whose core idea is to construct a large-scale synthetic RGB-T dataset using layout-to-image generation techniques and design an adaptive sampling mechanism that comprehensively considers class scarcity and segmentation difficulty. Specifically, we first improve the layout-to-image generation model FreestyleNet, enabling it to simultaneously generate RGB and thermal images based on semantic layouts, constructing a large-scale synthetic dataset containing over 31,000 image pairs to enhance data diversity and suppress overfitting. Second, we propose a novel sampling strategy that calculates a layout score by evaluating the scarcity of each semantic layout and segmentation difficulty, and determines its sampling frequency accordingly, thereby prioritizing the generation of synthetic data containing rare or difficult-to-segment samples. Furthermore, we oversample the real training set to balance the scale of synthetic and real data and avoid low-quality synthetic samples dominating the training process.

Chapter Organization. The reminder of this chapter is structured as follows: Section 5.2 describes the overall process of the SyntheticSeg framework and the details of constructing the synthetic dataset, and analyzes the sampling mechanism and its data distribution under different maximum sampling numbers. Section 5.3 verifies the effectiveness of the proposed method through comprehensive experiments. Section 5.4 summarizes the core contributions of this chapter and discusses

its limitations and future research directions.

5.2 The Proposed Method

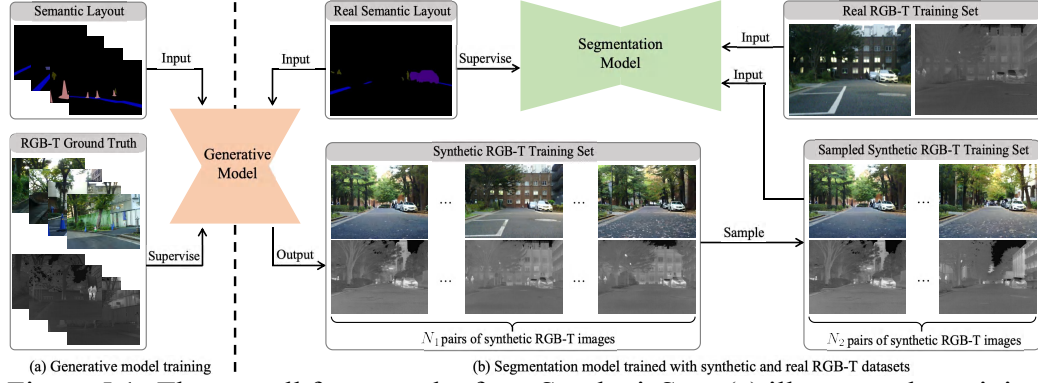


Figure 5.1: The overall framework of our SyntheticSeg. (a) illustrates the training process of RGB-T generative model; (b) illustrates the pipeline for joint training of segmentation models using both synthetic and real RGB-T datasets.

5.2.1 The Overall Framework

The MFNet dataset (Ha et al. 2017) contains only 2,353 fully annotated pairs (including 784 flipped pairs) of RGB-T images, with just 1,568 pairs in the training set. As networks used for RGB-T semantic segmentation grow more powerful, the limited scale of the dataset has become a critical factor affecting segmentation performance. With the advent of generative models like Diffusion (Dhariwal and Nichol 2021), it is now feasible to expand the training data by creating high-fidelity synthetic images.

We adapt the layout-to-image FreestyleNet (Xue et al. 2023) as our generative model, and the framework of our SyntheticSeg is illustrated in Fig. 5.1. FreestyleNet is trained and validated on COCO-Stuff (Caesar, Uijlings, and Fer-

rari 2018) and ADE20K (B. Zhou et al. 2017) to generate RGB images from semantic layouts. To apply it to the RGB-T image generation task on the MFNet dataset (Ha et al. 2017), we separately feed the RGB and thermal images from the MFNet dataset into FreestyleNet. The thermal images are first normalized and then converted into 3-channel images to match the input dimensions expected by FreestyleNet. Then, we modify the class indices of the dataset to define the number of output classes of the generative model.

As illustrated in Fig. 5.1 (b), we can feed the semantic layout from the real training set into the trained RGB-T generative model to obtain the corresponding RGB-T images. The semantic layout fed into the model and the generated RGB-T images form a pair of synthetic data suitable for training. In principle, for one semantic layout, we can generate an infinite number of corresponding RGB-T images by varying the random seed. So, if the real dataset contains N_0 pairs of training data, we can generate $N_0 \times N_1$ pairs of synthetic data, where N_1 is the number of selected seeds. This method can significantly expand the scale of fully annotated data available for training. To alleviate the class imbalance problem, we propose a metric to assess the scarcity and segmentation challenge of each semantic layout in the real dataset. Based on this metric, we determine N_2 , the number of synthetic data to be sampled for each real semantic layout. In this way, we have one pair of real RGB-T image and N_2 pairs of synthetic RGB-T images for each semantic layout.

Then, we feed RGB-T images from both the real and sampled synthetic datasets into the existing segmentation model, perform feature extraction and fusion on the RGB and thermal images, and decode them to obtain the segmentation results.

5.2.2 Sampling Mechanism

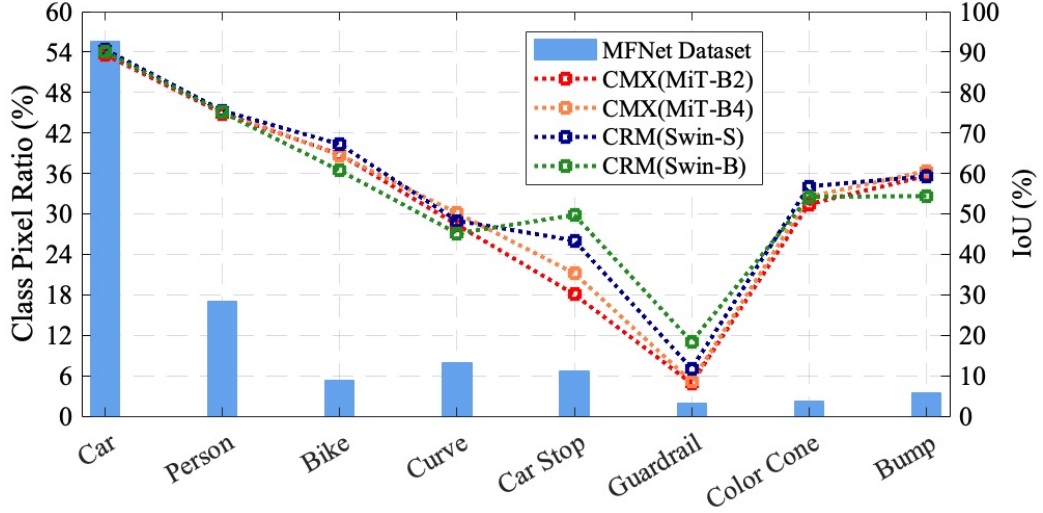


Figure 5.2: The pixel ratio of each class in the training set of MFNet dataset (Ha et al. 2017) and the IoU for each class of CMX (J. Zhang, Huayao Liu, et al. 2023) and CRM (Shin et al. 2024). We choose CMX using MiT-B2 and MiT-B4 as backbone, and CRM using Swin-S and Swin-B as backbone. This figure shows the correlation between the class pixel ratio in the training set and the IoU for that class.

As shown in Fig. 5.2, the segmentation performance of different classes in RGB-T semantic segmentation is related to their respective pixel ratio and segmentation difficulty. To improve the segmentation performance of classes with low pixel ratio and high segmentation difficulty, we design a metric to evaluate the scarcity and segmentation difficulty of each semantic layout in the training set. This metric determines the number of corresponding synthetic RGB-T images sampled from the generated synthetic dataset for that semantic layout. This sampling mechanism can improve the number and diversity of uncommon and challenging samples.

To evaluate the scarcity and segmentation challenge of each semantic layout, we first assess the pixel ratio and segmentation difficulty of the different classes.

In the MFNet dataset, for each class c ($c \in \{1, 2, \dots, C\}$, where C is the total number of classes excluding the background class 0), the pixel ratio of the class P_c can be defined as:

$$P_c = \frac{\sum_{i=1}^M |\Omega_c|}{\sum_{c=1}^C \sum_{i=1}^M |\Omega_c|}, \quad (5.1)$$

where M is the total number of images in the training set and Ω_c is the set of pixels belonging to class c in the layout. So, the numerator represents the sum of pixels for class c across all images, while the denominator represents the sum of pixels for all classes except the background across all images.

Then, we define μ_c as the class-wise mean loss for class c , representing the corresponding segmentation challenge. The μ_c can be defined as:

$$\mu_c = \frac{\sum_{i=1}^M \sum_{j \in \Omega_c} L_{ij}}{\sum_{i=1}^M |\Omega_c| + \epsilon}, \quad (5.2)$$

where L_{ij} is the loss for pixel j in image i , calculated by a pre-trained RGB-T semantic segmentation model. ϵ is a small constant to ensure numerical stability. So, the numerator represents the sum of losses for class c across all images, while the denominator represents the sum of pixels for class c across all images.

For each semantic layout i ($i \in \{1, 2, \dots, M\}$, where M is the total number of layouts in the training set), we can calculate the score for scarcity and segmentation challenge of the semantic layout based on the above obtained P_c and μ_c . The scarcer the semantic layout and the more difficult the segmentation, the higher the score. We can obtain the score β_i for layout i by:

$$\beta_i = \frac{\sum_{c=1}^C \sum_{j \in \Omega_c} \mu_c^{ij}}{\sum_{c=1}^C |\Omega_c| + \epsilon} \frac{1}{\min_{c \in C_i} P_c}, \quad (5.3)$$

where μ_c^{ij} is the class-wise mean loss for class c for pixel j in image i . C is the total number of classes excluding the background class 0. C_i denotes the set of classes present in semantic layout i , meaning the number of classes in C_i is less than or equal to C . So, we can use this formula as a metric to measure the scarcity and segmentation challenge of the semantic layout. This metric helps determine the number of synthetic images to sample from the high-fidelity synthetic dataset. By sorting all semantic layouts in the training set from small to large using this metric, the number of samples S_i corresponding to semantic layout i can be defined as:

$$S_i = \min\left(1 + \left\lfloor \frac{(i+1)}{M} \times R_{max} \right\rfloor, R_{max}\right), \quad (5.4)$$

where $\lfloor \cdot \rfloor$ indicates rounding to the nearest integer, M is the total number of layouts in the training set, and R_{max} is the maximum sampling number for a single semantic layout in the synthetic dataset. By leveraging this mechanism, we sample more synthetic images for semantic layouts that are both uncommon in the training set and present high segmentation challenges.

5.2.3 Data Distribution of Synthetic Dataset

Based on the sampling mechanism described above, we can create different scales of sampled synthetic datasets by setting different maximum sampling numbers R_{max} in Equ. (5.4) for the original large-scale synthetic dataset. We set R_{max} to values of 1, 2, 3, 4, and 5, which resulted in 1,568, 2,744, 3,660, 4,509, and 5,333 sampled synthetic images, respectively. When $R_{max} = 1$, the pixel ratio of each class in the sampled synthetic dataset is the same as that in the real dataset. In principle, increasing R_{max} can raise the pixel ratios in the uncommon and difficult-

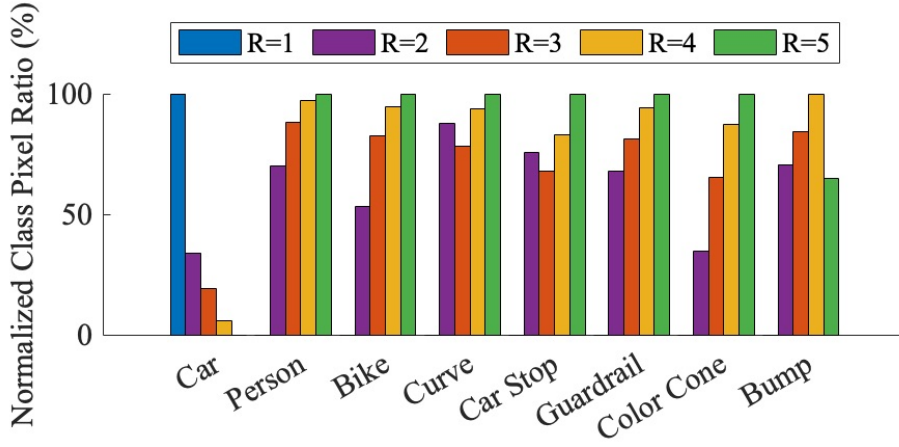


Figure 5.3: The normalized pixel ratio of each class in the synthetic dataset with different R_{max} .

to-segment classes. To more clearly illustrate the variation in pixel ratios as R_{max} increases, we normalize these ratios across different R_{max} . The normalized ratios are presented in Fig. 5.3. Take the Guardrail class as an example: prior to normalization. The pixel ratios for R_{max} values of 1, 2, 3, 4, and 5 are 1.86%, 2.22%, 2.29%, 2.36%, and 2.39%, respectively. These ratios show a slight upward trend, but the change is subtle when viewed as absolute values. However, after normalization, the pixel ratios for R_{max} values of 1, 2, 3, 4, and 5 become 0%, 67.92%, 81.13%, 94.34%, and 100%, respectively. This normalization makes the incremental changes clearer and highlights the trend. Fig. 5.3 shows that as R_{max} increases, the pixel ratio of the car class, which is the most common and easiest to segment, decreases. Meanwhile, the pixel ratios of other classes for $R_{max} > 1$ are higher than those for $R_{max} = 1$ (i.e., the pixel ratios in the real dataset).

5.3 Experimental Results and Discussions

5.3.1 Implementation Details

We use the MFNet dataset (Ha et al. 2017) as the real dataset for our experiments. The dataset consists of 2,353 pairs (including 784 flipped pairs) of RGB-T images and it includes 9 classes: Background, Car, Person, Bike, Curve, Car Stop, Guardrail, Color Cone, and Bump. We follow the same split scheme as that in (Ha et al. 2017): 1,568 pairs for training, 392 for validation, and 393 for testing. To construct a synthetic dataset, we feed 1,568 semantic layouts from the real training set into the trained layout-to-image generation model. We set the class indices of FreestyleNet (Xue et al. 2023) to 9 to adjust the number of output classes, aligning it with the MFNet dataset. Using an NVIDIA RTX 3090 graphics card, the generation speed for synthetic RGB and thermal images is 7.1 seconds per image. By setting 20 different random seeds, we obtain a large-scale synthetic dataset with 31,360 pairs of RGB-T synthetic images and corresponding semantic layouts. We set $R_{max} = 2$ in Equ. (5.4) to create the sampled synthetic training set, which includes 2,744 pairs of synthetic RGB-T images.

We train the existing RGB-T semantic segmentation model using both the sampled synthetic training set and the original real training set. Since the scale of the sampled synthetic training set is larger than that of the original real training set, and the quality of the synthetic images is lower than that of the real images, we oversample the real training set to match the scale of the sampled synthetic training set. Specifically, if two pairs of synthetic RGB-T images are sampled for a semantic layout during training, the corresponding real RGB-T images are also sampled twice to ensure same sizes of the real and synthetic training data. This

prevents the low-quality synthetic images from dominating the training process.

5.3.2 Ablation Study

To balance performance and model complexity, we use CMX (J. Zhang, Huayao Liu, et al. 2023) with a MiT-B2 backbone as the RGB-T semantic segmentation model for ablation experiments.

5.3.2.1 Ablation on Training Data

Table 5.1: The results (%) of the ablation study on the training data. $1\times$ and $2\times$ indicate using the real or synthetic training data once or twice for training, respectively. The best results are highlighted in **bold**.

Training Data		mAcc	mIoU
Real	Synthetic		
$1\times$		67.3	58.2
	$1\times$	65.6	52.8
$1\times$	$1\times$	72.2	59.9
$2\times$		68.5	57.5

To examine the effects of various training sets on the RGB-T semantic segmentation model, we first conduct an ablation study presented in Tab. 5.1. The first row of Tab. 5.1 shows the segmentation results obtained by training with 1,568 pairs of real RGB-T images from the original MFNet dataset. The second row shows the results from training with 1,568 RGB-T images generated from 1,568 semantic layouts. Comparing the results, we observe that the mIoU of training with synthetic data is 5.4% lower than that of training with real data. This indicates that although the generated synthetic RGB-T images are visually similar to real images, there remains a quality gap between them and real images. The

third row of Tab. 5.1 shows the segmentation results obtained by training with 1,568 pairs of real RGB-T images and the corresponding 1,568 pairs of synthetic RGB-T images. Training with a combination of real and synthetic images results in a higher mIoU compared to training with only real images. This indicates that despite synthetic images not being as high-quality as real images, joint training enhances the diversity of the training set. This allows the model to learn from a wider variety of samples, thereby improving its performance. To demonstrate that the mIoU improvement in the third row is not solely due to increased amount of training data, we conduct the fourth experiment using twice the original real training set, ensuring the same amount of training data as in the third experiment. The mIoU from the fourth experiment is decreased by 2.4% compared to the third and is even lower than the mIoU obtained with just the original real training set. This indicates that merely increasing the training data without enhancing its diversity does not effectively improve the segmentation performance.

5.3.2.2 Ablation on Sampling Mechanism

Table 5.2: The results (%) of the ablation study on the maximum sampling number R_{max} . The best results are highlighted in **bold**.

R_{max}	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		
1	93.0	88.3	83.0	73.5	74.2	65.0	66.1	51.2	41.1	31.4	36.1	8.2	56.8	52.9	69.2	57.7	68.8	58.5
2	92.9	88.3	84.2	74.9	73.4	65.2	64.6	51.3	43.5	36.3	63.6	14.2	57.9	52.8	74.1	62.1	72.6	60.4
3	93.0	88.4	83.3	74.0	72.7	64.0	65.8	51.7	44.7	36.1	68.2	14.3	56.4	52.6	68.3	57.4	72.4	59.7
4	93.7	88.9	82.7	73.5	73.8	65.0	62.8	51.6	50.6	38.6	53.6	12.3	54.2	51.7	69.3	54.6	71.1	59.4
5	92.8	88.0	83.7	74.4	74.0	64.9	63.4	51.0	46.3	34.5	59.1	12.3	55.3	50.8	70.7	56.6	71.6	59.0

We set 20 different random seeds and generate a synthetic dataset 20 times the scale of the original dataset. Our goal is to enhance the segmentation performance of uncommon and challenging classes through the sampling mechanism, thereby improving overall performance. We design an ablation experiment, shown

in Tab. 5.2, to analyze the impact of the maximum number of samples R_{max} on the results. $R_{max} = 1, 2, 3, 4, 5$ means that for each semantic layout in the original training set, a maximum of 1, 2, 3, 4, or 5 pairs of synthetic RGB-T images are sampled, respectively. The higher the R_{max} , the larger the sampled synthetic training set. Tab. 5.2 indicates that the overall segmentation performance of the model improves when $R_{max} > 1$ compared to $R_{max} = 1$. The accuracy (Acc) and IoU for the Car Stop and Guardrail classes, which have the worst segmentation results, are significantly improved when $R_{max} > 1$. The model achieves the highest mAcc and mIoU when $R_{max} = 2$, but segmentation performance gradually deteriorates as R_{max} increases further. This indicates that while training with both synthetic and real images can enhance segmentation performance, an excessive number of synthetic images can lead to low quality issues and overfitting, thus degrading model performance. Specifically, although synthetic data increases the diversity of the training set, the quality of synthetic images is generally inferior to that of real images. As the number of synthetic data increases, the model may learn more noises and inaccurate features.

Table 5.3: The results (%) of the ablation study on the sampling mechanism. P_c is the pixel ratio of each class. μ_c is the mean loss of each class. A $\checkmark$ under P_c or μ_c indicates whether P_c or μ_c is included in the sampling mechanism for synthetic data. 'OverSamp' indicates whether the real training set is oversampled to ensure the training size of synthetic and real data are the same. The best results are highlighted in **bold**.

Sampling			Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		mAcc	mIoU
P_c	μ_c	OverSamp	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
$\checkmark$		$\checkmark$	93.1	88.3	83.1	73.9	73.3	64.8	64.8	50.5	48.2	37.5	54.9	14.0	54.7	49.9	71.5	55.3	71.4	59.2
	$\checkmark$	$\checkmark$	92.7	87.9	82.6	73.7	71.2	63.3	62.9	51.0	44.0	36.4	65.1	16.0	56.4	51.7	64.9	57.1	71.0	59.5
$\checkmark$	$\checkmark$		92.3	87.9	83.1	74.0	74.8	64.7	67.0	51.1	49.4	34.0	53.4	12.0	55.9	51.9	70.9	59.2	71.8	59.2
$\checkmark$	$\checkmark$	$\checkmark$	92.9	88.3	84.2	74.9	73.4	65.2	64.6	51.3	43.5	36.3	63.6	14.2	57.9	52.8	74.1	62.1	72.6	60.4

As shown in Tab. 5.3, we conduct the experiments to analyze the effect of different sampling mechanisms for synthetic and real data on model performance. We set $R_{max} = 2$ and sampled 2,744 synthetic RGB-T images for joint training. In Tab. 5.3, the first row shows the segmentation results when the sampling mecha-

nism for synthetic data considers only the pixel ratio of each class. The second row shows the results when it considers only the segmentation difficulty of each class. Both rows oversample the real training set, resulting in 2,744 pairs of real RGB-T images used for joint training, to ensure the training size of synthetic and real data are the same. The results in the first two rows are better than those obtained by training only with real datasets (first row of Tab. 5.1). This demonstrates that using either the pixel ratio or segmentation difficulty for synthetic data improves the ability of the model. However, their segmentation performance is inferior to the fourth row, which considers both pixel ratio and segmentation difficulty in the synthetic data sampling mechanism. This proves that considering both factors can simultaneously maximize the segmentation ability of the model.

Compared to the fourth row, which achieves the best segmentation performance, the third row does not oversample the real data. It retains 1,568 pairs of real RGB-T images for joint training and, as a result, obtains worse mIoU, particularly for the Car Stop and Guardrail classes. These classes have scarce samples and are more difficult to segment. Sample scarcity means fewer training samples for the generative model, and harder segmentation indicates greater difficulty in extracting semantic features. Consequently, such classes yield poorer quality synthetic data from the generative model. When the scale of synthetic data outweighs that of real data in joint training, these classes further suppress segmentation accuracy compared to other classes. The Car Stop and Guardrail classes unexpectedly achieved the highest Acc while obtaining the lowest IoU. This suggests that false positives dominated their segmentation results, incorrectly judging negative cases as positive. This further validates the effect of lower-quality synthetic data on the segmentation results. The above results demonstrate the effectiveness of

oversampling real data to match the scale of synthetic data.

5.3.3 Comparative Study

Based on the ablation experiments described above, we construct the synthetic dataset used in the comparative experiments by setting $R_{max} = 2$ and sampling from the generated large-scale synthetic dataset. We also oversample the real dataset to match the size of the synthetic dataset.

5.3.3.1 The Quantitative Results

Table 5.4: The per-class results (%) of models using different method to alleviate class imbalance. The best results are highlighted in **bold**.

Methods	Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
CMX(MiT-B2) (J. Zhang, Huayao Liu, et al. 2023)	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
Focal Loss (Lin et al. 2017)	93.9	88.1	84.8	73.6	74.5	63.9	68.8	49.1	42.6	31.7	51.6	9.4	55.6	51.0	70.6	59.5	71.3	58.3
Resampling	93.4	87.7	84.8	74.1	73.5	64.1	63.1	47.7	45.0	27.0	66.1	11.0	58.3	52.4	68.3	54.7	72.4	57.4
SyntheticSeg (Ours)	92.9	88.3	84.2	74.9	73.4	65.2	64.6	51.3	43.5	36.3	63.6	14.2	57.9	52.8	74.1	62.1	72.6	60.4

In Tab. 5.4, we compare our SyntheticSeg with focal loss (Lin et al. 2017) and resampling. All the three methods use CMX (J. Zhang, Huayao Liu, et al. 2023) with MiT-B2 as the baseline. Focal loss is implemented with $\gamma = 2$, emphasizing hard-to-classify examples. The resampling method oversamples rare samples during training to emphasize classes with fewer instances. While focal loss achieves a higher IoU than the baseline for Curve, Car Stop Guardrail and Bump, it performs poorly on other classes. Resampling only outperforms the baseline in Curve and Guardrail, resulting in a lower mIoU than the baseline. Our method, however, not only significantly improves segmentation accuracy for objects with few samples and high segmentation difficulty but also performs well across other classes.

Table 5.5: The per-class results (%) of models using different training sets. We used RTFNet (Sun, Zuo, and M. Liu 2019), CMX (J. Zhang, Huayao Liu, et al. 2023), and CRM (Shin et al. 2024) with different backbones as the RGB-T segmentation models. Δ represents the mIoU improvement when training with both synthetic and real data compared to using only real data. The best results of each backbone are highlighted in **bold**.

Model	Backbone	Training Data		IoU								mIoU	Δ
		Real	Synthetic	Car	Person	Bike	Curve	Car Stop	Guardrail	Color Cone	Bump		
RTFNet (Sun, Zuo, and M. Liu 2019)	ResNet-50	✓		86.3	67.8	58.2	43.7	24.3	3.6	26.0	57.2	51.7	↑ 4.4
		✓	✓	86.5	69.9	60.8	49.5	27.8	8.7	46.0	58.0	56.1	
	ResNet-152	✓		87.4	70.3	62.7	45.3	29.8	0.0	29.1	55.7	53.2	↑ 3.7
		✓	✓	88.6	71.1	63.3	50.2	30.5	8.0	50.2	51.7	56.9	
CMX (J. Zhang, Huayao Liu, et al. 2023)	MiT-B2	✓		89.4	74.8	64.7	47.3	30.1	8.1	52.4	59.4	58.2	↑ 2.2
		✓	✓	88.3	74.9	65.2	51.3	36.3	14.2	52.8	62.1	60.4	
	MiT-B4	✓		90.1	75.2	64.5	50.2	35.3	8.5	54.2	60.6	59.7	↑ 1.2
		✓	✓	89.4	74.9	65.0	54.6	38.7	13.6	53.1	60.8	60.9	
	Swin-T	✓		90.0	73.1	63.7	47.9	40.7	9.9	54.4	54.2	59.1	↑ 0.8
		✓	✓	90.3	75.0	64.1	49.4	48.6	11.1	50.6	52.1	59.9	
CRM (Shin et al. 2024)	Swin-S	✓		90.6	75.5	67.2	48.3	43.4	11.8	56.8	59.3	61.2	↑ 0.8
		✓	✓	90.7	75.5	65.2	52.0	50.4	15.4	54.8	55.2	62.0	
	Swin-B	✓		90.0	75.1	67.0	45.2	49.7	18.4	54.2	54.4	61.4	↑ 0.7
		✓	✓	91.2	74.4	65.2	50.9	46.4	15.8	54.2	62.0	62.1	

In Tab. 5.5, we select three RGB-T semantic segmentation models and compare them by training only with real data versus training with both synthetic data from our proposed method and real data. We train each model with multiple backbones to analyze the impact of different training data on various architectures and backbone models. The first row for each model and backbone uses only the real training set, while the second row uses both the real and synthetic training sets. After incorporating synthetic training data, the segmentation ability of RTFNet has been significantly improved. Specifically, the IoU of the Bump class in RTFNet with ResNet-152 is the only one that decreases, while other segmentation results show improvement. For the three models with different backbones, the mIoU has consistently been improved after training with both synthetic and real data. More importantly, these models (except for CRM with Swin-B) notably enhanced the segmentation IoU of the Car Stop and Guardrail classes, which have few samples and are challenging to segment. This aligns with the intended outcomes of our proposed sampling mechanism. The results show that our method consistently improves the overall segmentation performance.

5.3.3.2 The Qualitative Demonstrations

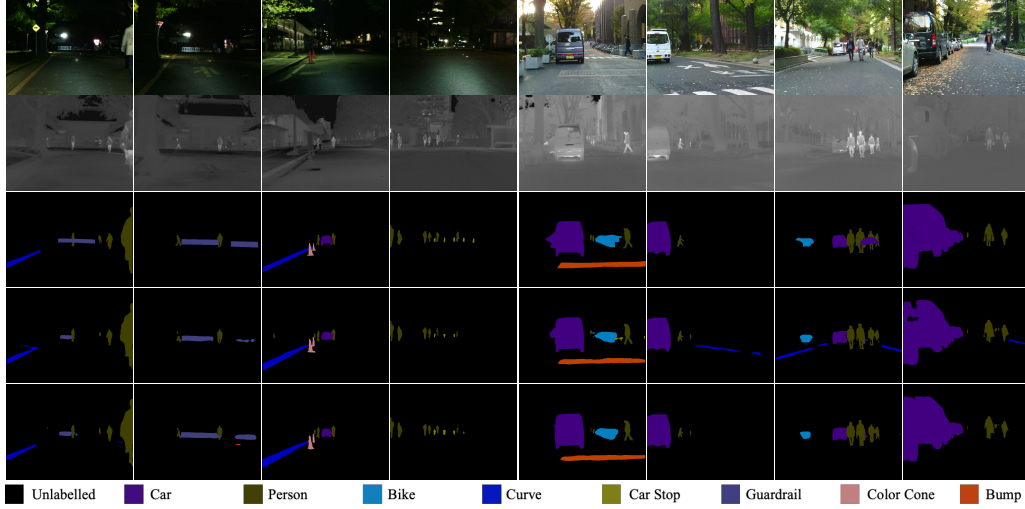


Figure 5.4: Qualitative comparison for semantic segmentation of RGB-T images on MFNet (Ha et al. 2017) dataset. The rows from top to bottom are RGB images, thermal images, ground truth of semantic layouts, results obtained by training CMX (J. Zhang, Huayao Liu, et al. 2023) with MiT-B2 using real data, and results obtained by training CMX (J. Zhang, Huayao Liu, et al. 2023) with MiT-B2 using real and sampled synthetic data. The first four columns and the last four columns are the samples of nighttime and daytime, respectively.

To better demonstrate the effectiveness of our SyntheticSeg, we select 4 samples taken at night and 4 samples taken during the day, which are displayed in Fig. 5.4. After training with synthetic and real data, the ability to segment uncommon and difficult-to-segment classes of the existing models has significantly been improved. For example, the models trained with synthetic and real data in the second cases can better segment the Guardrail class, which has the lowest pixel ratio. This result aligns with the conclusions from the quantitative experiments above. We also find that training with synthetic and real data improved the resistance to false detection of the models. For example, CMX trained with only real data in the sixth, seventh, and eighth cases detect the background as the

Curve class. This issue is significantly reduced after adding synthetic data to the training process. This improvement is due to the increased number and diversity of training samples from the synthetic dataset.

5.3.3.3 Suboptimal Case Analysis

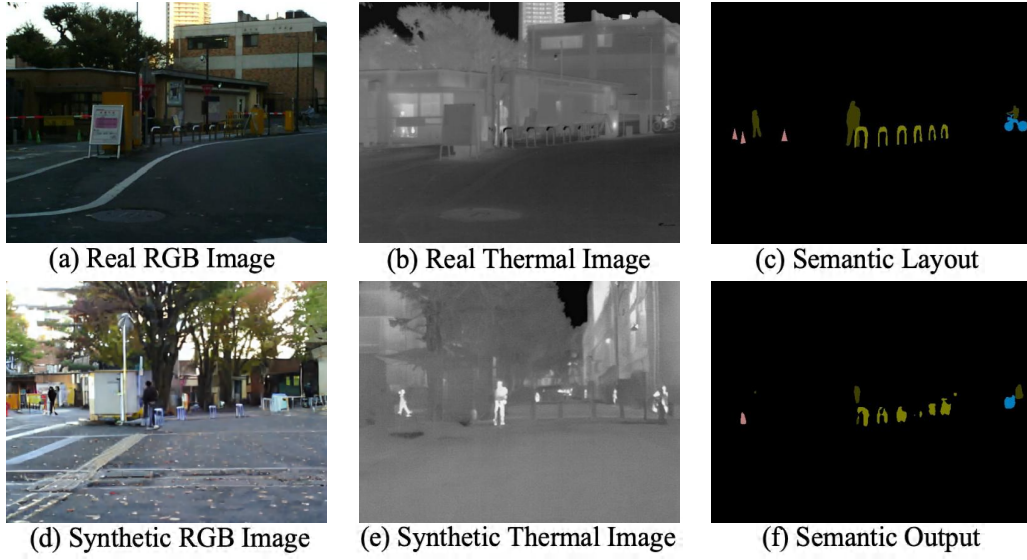


Figure 5.5: Visualization of the segmentation suboptimal case. (a) and (b) are real RGB-T images from MFNet dataset; (c) is the corresponding semantic layout; (d) and (e) are synthetic RGB-T images generated from (c); (f) is the prediction of segmentation model.

Although the above quantitative and qualitative analyses have proved the effectiveness of our method, there are still cases where the segmentation performance is not satisfactory, as shown in Fig. 5.5. Specifically, Fig. 5.5(f) illustrates that the prediction results are not accurate for both the Person class with a large sample size and the Car Stop class with a small sample size. The first reason is that the image quality of distant objects in real RGB images is poor. As seen in Fig. 5.5(a), the model struggles to extract rich semantic information and accurate contours of distant persons in dark environments. Besides, for classes with a

small sample size, the quality of synthetic images generated by the model is sub-optimal. The generated Car Stop and Color Cone classes in Fig. 5.5(d) and (e) and their corresponding semantic layouts in Fig. 5.5(e) lack spatial consistency, which means that if the synthetic images dominate the training, the segmentation accuracy will be suppressed. In addition, since the synthetic RGB and thermal images are obtained separately through the generative model, their background parts cannot guarantee spatial consistency. Despite our method has been proven to effectively improve the segmentation performance, this spatial inconsistency may limit the further improvement.

5.4 Conclusions and Future Work

We proposed here a data-driven method, SyntheticSeg, to enhance RGB-T segmentation by using synthetic data augmentation. We created high-quality synthetic RGB-T images and built a large-scale dataset to diversify training samples. The new metric we introduced effectively guides sampling from synthetic datasets by considering the scarcity of semantic layouts and the difficulty of segmentation. This method not only alleviates the class imbalance problem but also improves the overall segmentation accuracy. The experimental results show that our method achieves state-of-the-art performance on the MFNet dataset.

However, our method has some limitations. First, the quality of synthetic images is still inferior to real images, particularly for scarce classes, leading to suppressed segmentation performance when synthetic images dominate the training. Second, the synthetic RGB and thermal images cannot be accurately aligned, especially in the background, and their physical consistency is uncertain. Our fu-

ture work would focus on developing a generative model that generates synthetic RGB-T image pairs with improved physical and spatial consistency, potentially enhancing the segmentation performance.

Chapter 6

Spatial Balancing

6.1 Motivation

In previous chapters, we explored three major challenges in RGB-T semantic segmentation: improving interpretability through illumination-guided fusion (Chapter 3), enhancing temporal consistency through virtual view generation (Chapter 4), and alleviating data scarcity through synthetic data augmentation (Chapter 5). While these methods have progressively improved overall segmentation accuracy, they all neglect the spatial distribution of segmentation performance across different regions. As discussed in Section 1.2, we observed differences in segmentation performance across different spatial regions. Specifically, in autonomous driving scenarios, performance in the central region consistently lags behind that in the edge regions; we term this phenomenon "spatial bias."

Spatial bias is crucial for autonomous driving applications because objects crucial to driving safety, such as vehicles, pedestrians, and bicycles, are often located in the central region of an image. If the segmentation model performs poorly in

critical areas, it can lead to a decrease in the reliability of the entire perception system. As shown in Figure 1.3, this spatial bias arises contrary to conventional wisdom: the central region has more training samples, yet its segmentation performance is worse.

The main objective of this chapter is to systematically investigate the causes of spatial bias and develop methods to mitigate the problem. The study found that spatial bias is closely related to object complexity and image quality. In the central region, objects are typically smaller, more densely distributed, and more easily occluded, thus increasing visual complexity. Simultaneously, spatial differences exist in RGB image quality, with the signal-to-noise ratio (SNR) typically lower in the central region. However, thermal imaging images exhibit the opposite characteristics, maintaining a higher SNR in the central region.

Our core idea is that spatial bias can be mitigated by adaptively adjusting the fusion strategy based on spatial location. Specifically, in central regions where RGB image is noisy but thermal image is more reliable, we increase the reliance on thermal features; in edge regions we utilize the two modalities more evenly. This spatially aware fusion strategy is similar to the illumination-guided fusion strategy in Chapter 3, but it uses spatial dimensions rather than illumination conditions as the guiding principle. To realize this idea, we introduce Gaussian-guided Regional Balancing Masking (GRBM), which generates complementary binary masks for RGB and thermal images based on a Gaussian distribution centered at the image center. These masks mask the RGB features of image center regions with a higher probability, forcing the network to rely more on thermal features in those regions. Besides, we propose a spatial-weighted loss function that assigns higher weights to edge regions during training. This encourages the model to learn more from

clearer and simpler edge regions.

Chapter Organization. The remainder of this chapter is organized as follows. Section 6.2 provides a detailed analysis of the causes of spatial bias, as well as a theoretical interpretation using information theory and feature space analysis. Section 6.3 provides an overview of our proposed framework and detailed introduces the GRBM method. Besides, it describes the spatial-weighted loss function and presents the region-based evaluation metrics to measure spatial bias. Section 6.4 presents comprehensive experimental results on two datasets, as well as extensive ablation studies. Finally, Section 6.5 summarizes the main contributions and discusses their significance for future research.

6.2 Spatial Bias in Segmentation

6.2.1 Differences from Class Imbalance

Class imbalance caused by long-tail distribution may lead to inferior segmentation performance for tail classes. Let $D = \{(x_i, y_i) | i = 1, 2, \dots, N_s\}$ 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 $c \in \{1, 2, \dots, C\}$, the set D_c is defined as $D_c = \{(x_i, y_i) | y_i = c, i = 1, 2, \dots, N_c\}$, 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 from (H. Li, Chu, and Sun 2024).

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 $S = \{(x_i, y_i) | i = 1, 2, \dots, M_s\}$, 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 $r \in \{1, 2, \dots, K\}$, the set S_r is defined as $S_r = \{(x_i, y_i) | \text{region}(x_i) = r, i = 1, 2, \dots, M_r\}$, 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.

6.2.2 Reasons Causing Spatial Bias

To thoroughly investigate the reason for spatial bias, we first partition an image into distinct regions. There are several candidate strategies for partitioning. For example, vertical and horizontal partitioning could reveal directional biases, but ring-shaped partitioning is suitable for analyzing central-versus-edge performance differences particularly in autonomous driving domains, since the perspective geometry of forward-facing cameras causes distant objects on roads to be projected around image central regions, while nearby objects are more likely to appear around edge regions. In this section, we evaluate the three strategies to determine which could best reveals the spatial bias phenomenon.

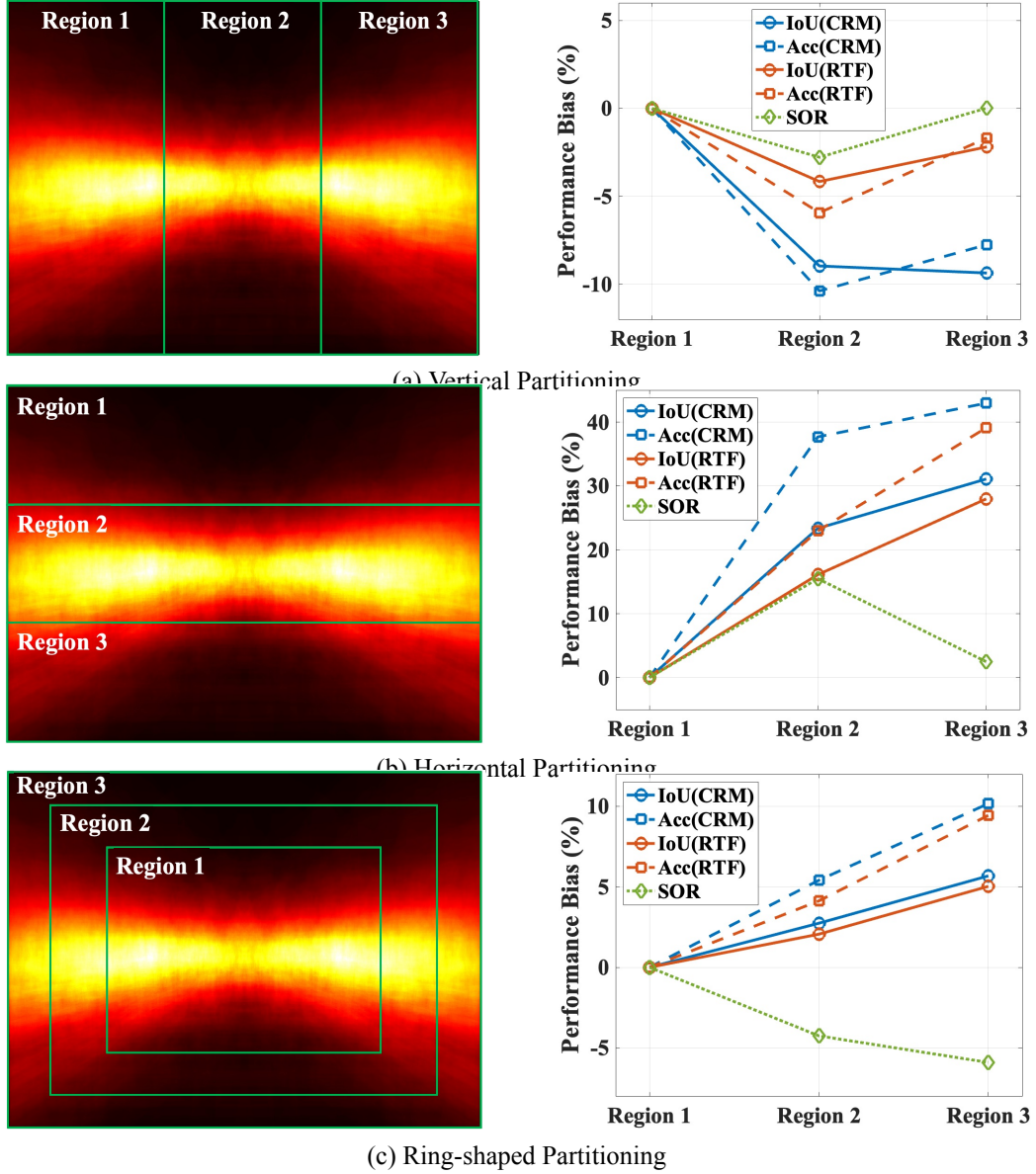


Figure 6.1: Visualization of three partitioning strategies and their corresponding performance variations in different regions. CRM-B (Shin et al. 2024) and RTFNet-152 (Sun, Zuo, and M. Liu 2019) are used as the segmentation networks. The segmentation performance is represented by the average IoU and 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.

6.2.2.1 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 (Ha et al. 2017) using the aforementioned three partitioning strategies. As illustrated in Fig. 6.1, 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 $\{L_1, L_2, \dots, L_{N_{mask}}\}$. 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:

$$P(x, y) = \frac{1}{N_{mask}} \sum_{i=1}^{N_{mask}} \mathbb{I}\{L_i(x, y) \neq 0\}, \quad (6.1)$$

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 $\{R_1, R_2, \dots, R_K\}$, 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:

$$SOR_k = \frac{1}{|R_k|} \sum_{(x,y) \in R_k} P(x,y), \quad (6.2)$$

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. 6.1, 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 suitable 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.

6.2.2.2 Class Distribution

It is known that class imbalance (H. Li, Chu, and Sun 2024) 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.

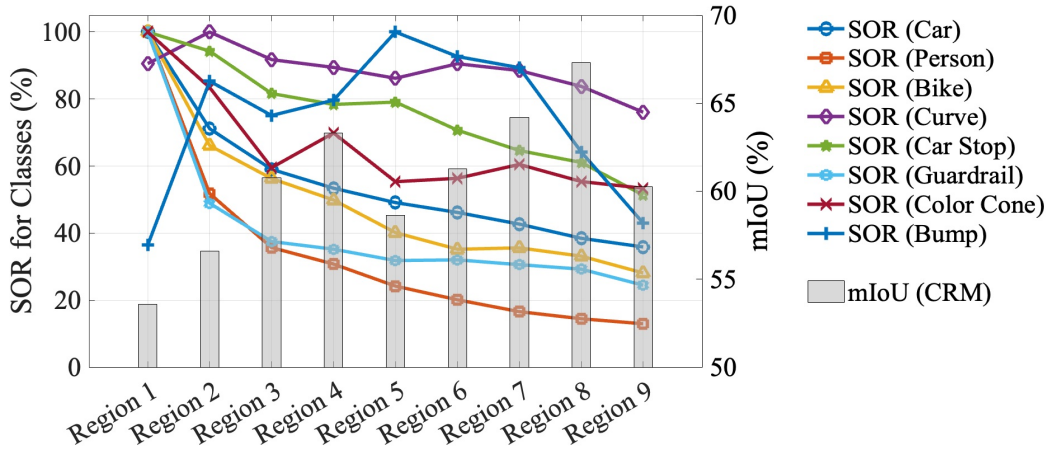


Figure 6.2: The SOR (%) distribution for different classes across various regions and the mIoU of all classes for each region. Region 1 represents the innermost region, while region 9 represents the outermost region. We normalize the SOR for each class in different regions to show the sample proportion 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.

To verify this hypothesis, we display the SOR of different classes in each region and the IoU of each region (see Fig. 6.2). 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. 6.3. 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.

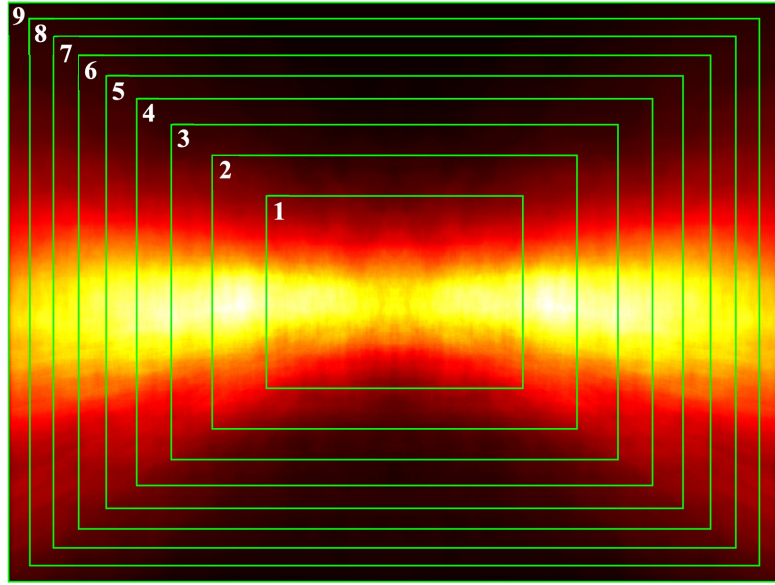


Figure 6.3: Visualization of nine ring-shaped regions with equal areas. Regions are numbered from 1 to 9. Region 1 is the innermost region and region 9 to the outermost region.

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 can not 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. So, we further illustrate the IoU of different classes across various regions in Tab. 6.1. It is evident that classes with a more evenly distributed sample size (e.g., Curve) have signifi-

cantly 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.

Table 6.1: The IoU (%) of different classes across various regions. Region 1 represents the innermost region, while region 9 represents the outermost region. CRM-B (Shin et al. 2024) is used as the segmentation network. For each class, the two lowest IoU values in different regions are denoted in **bold** and underline, respectively.

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

6.2.2.3 Target Object Complexity

From previous analysis, we have already found that the image central regions contains more objects but exhibits 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 more number of training sam-

ples in the central regions.

To quantitatively measure visual complexity, we adopt Laplacian Variance (LV) as a metric. The rationale for this choice is well founded in image processing literature (Pech-Pacheco et al. 2000; Pertuz, Puig, and Garcia 2013). 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 visual complexity in a region. This principle is widely used in applications like auto-focusing to quantify image sharpness (Pech-Pacheco et al. 2000). More importantly, the effectiveness of this type of operator has been well studied. In a detailed review, Pertuz et al. (Pertuz, Puig, and Garcia 2013) 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 visual complexity within a region. It is calculated as follows:

$$\sigma_{\text{Laplacian}}^2 = \frac{1}{W_{img}H_{img}} \sum_{i=1}^{W_{img}} \sum_{j=1}^{H_{img}} \left[\mathcal{L}(i, j) - \mu_{\text{Laplacian}} \right]^2, \quad (6.3)$$

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. 6.4 displays the LV values of RGB and thermal images during daytime and nighttime from the MFNet dataset. The conclusions are listed

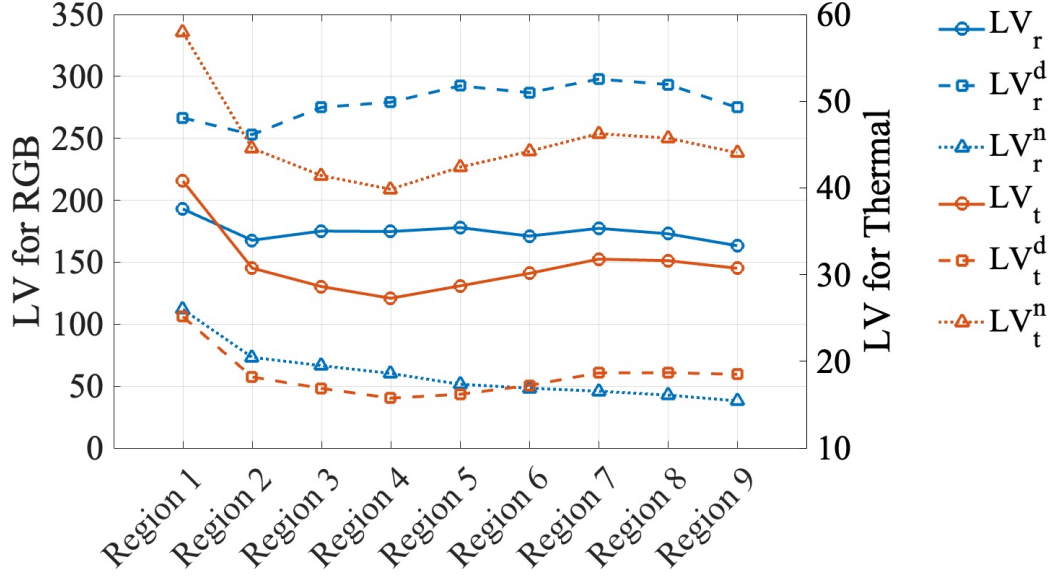


Figure 6.4: Laplacian Variance (LV) for each region. Region 1 represents the innermost region, while region 9 represents the outermost region. LV_r , LV_r^d , and LV_r^n represent the LV of all RGB images, daytime RGB images, and nighttime RGB images in the dataset, respectively. LV_t , LV_t^d , and LV_t^n represent the LV of all thermal images, daytime thermal images, and nighttime thermal images in the dataset, respectively.

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.
- The LV values of daytime RGB images are higher than those of nighttime RGB images, indicating that daytime RGB images have more visual tex-

tures.

- 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.

6.2.2.4 Image Quality

Although in Fig. 6.4 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:

$$\text{SNR} = \frac{\mu_{\text{signal}}}{\sigma_{\text{noise}} + \epsilon}, \quad (6.4)$$

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 noises, reflecting the intensity of noises in an image. ϵ is a very small number to prevent division by zero. Higher SNR indicates better image quality and less noises in that region, and vice versa.

Fig. 6.5 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:

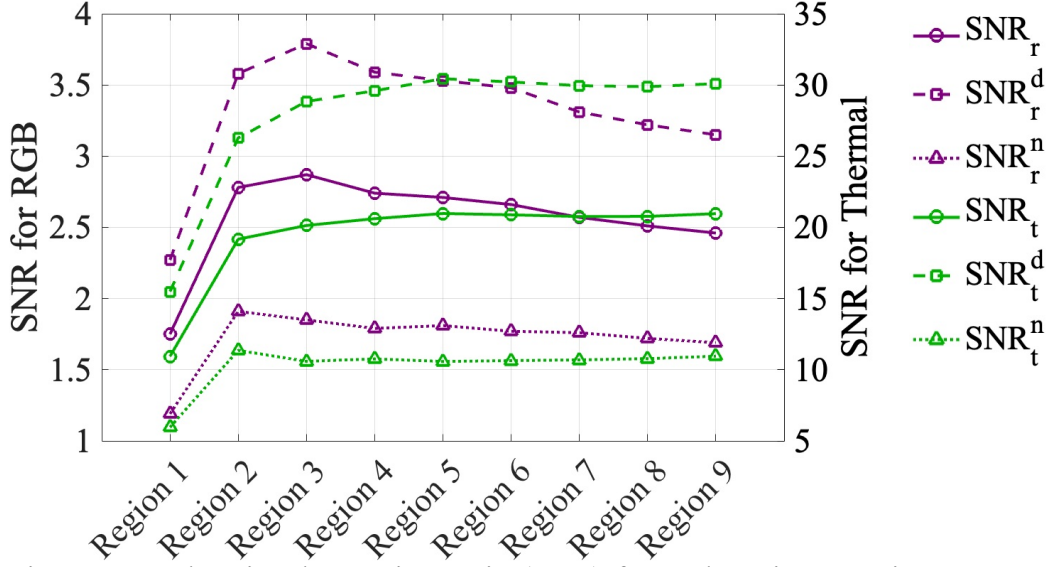


Figure 6.5: The Signal-to-Noise Ratio (SNR) for each region. Region 1 represents the innermost region, while region 9 represents the outermost region. SNR_r , SNR_r^d , and SNR_r^n represent the SNR of all RGB images, daytime RGB images, and nighttime RGB images in the dataset, respectively. SNR_t , 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.

- Both RGB and thermal images show more noises 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 noises. 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.

6.2.3 Theoretical Interpretation

Our analysis in Sec. 6.2.2 reveals a close correlation between spatial segmentation performance degradation and two key factors: high object complexity and inferior image quality. Here, we provide a rigorous theoretical interpretation for the correlation using information theory and feature space analysis.

6.2.3.1 Spatial Bias as a Problem of High Local Entropy

Information entropy measures the uncertainty within a system. In machine learning, higher entropy in input data usually corresponds to increased difficulty for learning. We adopt Shannon Entropy to quantify this uncertainty. For a discrete feature variable X with probability mass function $p(x)$, the entropy $\mathcal{H}(X)$ is defined as:

$$\mathcal{H}(X) = - \sum_{x \in X} p(x) \log p(x). \quad (6.5)$$

A higher value of $\mathcal{H}(X)$ indicates higher uncertainty, implying that the learning task is more complex. We argue that the lower segmentation performance observed in central regions is due to it being a high-entropy area, which results from two contributing factors:

- **Intrinsic Entropy from Complexity:** High target complexity (high LV) implies a larger variety of classes in a local neighborhood. This leads to a flatter feature distribution $p(x)$, where many classes have non-negligible probabilities rather than a few dominant ones. According to Eq. 6.5, a more uniform distribution maximizes entropy, making the feature space more ambiguous and the class boundaries harder to separate.
- **Noise-Induced Entropy from Inferior Quality:** Inferior image quality (low SNR) introduces random noise, which can be modeled as corrupting true feature signals. This convolution of the signal distribution with a noise distribution "smears out" $p(x)$, further increasing its uniformity and thus its entropy. More formally, noise increases the conditional entropy $\mathcal{H}(X|Y)$, where Y is the true class label, meaning that even when the true class is known, the observed feature x is still highly uncertain. The network's challenge is to minimize the posterior uncertainty $\mathcal{H}(Y|X)$, which is fundamentally harder when the input itself is noisy.

6.2.3.2 Modality Fusion and Mutual Information

Effective RGB-T fusion should leverage the complementary information between the two modalities. This can be conceptualized using Mutual Information (MI), which quantifies the information that one random variable contains about another. For the feature spaces of RGB and thermal modalities, denoted by $\mathbf{R}$ and $\mathbf{T}$ respectively, the mutual information $I(\mathbf{R}; \mathbf{T})$ is formally defined as the Kullback-Leibler divergence between the joint distribution $P(r, t)$ and the product of the marginal distributions $P(r)P(t)$:

$$I(\mathbf{R}; \mathbf{T}) = \sum_{r \in \mathbf{R}} \sum_{t \in \mathbf{T}} P(r, t) \log \left(\frac{P(r, t)}{P(r)P(t)} \right). \quad (6.6)$$

This definition can be equivalently expressed in terms of Shannon Entropy, which provides a more intuitive interpretation for our analysis:

$$I(\mathbf{R}; \mathbf{T}) = \mathcal{H}(\mathbf{R}) + \mathcal{H}(\mathbf{T}) - \mathcal{H}(\mathbf{R}, \mathbf{T}), \quad (6.7)$$

where $\mathcal{H}(\mathbf{R})$ and $\mathcal{H}(\mathbf{T})$ represent the total uncertainty or information content within the RGB feature space and thermal feature space, respectively. $\mathcal{H}(\mathbf{R}, \mathbf{T})$ represents the joint entropy, measuring the total uncertainty of the combined RGB-T system. Because the two modalities capture the same underlying physical scene, they share redundant information. This redundancy makes their joint entropy less than the summation of their individual entropies. Therefore, $I(\mathbf{R}; \mathbf{T})$ represents this redundancy, the amount of information that is shared between the RGB and Thermal modalities. It quantifies the reduction in uncertainty about modality $\mathbf{R}$ when modality $\mathbf{T}$ is known (and vice versa).

However, the goal is not merely to fuse all information, but to fuse useful information while rejecting noises. We model the observed features as $\mathbf{R} = S_{\mathbf{R}} + N_{\mathbf{R}}$ and $\mathbf{T} = S_{\mathbf{T}} + N_{\mathbf{T}}$, where $S_{\mathbf{R}}$, $N_{\mathbf{R}}$, $S_{\mathbf{T}}$ and $N_{\mathbf{T}}$ represent the signal of RGB image, the noise of RGB image, the signal of thermal image, and the noise of thermal image, respectively. The true objective is to maximize the information about the ground-truth labels Y , that is, maximize $I(f(\mathbf{R}, \mathbf{T}); Y)$, where f is the fusion function.

Our analysis shows a regional disparity in signal quality: in central regions, $N_{\mathbf{R}}$ is large while $N_{\mathbf{T}}$ is small. A naive fusion function f might allow high noises from the RGB channel to corrupt the relatively clean signal from the thermal channel. Our proposed Gaussian-guided Regional Balancing Masking (GRBM) method acts as a spatially-aware information gate. By probabilistically masking the noisy RGB features in the center, it forces the fusion function f to rely more heavily on the thermal stream. This can be interpreted as a practical heuristic for optimizing the information flow, aiming to maximize the influence of the modality with the highest signal-to-noise ratio in any given region.

6.3 The Proposed Method

6.3.1 Method Overview

We use our findings to improve RGB-T semantic segmentation. The overall framework of our method is illustrated in Fig. 6.6. 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

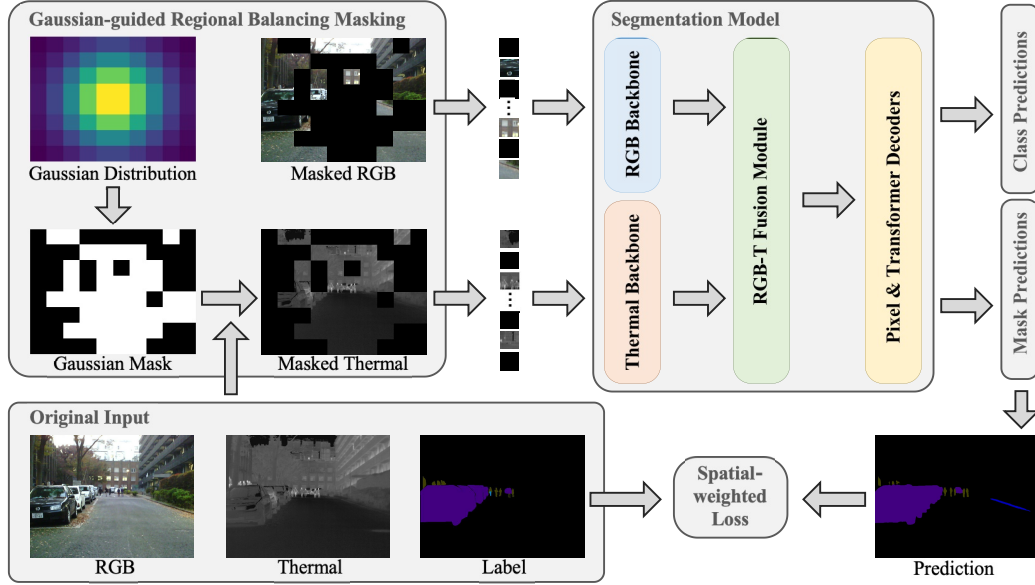


Figure 6.6: The overall framework of our method to improve the RGB-T semantic segmentation. Complementary masks for RGB and thermal images are generated through the Gaussian-guided Regional Balancing Masking (GRBM) method. We use CRM (Shin et al. 2024) as the segmentation model. The spatial-weighted loss is calculated based on the model’s predictions and the ground truth labels.

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 (Shin et al. 2024). Finally, we calculate our designed spatial-weighted loss between model predictions and the corresponding ground truth to improve the segmentation performance.

6.3.2 Gaussian-guided Regional Balancing Masking (GRBM)

From our previous analysis, we have showed that segmentation performance in central regions is relatively inferior. Our theoretical interpretation in Sec. 6.2.3 further posits that this is due to central regions being a high-entropy space, where the

RGB modality is corrupted by noises 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:

$$M_{rgb} + M_{th} = \mathbf{1}, \quad (6.8)$$

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

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

$$N_x = \lfloor \frac{W_{img}}{PS} \rfloor, \quad N_y = \lfloor \frac{H_{img}}{PS} \rfloor, \quad (6.9)$$

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:

$$\mathcal{W}(i, j) = \exp \left(-\frac{(i - \frac{N_x}{2})^2 + (j - \frac{N_y}{2})^2}{2\sigma_p^2} \right), \quad (6.10)$$

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. 6.6, 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:

$$M_{rgb} = \text{Bernoulli}(\mathcal{W}(x, y)), \quad (6.11)$$

$$M_{th} = \mathbf{1} - M_{rgb}, \quad (6.12)$$

where $\text{Bernoulli}(\cdot)$ 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:

$$\langle M_{rgb}, M_{th} \rangle_i = \begin{cases} \langle \mathbf{0}, M_{th} \rangle & \text{if } i = 0 \\ \langle M_{rgb}, \mathbf{0} \rangle & \text{if } i = 1, \\ \langle M_{rgb}, M_{th} \rangle & \text{if } i = 2 \end{cases} \quad (6.13)$$

where $i \in 0, 1, 2$ is a random number. $\mathbf{0}$ represents an all-zero matrix, indicating that no mask is used. This stochastic selection of masking patterns is a critical component for regularization (Shin et al. 2024). 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:

$$I'_{rgb} = (\mathbf{1} - M_{rgb}) \odot I_{rgb} + M_{rgb} \odot \tau, \quad (6.14)$$

$$I'_{th} = (\mathbf{1} - M_{th}) \odot I_{th} + M_{th} \odot \tau, \quad (6.15)$$

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. Then, 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, par-

ticularly those employing cross-modal attention, 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 from 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 noises.

6.3.3 Spatial-Weighted Loss

The central regions of RGB and thermal images contain more noises 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 noises, leading to degradation in the overall segmentation performance (details of this conclusion is 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:

$$d_{ij} = \sqrt{(y_i - y_c)^2 + (x_i - x_c)^2}, \quad (6.16)$$

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 $[1, 1 + \alpha]$:

$$w_{ij} = 1 + \frac{\alpha d_{ij}}{d_{max}}, \quad (6.17)$$

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:

$$\begin{aligned} \ell_{wce} = \frac{1}{W_{img} \times H_{img}} \sum_{i=1}^{W_{img}} \sum_{j=1}^{H_{img}} w_{i,j} & \left[t_{i,j} \log(\text{Sigmoid}(s_{i,j})) \right. \\ & \left. + (1 - t_{i,j}) \log(1 - \text{Sigmoid}(s_{i,j})) \right], \end{aligned} \quad (6.18)$$

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

$$\begin{aligned} \ell_{wdice} = 1 - & \\ & \frac{2 \sum_{i=1}^{W_{img}} \sum_{j=1}^{H_{img}} w_{i,j} \odot t_{i,j} \odot \text{Sigmoid}(s_{i,j})}{\sum_{i=1}^{W_{img}} \sum_{j=1}^{H_{img}} w_{i,j} \odot t_{i,j} + \sum_{i=1}^{W_{img}} \sum_{j=1}^{H_{img}} w_{i,j} \odot \text{Sigmoid}(s_{i,j})}, \end{aligned} \quad (6.19)$$

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 signals and encourages it to prioritize learning features from the cleaner data at the edges. This method is similar to curriculum learning strategies (Bengio et al. 2009), 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 (Shin et al. 2024).

6.3.4 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. 6.1 (c), which partitions an image into K regions of equal area $\mathcal{R} = \{R_1, R_2, \dots, R_K\}$ from the center to edges. The IoU for each region is calculated as follows:

$$\text{IoU}_k = \frac{\|\text{TP}_k\|}{\|\text{TP}_k\| + \|\text{FP}_k\| + \|\text{FN}_k\|}, \quad (6.20)$$

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:

$$\sigma_{\text{IoU}}^2 = \frac{1}{K} \sum_{k=1}^K (\text{IoU}_k - \mu_{\text{IoU}})^2, \quad (6.21)$$

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_{IoU}^2) indicates more severe spatial bias in the segmentation results, and vice versa.

6.4 Experimental Results and Discussions

6.4.1 Datasets

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

6.4.1.1 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 (Ha et al. 2017), using 1568 pairs of images (including flipped images) for training, 392 pairs for validation, and 393 pairs for testing.

6.4.1.2 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. (Y.-H. Kim et al. 2021) annotate 503 pairs of daytime and 447 pairs of nighttime RGB-T images, with 19 semantic classes identical to those in the Cityscapes dataset (Cordts et al. 2016). We adopt the same dataset split method as CRM (Shin et al. 2024), using 499 annotated pairs for training, 140 pairs for validation, and 311 pairs for testing.

6.4.2 Implementation Details

We set the patch size $PS = 64$ in Eq. (6.9), the scaling factor $\alpha = 2$ in Eq. (6.17), and the standard deviation $\sigma_p = N_y/4$ in Eq. (6.10), which is varying with the height of the images in the training set. In Eq. (6.14) and Eq. (6.15), 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 (Ze Liu et al. 2021) 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 an 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.

6.4.3 Ablation Study

Table 6.2: Results (%) of the ablation study on standard deviation (σ_p) in Gaussian distribution. Var. is the variance of IoU across all regions (σ_{IoU}^2). mIoU is the IoU of the whole image. 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

6.4.3.1 Ablation on Gaussian-guided Regional Balancing Masking

To trade-off performance and model complexity, we use the Swin-S transformer (Ze Liu et al. 2021) as the backbone for our ablation 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. (6.10), 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. 6.2. The first row displays the baseline method for comparison, which is CRM (Shin et al. 2024) 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. (6.9) 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

Table 6.3: Results (%) of the ablation study on patch size (PS) in complementary masking generation. Var. is the variance of IoU across all regions (σ_{IoU}^2). mIoU is the IoU of the whole image. 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

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. 6.3. 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.

6.4.3.2 Ablation on Spatial-Weighted Loss

In Eq. (6.17), 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. 6.4. 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,

Table 6.4: Results (%) of the ablation study on scaling factor (α) in spatial-weighted loss. Var. is the variance of IoU across all regions (σ_{IoU}^2). mIoU is the IoU of the whole image. 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

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, as the spatial bias in the segmentation results shows that the performance at the image central regions is better than that at the image edge regions. By emphasizing the attention to edge information in the loss function, the spatial bias in the segmentation results is increased.

Table 6.5: Results (%) of different spatial weighting strategies on the loss function. Var. is the variance of IoU across all regions (σ_{IoU}^2). mIoU is the IoU of the whole image. 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

Tab. 6.5 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. 6.4). The third row adopts the opposite strategy that focuses more on im-

age 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 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 noises 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. 6.4 as our best model.

6.4.4 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.

Table 6.6: The comparative per-class results (%) on the MFNet dataset. We report the Acc (%) and IoU (%) for each class and the mAcc (%) and mIoU (%) for all the classes. The results demonstrate the superiority of our method, with the top two results in each column highlighted 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 (Sun, Zuo, and M. Liu 2019)	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 (Jinyuan Liu et al. 2023)	ICCV'23	MiT-B3	<u>96.3</u>	87.8	89.6	71.4	81.2	63.2	63.5	47.5	66.7	31.1	-	-	85.3	48.9	<u>84.8</u>	50.3	74.8	56.1
CENet (Z. Feng, Guo, and Sun 2023)	RAL'23	ResNet	92.0	85.8	78.9	70.0	74.9	61.4	64.8	46.8	39.8	29.3	<u>65.7</u>	8.7	54.1	47.8	77.1	56.9	71.8	56.1
CACFNet (W. Zhou, Dong, et al. 2023)	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	<u>76.7</u>	57.8
EAEFNet (Liang et al. 2023)	RAL'23	ResNet-152	95.4	87.6	85.2	72.6	79.9	63.8	70.6	<u>48.6</u>	47.9	35.0	62.8	<u>14.2</u>	62.7	52.4	71.9	58.3	75.1	58.9
CMX (J. Zhang, Huayao Liu, et al. 2023)	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 (J. Zhang, R. Liu, et al. 2023)	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 (Lv, Zhi Liu, and G. Li 2024)	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 (Dong et al. 2024)	TITS'24	ConvNeXt	96.5	89.8	<u>92.1</u>	71.6	<u>84.8</u>	63.9	76.1	46.7	44.6	31.3	38.7	6.7	<u>71.1</u>	52.0	78.1	57.4	75.6	57.5
CRM (Shin et al. 2024)	ICRA'24	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
		Swin-S	95.1	<u>90.6</u>	88.5	75.5	78.5	<u>67.2</u>	69.3	48.3	52.3	43.4	62.6	11.8	60.9	<u>56.8</u>	77.2	59.3	76.0	61.2
		Swin-B	95.2	90.8	85.6	74.8	78.1	66.6	<u>74.9</u>	45.9	52.7	<u>43.5</u>	70.7	12.2	65.1	57.4	80.8	62.1	78.0	<u>61.3</u>
Ours		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
		Swin-S	95.2	90.3	85.9	<u>74.9</u>	78.0	66.9	68.1	48.3	<u>60.1</u>	47.6	53.1	15.8	62.1	56.2	80.8	<u>60.9</u>	75.8	62.1

6.4.4.1 Evaluation on MFNet Dataset

We choose Swin-T and Swin-S (Ze Liu et al. 2021) as the backbones for our semantic segmentation network and compare our network with RTFNet (Sun, Zuo, and M. Liu 2019), SegMiF (Jinyuan Liu et al. 2023), CENet (Z. Feng, Guo, and Sun 2023), CACFNet (W. Zhou, Dong, et al. 2023), EAEFNet (Liang et al. 2023), CMX (J. Zhang, Huayao Liu, et al. 2023), CMNeXt (J. Zhang, R. Liu, et al. 2023), CAINet (Lv, Zhi Liu, and G. Li 2024), EGFNet (Dong et al. 2024), and CRM (Shin et al. 2024). The results are shown in Tab. 6.6. Note that several state-of-the-art methods, such as CMX (J. Zhang, Huayao Liu, et al. 2023) and CAINet (Lv, Zhi Liu, and G. Li 2024), employ sophisticated cross-modal attention mechanisms to dynamically learn feature fusion. Despite the sophisticated designs of these modules, our method, which employs a much simpler max-operation fusion, demonstrates superior overall segmentation performance. In addition, our method with the lightweight Swin-S achieves higher mIoU than the CRM with Swin-B. Notably, our method outperforms CRM in terms of IoU for some challenging classes, including Curve, Guardrail, and Bump. However, CRM achieves higher Acc in these classes. This suggests that CRM performs better in identifying positive samples but has lower overlap in the overall predictions, possibly due to more false positives, where it incorrectly predicts negative samples as positive. Our method effectively mitigates this issue.

6.4.4.2 Evaluation on KP Dataset

We select Swin-T, Swin-S, and Swin-B (Ze Liu et al. 2021) as the backbones for our semantic segmentation network, and compare them with MFNet (Ha et

Table 6.7: The comparative per-class IoU (%) on the KP dataset. We report the IoU (%) for each class and the mIoU (%) for all the classes. The results demonstrate the superiority of our method, with the top two results in each column highlighted 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 (Ha et al. 2017)	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	<u>0.3</u>	0.3	0.0	0.0	0.6	24.0
RTFNet (Sun, Zuo, and M. Liu 2019)	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 (J. Zhang, Huayao Liu, et al. 2023)	TITS'23	MiT-B4	97.7	53.8	90.2	0.0	47.1	46.2	10.9	45.1	87.2	<u>34.3</u>	93.5	74.5	0.0	91.6	0.0	59.7	0.0	46.1	0.2	46.2
CRM (Shin et al. 2024)	ICRA'24	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
		Swin-S	98.8	60.7	<u>92.1</u>	0.0	60.4	<u>55.1</u>	31.1	<u>53.2</u>	89.1	27.7	<u>95.0</u>	81.4	17.7	95.2	1.1	<u>83.3</u>	0.0	49.9	42.3	54.4
		Swin-B	99.0	<u>61.8</u>	91.8	0.0	58.7	50.6	39.1	55.4	<u>89.2</u>	23.2	94.3	85.2	2.9	<u>95.3</u>	0.0	80.5	0.0	66.2	54.6	<u>55.1</u>
Ours		Swin-T	98.7	55.0	91.3	0.0	<u>63.8</u>	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
		Swin-S	<u>98.9</u>	62.6	91.6	0.0	57.2	51.4	<u>39.4</u>	52.3	88.3	24.3	94.8	84.5	<u>20.6</u>	95.6	0.0	84.9	0.0	49.3	<u>47.7</u>	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	<u>85.0</u>	39.2	<u>95.3</u>	0.0	77.4	0.0	<u>52.6</u>	38.8	56.8

al. 2017), RTFNet (Sun, Zuo, and M. Liu 2019), CMX (J. Zhang, Huayao Liu, et al. 2023), and CRM (Shin et al. 2024). Tab. 6.7 shows the results. Notably, our method consistently outperforms CRM when using the same backbones. The model with Swin-B as the backbone achieves significant advantages in most classes, especially in improving the IoU of the challenging Rider class by 21.5%.

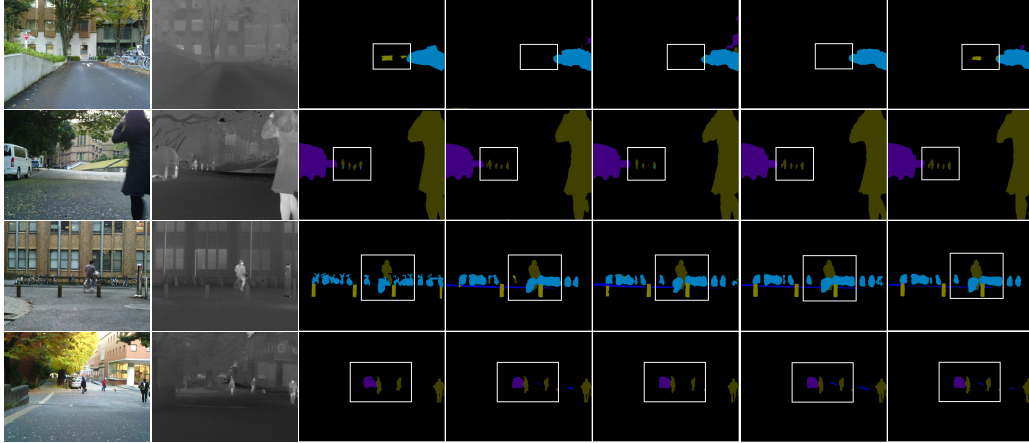


Figure 6.7: Qualitative demonstration for RGB-T semantic segmentation on MFNet (Ha et al. 2017) dataset. From left to right: (1) RGB images, (2) thermal images, (3) ground truth labels, (4) results of CRM with Swin-T backbone, (5) results of our method with Swin-T backbone, (6) results of CRM with Swin-S backbone, (7) results of our method with Swin-S backbone.

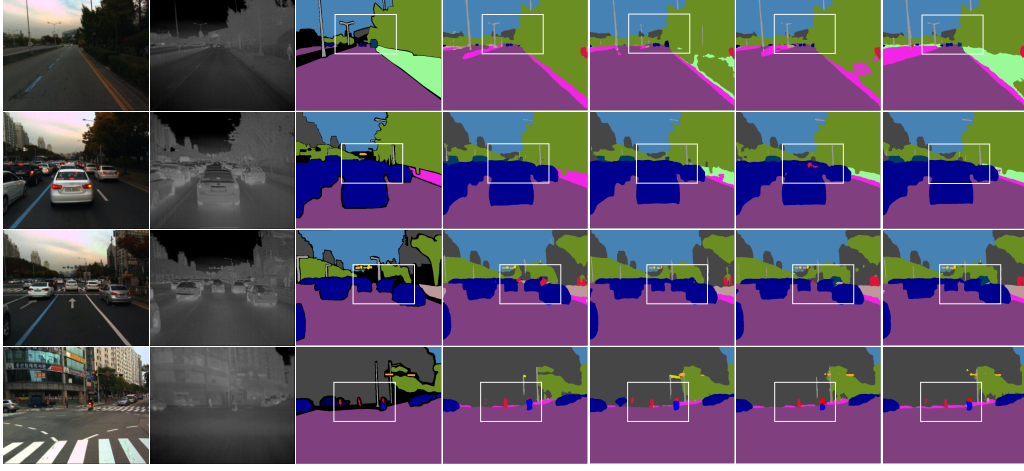


Figure 6.8: Qualitative demonstration for RGB-T semantic segmentation on the KP (Y.-H. Kim et al. 2021) dataset. From left to right: (1) RGB images, (2) thermal images, (3) ground truth labels, (4) results of CRM with Swin-T backbone, (5) results of our method with Swin-T backbone, (6) results of CRM with Swin-B backbone, (7) results of our method with Swin-B backbone.

6.4.4.3 Qualitative Demonstrations

Fig. 6.7 and Fig. 6.8 qualitatively compares the segmentation results of our method and CRM (Shin et al. 2024) on the MFNet (Ha et al. 2017) dataset and the KP (Y.-H. Kim et al. 2021) dataset, respectively.

For the results on the MFNet dataset, the first row shows that only our method with Swin-S successfully detects the 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 superior 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 and superiority of our method, it still has limitations. As indicated in Tab. 6.7, our method’s performance for the Motorcycle class on the KP dataset is lower than the baseline CRM with the same backbone. This scenario is also illustrated in the fourth row of Fig. 6.8. For very small and complex objects, such as the motorcycle highlighted in the box, 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.

6.5 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 of 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 noises, further boosting its performance. Experimental results demonstrate that our method achieves superior performance compared with state-of-the-arts on 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 from autonomous driving. However, we believe that the proposed analytical framework for spatial bias would be still 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.

Chapter 7

Conclusions, Limitations, and Future Work

7.1 Conclusions

This dissertation systematically studies RGB-T semantic segmentation for autonomous driving from four dimensions: illumination-guided fusion, temporal consistency, data augmentation, and spatial balancing, proposing a series of methods to address key challenges in practical applications.

Chapter 3 proposes the IGFNet, which achieves interpretable multi-modal fusion by modeling the illumination characteristics of RGB and thermal images. Its IEM module generates pixel-level weights to quantify RGB reliability, and the IGCM-RM adaptively fuses RGB and thermal features accordingly. It achieves 59.0% mIoU on the MFNet dataset, outperforming existing methods.

Chapter 4 proposes a temporal consistency framework, synthesizing the next frame through a VVIG module. Combining the designed consistency losses and

proposed CA metric, it significantly improves temporal stability while maintaining accuracy. Experimental results show that this method achieves 59.9% mIoU while improving temporal consistency to TC 36.40% and CA 23.29%, and exhibits good scalability.

Chapter 5 proposes SyntheticSeg, constructing a large-scale synthetic RGB-T dataset containing 31,360 image pairs. It also proposes a sampling mechanism that simultaneously considers class scarcity and segmentation difficulty, effectively mitigating class imbalance. The final model achieves an mIoU of 62.0%, with the IoU for rare classes such as guardrail improving from 11.8% to 15.4%, validating the potential of data-centric methods.

Chapter 6 systematically reveals the spatial bias in RGB-T segmentation: despite having more samples in the central region, its performance is lower than that of the edges. The study finds that this bias stems from object complexity and image quality, and proposes a GRBM method and spatial-weighted loss function to alleviate this problem.

This paper, through exploration in four dimensions, systematically improves RGB-T segmentation performance from a baseline of 58.2% to 62.1%. These achievements construct a robust RGB-T semantic segmentation framework for autonomous driving scenarios, providing a theoretical and technical foundation for multi-modal perception research.

7.2 Limitations

Despite its significant contributions, this study has several limitations. First, our SyntheticSeg method generates RGB and thermal images independently, poten-

tially leading to spatial inconsistencies in background regions and affecting the preservation of physical correspondences. Furthermore, the generative model exhibits lower synthesis quality in rare classes due to limited training data, creating a cyclical contradiction between data dependency and insufficient quality.

Regarding temporal consistency, while the current framework validates consistency between consecutive frames, it does not explicitly model long-term temporal relationships. The Euler angle-based transformation of the VVIG module is suitable for short-term camera motion but struggles to capture complex long-term dynamics, such as multi-object interactions and occlusion. Additionally, the framework synthesizes only single frames rather than continuous video streams, limiting its ability to model complex temporal relationships in real-world driving scenarios.

Regarding generalization ability, the experiments are primarily based on the MFNet dataset, which has specific environment and sensor conditions and a limited number of classes, restricting our assessment of the method’s adaptability across a wider range of configurations and class sizes.

Finally, the spatial balancing strategy is designed for forward-looking cameras and is difficult to directly apply to other camera configurations (such as fisheye or multi-camera systems) or other domains (such as aerial or medical imaging). Furthermore, this dissertation only focuses on RGB-T fusion and has not yet explored how to extend it to multimodal sensor combinations such as LiDAR and radar.

7.3 Future Work

Based on the contributions and limitations of this study, future work can be explored in the following directions:

First, a unified multi-dimensional perception framework can be constructed, integrating illumination-guided fusion, temporal consistency modeling, synthetic data augmentation, and spatial balancing strategies to fully leverage the synergistic effects between dimensions. For example, the synthetic quality of key regions can be improved under spatial balancing guidance, or differentiated illumination correction strategies can be applied to different regions.

Second, an end-to-end generation model capable of jointly generating RGB and thermal image pairs should be developed to ensure physical consistency between modalities, thereby improving synthetic data quality and enhancing the segmentation network's ability to model multi-modal complementarity relationships.

Regarding multi-modal robustness and security, further exploration is needed to understand the resistance of thermal modalities to RGB adversarial attacks, develop uncertainty methods that can quantify regional reliability differences, and design a perception architecture capable of stable operation under sensor failure or extreme conditions.

Furthermore, the methods used in this study can be extended to other fields, such as indoor robotics, aerial remote sensing, security monitoring, agricultural monitoring, and medical imaging. Cross-domain validation can identify universally applicable design principles and promote adaptation and optimization for specific scenarios.

In summary, this dissertation promotes the development of RGB-T segmen-

tation in terms of interpretable fusion, temporal consistency, data augmentation, and spatial balancing, laying the foundation for building more robust and universal autonomous perception systems in the future.

References

- Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla (2017). “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.12, pp. 2481–2495.
- Bengio, Yoshua et al. (2009). “Curriculum learning”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML ’09. Montreal, Quebec, Canada: Association for Computing Machinery, pp. 41–48. ISBN: 9781605585161. DOI: 10.1145/1553374.1553380. URL: <https://doi.org/10.1145/1553374.1553380>.
- Caesar, Holger, Jasper Uijlings, and Vittorio Ferrari (2018). “Coco-stuff: Thing and stuff classes in context”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1209–1218.
- Chen, Bike, Chen Gong, and Jian Yang (2019). “Importance-Aware Semantic Segmentation for Autonomous Vehicles”. In: *IEEE Transactions on Intelligent Transportation Systems* 20.1, pp. 137–148. DOI: 10.1109/TITS.2018.2801309.
- Chen, Liang-Chieh, George Papandreou, et al. (2017). “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully

- connected crfs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4, pp. 834–848.
- Chen, Liang-Chieh, Yukun Zhu, et al. (2018). “Encoder-decoder with atrous separable convolution for semantic image segmentation”. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818.
- Cheng, Jingchun et al. (2017). “Segflow: Joint learning for video object segmentation and optical flow”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 686–695.
- Choi, Hyeong Kyu, Joonmyung Choi, and Hyunwoo J Kim (2022). “Tokenmixup: Efficient attention-guided token-level data augmentation for transformers”. In: *Advances in Neural Information Processing Systems* 35, pp. 14224–14235.
- Cordts, Marius et al. (June 2016). “The Cityscapes Dataset for Semantic Urban Scene Understanding”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Cui, Yin et al. (2019). “Class-balanced loss based on effective number of samples”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277.
- Deng, Fuqin et al. (2021). “FEANet: Feature-enhanced attention network for RGB-thermal real-time semantic segmentation”. In: *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, pp. 4467–4473.
- Deng, Jia et al. (2009). “Imagenet: A large-scale hierarchical image database”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 248–255.

- Dhariwal, Prafulla and Alexander Nichol (2021). “Diffusion models beat gans on image synthesis”. In: *Advances in neural information processing systems* 34, pp. 8780–8794.
- Dong, Shaohua et al. (2024). “EGFNet: Edge-Aware Guidance Fusion Network for RGB–Thermal Urban Scene Parsing”. In: *IEEE Transactions on Intelligent Transportation Systems* 25.1, pp. 657–669. DOI: 10.1109/TITS.2023.3306368.
- Dosovitskiy, Alexey et al. (2020). “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929*.
- Feng, Yuchao, Wei Hua, and Yuxiang Sun (2023). “NLE-DM: Natural-Language Explanations for Decision Making of Autonomous Driving Based on Semantic Scene Understanding”. In: *IEEE Transactions on Intelligent Transportation Systems* 24.9, pp. 9780–9791.
- Feng, Yuchao and Yuxiang Sun (2024). “PolarPoint-BEV: Bird-eye-view Perception in Polar Points for Explainable End-to-end Autonomous Driving”. In: *IEEE Transactions on Intelligent Vehicles*, pp. 1–11.
- Feng, Zhen, Yanning Guo, and Yuxiang Sun (2023). “CEKD: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images”. In: *IEEE Robotics and Automation Letters* 8.4, pp. 2205–2212.
- Fu, Jun et al. (2019). “Dual attention network for scene segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3146–3154.
- Ha, Qishen et al. (2017). “MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes”. In: *2017 IEEE/RSJ Interna-*

- tional Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 5108–5115.
- Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). “Denoising diffusion probabilistic models”. In: *Advances in neural information processing systems* 33, pp. 6840–6851.
- Huang, Jianxin et al. (2024). “RoadFormer+: Delivering RGB-X Scene Parsing through Scale-Aware Information Decoupling and Advanced Heterogeneous Feature Fusion”. In: *IEEE Transactions on Intelligent Vehicles*, pp. 1–10. DOI: 10.1109/TIV.2024.3448251.
- Huang, Zilong et al. (2019). “Ccnet: Criss-cross attention for semantic segmentation”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 603–612.
- Karimi, Davood and Septimiu E. Salcudean (2020). “Reducing the Hausdorff Distance in Medical Image Segmentation With Convolutional Neural Networks”. In: *IEEE Transactions on Medical Imaging* 39.2, pp. 499–513. DOI: 10.1109/TMI.2019.2930068.
- Kim, Yeong-Hyeon et al. (2021). “MS-UDA: Multi-Spectral Unsupervised Domain Adaptation for Thermal Image Semantic Segmentation”. In: *IEEE Robotics and Automation Letters* 6.4, pp. 6497–6504.
- Kingma, Diederik P and Jimmy Ba (2014). “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980*.
- Li, Haotian, Henry K. Chu, and Yuxiang Sun (2024). “Temporal Consistency for RGB-Thermal Data-Based Semantic Scene Understanding”. In: *IEEE Robotics and Automation Letters* 9.11, pp. 9757–9764. DOI: 10.1109/LRA.2024.3458594.

- Li, Haotian and Yuxiang Sun (2023). “IGFNet: Illumination-Guided Fusion Network for Semantic Scene Understanding using RGB-Thermal Images”. In: *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1–6. DOI: 10.1109/ROBIO58561.2023.10354613.
- Liang, Mingjian et al. (2023). “Explicit Attention-Enhanced Fusion for RGB-Thermal Perception Tasks”. In: *IEEE Robotics and Automation Letters*.
- Lin, Tsung-Yi et al. (2017). “Focal loss for dense object detection”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988.
- Liu, Jinyuan et al. (2023). “Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 8115–8124.
- Liu, Yifan et al. (2020). “Efficient semantic video segmentation with per-frame inference”. In: *Proceedings of the European conference on computer vision (ECCV)*. Springer, pp. 352–368.
- Liu, Ze et al. (Oct. 2021). “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 10012–10022.
- Lv, Ying, Zhi Liu, and Gongyang Li (2024). “Context-Aware Interaction Network for RGB-T Semantic Segmentation”. In: *IEEE Transactions on Multimedia*.
- Ma, Weixin, Shoudong Huang, and Yuxiang Sun (2024). “Triplet-Graph: Global Metric Localization based on Semantic Triplet Graph for Autonomous Vehicles”. In: *IEEE Robotics and Automation Letters*.

- Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi (2016). “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”. In: *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. DOI: 10.1109/3DV.2016.79.
- Nilsson, David and Cristian Sminchisescu (2018). “Semantic video segmentation by gated recurrent flow propagation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6819–6828.
- Park, Hyojin et al. (2022). “Real-Time, Accurate, and Consistent Video Semantic Segmentation via Unsupervised Adaptation and Cross-Unit Deployment on Mobile Device”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21431–21438.
- Pech-Pacheco, J.L. et al. (2000). “Diatom autofocusing in brightfield microscopy: a comparative study”. In: *Proceedings 15th International Conference on Pattern Recognition*. Vol. 3, 314–317 vol.3. DOI: 10.1109/ICPR.2000.903548.
- Pertuz, Said, Domenec Puig, and Miguel Angel Garcia (2013). “Analysis of focus measure operators for shape-from-focus”. In: *Pattern Recognition* 46.5, pp. 1415–1432.
- Qiu, Shoumeng et al. (2024). “Subclassified Loss: Rethinking Data Imbalance From Subclass Perspective for Semantic Segmentation”. In: *IEEE Transactions on Intelligent Vehicles* 9.1, pp. 1547–1558. DOI: 10.1109/TIV.2023.3325343.
- Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241.

- Rota Bulo, Samuel, Gerhard Neuhold, and Peter Kotschieder (July 2017). “Loss Max-Pooling for Semantic Image Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Russakovsky, Olga et al. (2015). “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115, pp. 211–252.
- Sáez, José A et al. (2015). “SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering”. In: *Information Sciences* 291, pp. 184–203.
- Shin, Ukcheol et al. (2024). “Complementary Random Masking for RGB-Thermal Semantic Segmentation”. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11110–11117. DOI: 10.1109/ICRA57147.2024.10611200.
- Song, Kechen et al. (2023). “RGB-T image analysis technology and application: A survey”. In: *Engineering Applications of Artificial Intelligence* 120, p. 105919.
- Strudel, Robin et al. (2021). “Segmenter: Transformer for semantic segmentation”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 7262–7272.
- Sun, Yuxiang, Weixun Zuo, and Ming Liu (2019). “RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes”. In: *IEEE Robotics and Automation Letters* 4.3, pp. 2576–2583.
- Sun, Yuxiang, Weixun Zuo, Peng Yun, et al. (2021). “FuseSeg: Semantic Segmentation of Urban Scenes Based on RGB and Thermal Data Fusion”. In: *IEEE Transactions on Automation Science and Engineering* 18.3, pp. 1000–1011.
- Taghanaki, Saeid Asgari et al. (2019). “Combo loss: Handling input and output imbalance in multi-organ segmentation”. In: *Computerized Medical Imaging*

- and Graphics* 75, pp. 24–33. ISSN: 0895-6111. DOI: <https://doi.org/10.1016/j.compmedimag.2019.04.005>.
- Teng, Siyu et al. (2023). “Motion Planning for Autonomous Driving: The State of the Art and Future Perspectives”. In: *IEEE Transactions on Intelligent Vehicles* 8.6, pp. 3692–3711.
- Tian, Junjiao et al. (2022). “Striking the Right Balance: Recall Loss for Semantic Segmentation”. In: *2022 International Conference on Robotics and Automation (ICRA)*, pp. 5063–5069. DOI: 10.1109/ICRA46639.2022.9811702.
- Touvron, Hugo et al. (2021). “Training data-efficient image transformers & distillation through attention”. In: *International conference on machine learning*. PMLR, pp. 10347–10357.
- Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.
- Venkataaramanan, Shashanka et al. (2022). “Alignmixup: Improving representations by interpolating aligned features”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19174–19183.
- Wang, Weilun et al. (2022). “Semantic image synthesis via diffusion models”. In: *arXiv:2207.00050*.
- Xie, Enze et al. (2021). “SegFormer: Simple and efficient design for semantic segmentation with transformers”. In: *Advances in neural information processing systems* 34, pp. 12077–12090.
- Xu, Huaiyuan, Huaping Liu, Shoudong Huang, et al. (2024). “C2L-PR: Cross-modal Camera-to-LiDAR Place Recognition via Modality Alignment and Orientation Voting”. In: *IEEE Transactions on Intelligent Vehicles*, pp. 1–17.

- Xu, Huaiyuan, Huaping Liu, Shiyu Meng, et al. (2023). “A Novel Place Recognition Network using Visual Sequences and LiDAR Point Clouds for Autonomous Vehicles”. In: *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2862–2867.
- Xu, Jiangtao, Kaige Lu, and Han Wang (2021). “Attention fusion network for multi-spectral semantic segmentation”. In: *Pattern Recognition Letters* 146, pp. 179–184.
- Xue, Han et al. (2023). “Freestyle layout-to-image synthesis”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14256–14266.
- Zhang, Jiaming, Huayao Liu, et al. (2023). “CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers”. In: *IEEE Transactions on Intelligent Transportation Systems*.
- Zhang, Jiaming, Ruiping Liu, et al. (2023). “Delivering Arbitrary-Modal Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1136–1147.
- Zhang, Qiang et al. (2021). “ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2633–2642.
- Zhang, Yizhe et al. (2022). “Auxadapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2339–2348.

- Zheng, Sixiao et al. (2021). “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6881–6890.
- Zheng, Zhaohui et al. (2024). “Zone Evaluation: Revealing Spatial Bias in Object Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12, pp. 8636–8651. DOI: 10.1109/TPAMI.2024.3409416.
- Zhou, Bolei et al. (2017). “Scene parsing through ade20k dataset”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 633–641.
- Zhou, Heng et al. (2025). “Deformation-Resilient Multigranularity Learning for Unaligned RGB–T Semantic Segmentation”. In: *IEEE Transactions on Neural Networks and Learning Systems* 36.10, pp. 18530–18544. DOI: 10.1109/TNNLS.2025.3585105.
- Zhou, Wujie, Shaohua Dong, et al. (2023). “CACFNet: Cross-modal attention cascaded fusion network for RGB-T urban scene parsing”. In: *IEEE Transactions on Intelligent Vehicles*.
- Zhou, Wujie, Jinfu Liu, et al. (2021). “GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation”. In: *IEEE Transactions on Image Processing* 30, pp. 7790–7802.
- Zhou, Wujie, Han Zhang, et al. (2023). “MMSMCNet: Modal memory sharing and morphological complementary networks for RGB-T urban scene semantic segmentation”. In: *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhou, Xiaofei et al. (2025). “AGFNet: Adaptive Gated Fusion Network for RGB-T Semantic Segmentation”. In: *IEEE Transactions on Intelligent Transportation Systems* 26.5, pp. 6477–6492. DOI: 10.1109/TITS.2025.3528064.

Appendix

Assume that a 3-D object point P is captured in two consecutive frames (Img_1 and Img_2), and the pixels for P on the two images are p_1 and p_2 . From the pinhole camera model, we have: $p_1 \simeq KP$, $p_2 \simeq K(RP + t)$, where K is the camera intrinsic matrix, R and t are respectively the rotation matrix and translation vector between the consecutive frames, $\simeq$ denotes equivalence up to a scale.

Assume that the camera motion includes only rotations (i.e., $t = 0$), we have $p_2 \simeq KRP$. Since $P \simeq K^{-1}p_1$, we have $p_2 \simeq KRK^{-1}p_1$. Since $R = R_z(\alpha)R_y(\beta)R_x(\gamma)$, we have $p_2 \simeq KR_z(\alpha)R_y(\beta)R_x(\gamma)K^{-1}p_1$, where α, γ, β are respectively the Euler angles about the $\hat{Z}, \hat{Y}, \hat{X}$ axes. Since $p_2 \simeq M_{\text{VVT}}p_1$, we have $M_{\text{VVT}} = KR_z(\alpha)R_y(\beta)R_x(\gamma)K^{-1}$.

Comparing M_{VVT} to the homography matrix $H = K(R - \frac{tn^T}{d})K^{-1}$, where n^T and d are respectively the normal vector and the distance to the plane from the origin point, we can find that $M_{\text{VVT}} = H$ if there is no translational movement (i.e., $t = 0$). So, we use the homography matrix as the VVT matrix in this work.