electronics

MDPI

Article

Dynamic Frequency Separated Fusion Network for RGB-T Semantic Segmentation

Shaochuan Zhao [1], Xiaotian Wang [1], Jiawei Jiang [2], Hefeng Yin [3,*], Bing Liu [1] and Yong Zhou [1]

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China; shaochuan_zhao@cumt.edu.cn (S.Z.); wangxiaotian@cumt.edu.cn (X.W.); liubing@cumt.edu.cn (B.L.); yzhou@cumt.edu.cn (Y.Z.)
[2] School of Artificial Intelligence and Computer Science, Chongqing Institute of Engineering, Chongqing 404100, China; jiangjiawei@cqie.edu.cn
[3] School of Automation, Wuxi University, Wuxi 214105, China
* Correspondence: yinhefeng@cwxu.edu.cn

Abstract

RGB-T semantic segmentation fuses visible (RGB) and thermal infrared (TIR) images to enhance model robustness under complex lighting conditions. However, most existing methods employ unified fusion strategies in the spatial domain, limiting their ability to fully exploit the complementary characteristics of the two modalities across different frequency components. To address this issue, we propose a novel fusion model that achieves finer cross-modal collaboration through frequency-aware feature disentanglement and adaptive fusion mechanisms. Specifically, the model first enhances the multi-granularity representation of dual-modality features using attention mechanisms. The features are then transformed into the frequency domain, where a dynamic band disentanglement mechanism adaptively separates high-frequency details from low-frequency semantics based on the input content. A heterogeneous fusion pathway is further designed to differentially aggregate high- and low-frequency components, emphasizing cross-modal detail alignment in the high-frequency path and semantic consistency in the low-frequency path. The proposed model is evaluated on three widely-used RGB-T segmentation benchmarks, namely MFNet, PST900, and FMB, and achieves state-of-the-art performance, demonstrating its effectiveness and superiority.

Keywords: semantic segmentation; frequency-domain learning; multi-modal

Check for updates

Academic Editors: Donglin Zhang and Zhen Liu

Received: 4 October 2026
Revised: 3 November 2026
Accepted: 4 November 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Semantic segmentation assigns a semantic label to each pixel in an image, playing a critical role in applications such as autonomous driving [1], robotic perception [2], and medical image analysis [3]. Significant progress has been made in RGB-based semantic segmentation, largely driven by the development of fully convolutional networks [4] and deep convolutional neural networks [5], particularly on large-scale benchmarks like Cityscapes [6], where breakthrough performance has been achieved. However, these methods often suffer from severe performance degradation under adverse conditions—such as nighttime or poor weather—where RGB image quality deteriorates. This limitation stems from the high dependence of RGB sensors on external illumination [7,8], rendering monomodal RGB systems inherently unstable for all-weather and all-day segmentation tasks. Consequently, RGB-only approaches struggle to deliver consistent and reliable results in real-world deployment scenarios.

To overcome this limitation, researchers have increasingly turned to multi-modal methods that combine RGB with additional sensing modalities. Among them, thermal infrared (TIR) sensors have drawn significant attention due to their robustness against lighting variations and adverse weather conditions [9,10]. Unlike RGB cameras, TIR sensors capture thermal radiation emitted by objects, enabling stable perception across diverse and challenging environments. With the growing availability of aligned RGB-T datasets, RGB-T semantic segmentation has emerged as a promising research direction, particularly suitable for low-light or complex outdoor scenarios.

Mainstream RGB-T segmentation methods [11–15] typically employ separate encoders to extract unimodal features from RGB and TIR inputs. Given the significant modality gap, recent efforts have focused on cross-modal fusion strategies in the embedding space, aiming to exploit complementary information through spatial or channel-wise correlations. These fusion schemes can be broadly categorized into two groups: (1) applying a unified fusion mechanism across different feature levels to jointly integrate low-level details and high-level semantics [16]; and (2) designing hierarchical fusion strategies tailored to specific stages of the feature hierarchy [14].

Although the aforementioned methods effectively integrate multi-modal information in the feature space, they largely implicitly assume that the two modalities exhibit a certain degree of symmetry or uniform complementarity in their information distribution. However, empirical observations reveal that RGB and TIR modalities demonstrate significantly asymmetric complementary characteristics in the frequency domain. We conduct a statistical analysis of radial DCT energy distributions on the training sets of three mainstream RGB–TIR datasets, with results shown in Figure 1. A highly consistent pattern emerges across all datasets: in the ultra-low-frequency region, the energy spectra of RGB and TIR nearly overlap, indicating strong agreement in coarse-grained semantics such as overall structure, object silhouettes, and global thermal distribution. In contrast, a persistent and substantial gap exists between the two modalities in the mid- to high-frequency regions, revealing that high-frequency components, such as textures and fine details, exhibit markedly greater divergence. To further confirm that this asymmetry persists at the feature level where fusion actually operates, as shown in Table 1, we compute the position-wise spectral cosine similarity between RGB and TIR features at each encoder stage on the MFNet dataset. The results show a consistent trend: low-frequency cosine similarity increases from 0.71 (Stage 1) to 0.88 (Stage 4), while high-frequency similarity decreases from 0.43 to 0.23, indicating that the two modalities share stronger semantic agreement in the low-frequency band and become increasingly divergent in the high-frequency band as features deepen. These observations jointly motivate our design of distinct fusion strategies for low- and high-frequency components.

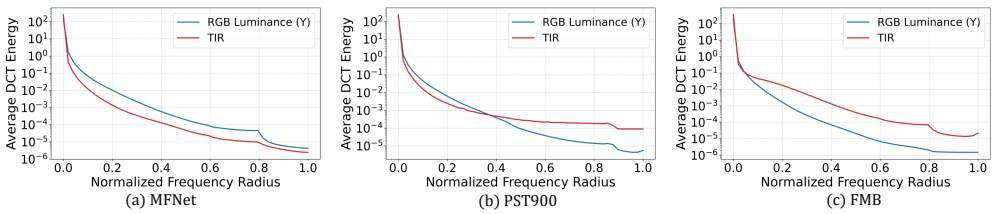


Figure 1. The average radial DCT energy distributions of the RGB luminance channel and thermal infrared images computed on the training sets of MFNet, PST900 and FMB. The vertical axis uses a log scale to represent energy magnitude. The horizontal axis denotes the normalized frequency radius ranging from zero for the DC component to one for the highest frequency. All three datasets consistently show that the two modalities exhibit nearly identical energy in the ultra low-frequency region while their energy distributions diverge significantly in the mid- to high- frequency range.

Table 1. Position-wise spectral cosine similarity between RGB and TIR features at various encoder stages on MFNet.

	Stage 1	Stage 2	Stage 3	Stage 4
Low frequency	0.71	0.77	0.83	0.88
High frequency	0.43	0.37	0.30	0.23

This observation highlights a fundamental limitation of current fusion paradigms: most existing methods perform fusion directly in the spatial domain on raw features, without accounting for the distinct modality-specific characteristics embedded in different frequency bands. Consequently, they fail to fully exploit the complementary advantages inherent in the frequency-domain representations of RGB and TIR.

Based on the above-mentioned issue and inspired by the growing interest in frequency-domain analysis [17–20] in vision tasks, we propose a novel fusion model that explicitly leverages the frequency spectrum to enhance cross-modal collaboration. Our approach first strengthens the representational power of dual-modality features through multi-granularity attention mechanisms. Subsequently, instead of performing fusion in the spatial domain, we transform the features into the frequency domain and introduce a dynamic band disentanglement mechanism that adaptively separates high- and low-frequency components based on input content. This enables content-aware frequency decomposition, avoiding fixed or handcrafted thresholds. We then design a heterogeneous fusion pathway to differentially aggregate the separated bands: cross-attention is employed in the high-frequency path to align fine-grained structural details across modalities, while lightweight MLP fusion is applied in the low-frequency path to consolidate global semantic consistency. This dual-path design respects the distinct semantic roles of different frequency components, enabling more precise and context-aware fusion.

We evaluate the proposed model on three widely used RGB-T segmentation benchmarks, MFNet [9], PST900 [21] and FMB [22]. Experimental results demonstrate that our method outperforms existing approaches across multiple metrics, achieving state-of-the-art performance. Qualitative analysis further reveals improved boundary accuracy and enhanced robustness in challenging scenarios.

In summary, the main contributions of this work are as follows:

- We propose a frequency-aware feature disentanglement mechanism to adaptively decompose multi-modal features into high-frequency and low-frequency components, enabling context-aware modeling in the frequency domain.
- We design a dual-path heterogeneous fusion architecture that performs differentiated fusion of multi-modal information according to the frequency characteristics of features, enhancing cross-modal alignment and complementarity of fine details in the high-frequency branch and ensuring semantic consistency in the low-frequency branch.
- We achieve state-of-the-art performance on the MFNet, PST900, and FMB benchmarks, thoroughly validating the effectiveness of frequency-domain adaptive fusion and providing a new technical pathway for multi-modal semantic segmentation.

2. Related Work

2.1. RGB-T Semantic Segmentation

RGB-Thermal semantic segmentation has gained increasing attention in computer vision, driven by the robustness of thermal imaging under challenging conditions such as low illumination and adverse weather. Thermal infrared sensors capture emitted radiation rather than reflected light, which makes them insensitive to lighting variations and capable of revealing object presence even when visible cues are degraded or absent. This physical

complementarity between RGB and thermal modalities provides a strong motivation for their joint use in scene understanding tasks.

Early approaches typically rely on simple fusion strategies, such as element-wise addition [10] and feature concatenation [21]. These methods are straightforward to implement and can be easily integrated into standard network architectures. However, they treat both modalities as interchangeable sources of information and perform fusion without accounting for their distinct signal characteristics. As a result, they often fail to fully exploit the cross-modal complementarity that arises from the different physical origins of RGB and thermal data.

As research advances, a growing body of work has shifted focus toward modeling high-level semantic interactions between RGB and TIR modalities. Cross-modal attention mechanisms [2] enable dynamic weighting of features based on contextual relevance, allowing the network to adaptively select the most reliable modality at each spatial location. Reciprocal feature enhancement [23] promotes bidirectional interaction by iteratively refining representations through mutual guidance. The incorporation of generative adversarial network insights [24] helps align feature distributions across modalities in latent spaces, improving consistency without requiring pixel-level correspondence. Edge-aware supervision [14] further guides networks to preserve structural boundaries by introducing auxiliary geometric constraints during training.

Despite these advancements, the majority of existing methods perform fusion directly in the spatial domain, with limited consideration of the distinct frequency response characteristics across modalities. Due to fundamental differences in imaging mechanisms [1], RGB and TIR features exhibit heterogeneous distributions across frequency bands. RGB images contain rich high-frequency content corresponding to textures, edges, and fine details, while thermal images concentrate most of their energy in the ultra-low-frequency region, reflecting coarse object shapes and global thermal patterns. Naive fusion in the spatial domain implicitly assumes uniform representational properties across all frequencies, which can lead to redundancy in low-frequency semantics and interference or loss of discriminative details in high-frequency components.

To address this, we propose a novel frequency-domain fusion paradigm in which multi-modal features are first transformed into the frequency domain and adaptively disentangled into high- and low-frequency components. Heterogeneous fusion pathways are then designed to process each component according to its distinct characteristics. The low-frequency pathway emphasizes semantic alignment to reinforce shared structural cues, while the high-frequency pathway preserves modality-specific details through selective enhancement. This differential treatment enables adaptive and frequency-aware fusion that respects the intrinsic asymmetry between RGB and TIR signals. The proposed approach not only improves consistency at both the detail and semantic levels but also introduces a new perspective for modeling cross-modal interactions in RGB-T semantic segmentation by explicitly leveraging spectral complementarity.

2.2. Learning in Frequency Domain

In classical image processing, frequency-domain analysis based on transforms such as the Fourier transform and the discrete cosine transform has played a foundational role in spectral characterization, filter design, and image compression [25]. These tools enable the decomposition of images into constituent frequency components, revealing how energy is distributed across spatial scales. More recently, this analytical perspective has been integrated into deep learning frameworks [26,27], giving rise to frequency-aware modeling as a powerful paradigm for interpreting and refining internal feature representations in neural networks.

Through systematic frequency-domain investigations, researchers have identified a consistent spectral bias in deep models [28,29]. During optimization, these models exhibit a strong tendency to learn low-frequency patterns first, such as global scene layout and coarse semantic structures, while converging slowly or inadequately on high-frequency content that encodes fine details like object boundaries, textures, and sharp transitions. This imbalance becomes especially problematic in dense prediction tasks such as semantic segmentation, where accurate delineation of object contours is critical. As a consequence, model predictions often suffer from blurred edges, fragmented regions, or geometric inconsistencies, particularly under challenging imaging conditions.

To address this limitation, recent efforts have explored frequency-domain feature optimization to recalibrate the network's sensitivity to fine-scale structures. Certain approaches [17,30] introduce frequency-aware attention mechanisms that selectively amplify responses in informative spectral channels, thereby enhancing the representation of underlearned frequencies. Other methods [31] reconstruct features directly in the frequency domain to recover lost high-frequency information through inverse transformation. Furthermore, frequency-guided feature selection and enhancement strategies [20,32] have been developed to improve robustness against noise and degradation by prioritizing stable spectral components.

Despite these promising developments, nearly all existing frequency-aware techniques are designed for single-modality settings. There remains a significant gap in understanding how to leverage the distinct yet complementary spectral signatures of heterogeneous modalities—such as visible and thermal infrared—in the context of multi-modal semantic segmentation. The unique frequency distributions of RGB and TIR data suggest that a unified spatial fusion strategy may obscure their intrinsic advantages, whereas a frequency-resolved approach could unlock more principled cross-modal synergy.

Motivated by this insight, we propose a frequency-domain feature fusion framework specifically tailored for RGB-T semantic segmentation. Our method begins by transforming modality-specific features into the frequency domain and adaptively decomposing them into low-frequency and high-frequency components. We then apply a differential fusion strategy that aligns with the functional roles of each spectral band: low-frequency components undergo semantic alignment to reinforce consistent global understanding, while high-frequency components are fused by dynamically selecting complementary information to enhance edge sharpness and textural complementarity. This decouple-then-fuse mechanism enables a more precise and interpretable modeling of cross-modal interactions, leading to substantial improvements in segmentation accuracy, boundary sharpness, and robustness across diverse and adverse environmental conditions.

3. Method

3.1. Network Overview

Figure 2 illustrates the overall architecture of our network, which mainly consists of the following three key components:

- Independent Encoders: This paper employs two independent ConvNeXt [33] backbone networks to extract hierarchical features from RGB and thermal infrared images, denoted as F_{rgb} and F_{tir} , respectively.
- Two-Stage Feature Fusion: We design a two-stage feature fusion module. In the first stage, a Spatial and Channel Enhancement module employs spatial attention and channel attention to enhance the features of both modalities. In the second stage, a Dynamic Frequency Separation and Fusion module adaptively separates high-frequency and low-frequency components based on the input content, then designs heterogeneous fusion paths according to their frequency characteristics, enabling differential

fusion to enhance semantic consistency while adaptively achieving complementary detail integration.

- Decoder: We adopt a Transformer-based decoder [34] that progressively reconstructs the final prediction by leveraging the multi-scale fused features.

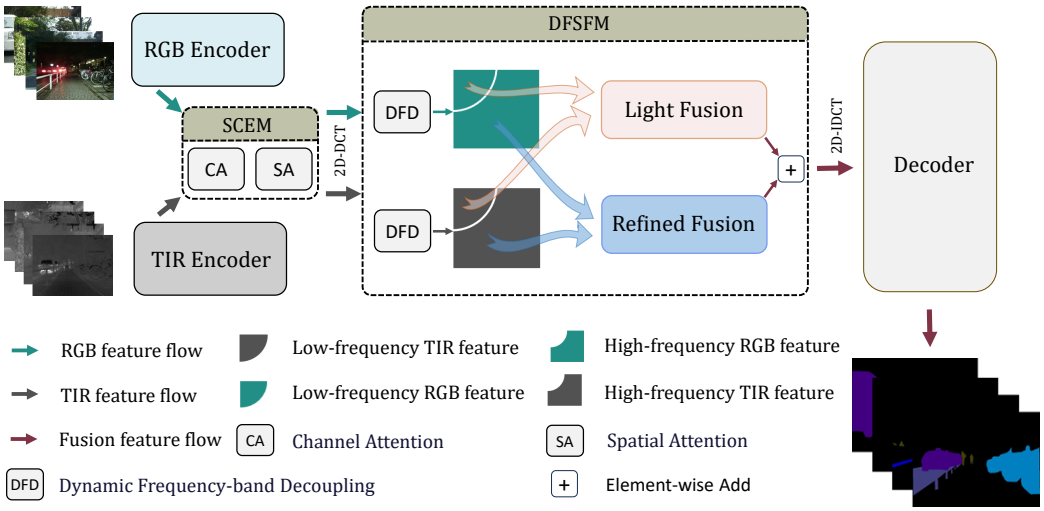


Figure 2. The overall architecture of our proposed Dynamic Frequency Separated Fusion Network for RGB-T semantic segmentation.

3.2. Spatial and Channel Enhancement Module

In multi-modal perception, features from different modalities exhibit variations and redundancies in spatial distribution and channel responses. To enhance key features and promote inter-modal synergy, we designed a spatial and channel enhancement module that optimizes feature maps through a dual attention mechanism. The detailed structure of this module is shown in Figure 3.

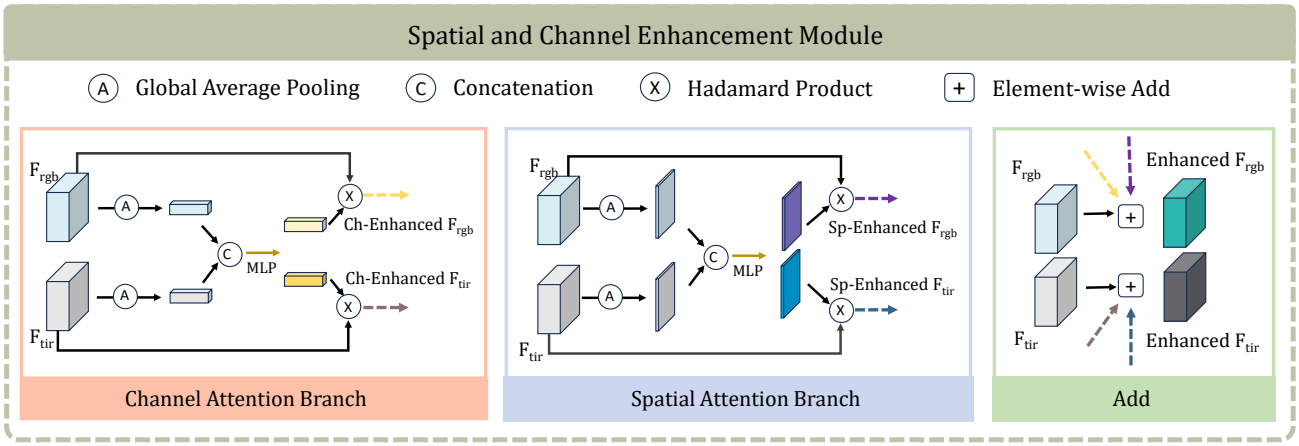


Figure 3. Spatial and Channel Enhancement Module.

Channel Attention Branch: Given input feature $F_{rgb} \in \mathbb{R}^{H\times W\times C}$ and $F_{tir} \in \mathbb{R}^{H\times W\times C}$ maps from two modalities, we first apply global average pooling along the channel dimension for each to capture channel-wise statistics. The resulting vectors are concatenated along the channel axis to form a 2C-dimensional compact descriptor $z_c \in \mathbb{R}^{2C}$. A shared MLP is then applied to transform z_c into a 2C-dimensional attention embedding, which is passed through a Sigmoid function to produce the joint channel attention weights $W_c \in \mathbb{R}^{2C}$. This

vector is evenly split into two parts $W_c^{rgb} \in \mathbb{R}^C$ and $W_c^{tir} \in \mathbb{R}^C$ which are used to weight F_{rgb} and F_{tir}:

$$\hat{F}_{rgb}^c, \hat{F}_{tir}^c = \text{Split}\Big(\sigma\Big(\mathcal{F}_{\text{mlp}}(z_c)\Big)\Big) \odot \Big(F_{rgb}, F_{tir}\Big), \tag{1}$$

where σ is the sigmoid function, Split divides the output vector into two equal parts for RGB and TIR branches, and $\odot$ denotes element-wise multiplication applied channel-wise to weight the input features.

Spatial Attention Branch: Similar to the Channel Attention Branch, given the original input features $F_{rgb} \in \mathbb{R}^{H\times W\times C}$ and $F_{tir} \in \mathbb{R}^{H\times W\times C}$, we first apply global average pooling along the channel dimension to obtain $\bar{F}_{rgb} \in \mathbb{R}^{H\times W\times 1}$ and $\bar{F}_{tir} \in \mathbb{R}^{H\times W\times 1}$. These are then concatenated along the channel axis to form a joint spatial descriptor $z_s \in \mathbb{R}^{H\times W\times 2}$. The spatial attention weights are generated through a lightweight network, and the final reweighted features are expressed as:

$$\hat{F}_{rgb}^s, \hat{F}_{tir}^s = \text{Split}\Big(\sigma\Big(\mathcal{F}_{\text{mlp}}(z_s)\Big)\Big) \odot \Big(F_{rgb}, F_{tir}\Big), \tag{2}$$

where σ denotes the sigmoid function, and $\odot$ represents element-wise multiplication applied across the spatial dimensions.

Then, the bi-modal enhanced features are represented as:

$$\hat{F}_{rgb} = F_{rgb} + \hat{F}_{rgb}^c + \hat{F}_{rgb}^s,\ \hat{F}_{tir} = F_{tir} + \hat{F}_{tir}^c + \hat{F}_{tir}^s. \tag{3}$$

3.3. Dynamic Frequency Separated Fusion Module

In multi-modal perception, RGB images excel at capturing rich texture and color information, while thermal infrared (TIR) images are insensitive to illumination changes and exhibit stronger environmental robustness. However, traditional fusion methods typically perform feature fusion directly in the spatial domain, which can lead to interference from high-frequency details or confusion of low-frequency semantic information across modalities, making it difficult to fully exploit the complementary nature between the two modalities.

To address this issue, we propose a Dynamic Frequency Separation Fusion Module, which enables more fine-grained and efficient feature interaction by decoupling and differentially fusing dual-modal features in the frequency domain. The core idea of the module is that different frequency components carry distinct types of visual information: low-frequency components primarily contain structural and semantic content, making them suitable for robust fusion via channel-wise transformation applied independently at each position; in contrast, high-frequency components correspond to edges and textural details, and are thus better suited for refined fusion via global modeling mechanisms to achieve cross-modal alignment and enhancement.

Specifically, as shown in Figure 4, given the enhanced features from the RGB and thermal infrared (TIR) modalities, denoted as $\hat{\mathbf{F}}_{rgb}$ and $\hat{\mathbf{F}}_{tir}$, which have been processed by spatial and channel enhancement modules, we first apply the 2D Discrete Cosine Transform (2D-DCT) to convert these spatial-domain features into the frequency domain:

$$\tilde{F}_{\text{rgb}} = \mathcal{J}_{\text{DCT}}(\hat{\mathbf{F}}_{rgb}), \tilde{F}_{\text{tir}} = \mathcal{J}_{\text{DCT}}(\hat{\mathbf{F}}_{tir}), \tag{4}$$

where $\mathcal{J}_{\text{DCT}}$ denotes the 2D-DCT operation.

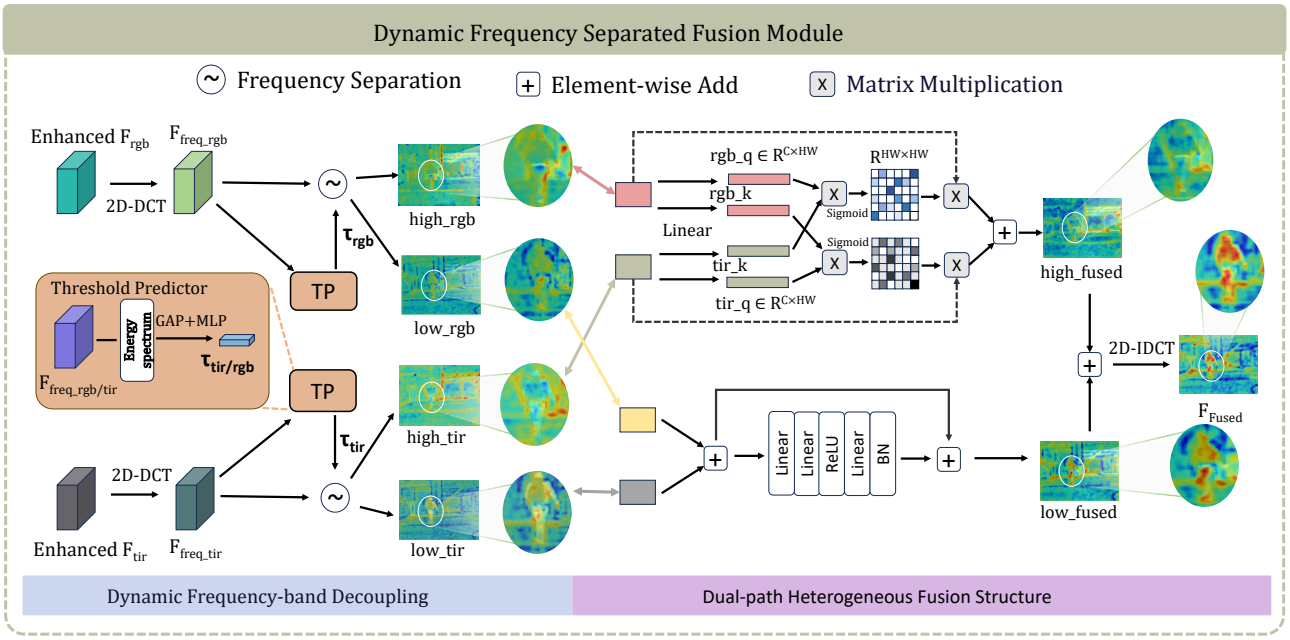


Figure 4. Dynamic Frequency Separated Fusion Module.

For each transformed feature map, the energy spectrum is computed as $|\tilde{F}|^2$, followed by global average pooling to capture the overall frequency distribution characteristics. Based on the pooled energy vector, a lightweight MLP network $\mathcal{M}_{\text{thres}}$ is employed to dynamically predict a modality-specific separation threshold:

$$\tau_{\text{rgb}} = \sigma(\mathcal{M}_{\text{thres}}\left(\text{GAP}(|\tilde{F}_{\text{rgb}}|^2)\right)), \tau_{\text{tir}} = \sigma(\mathcal{M}_{\text{thres}}\left(\text{GAP}(|\tilde{F}_{\text{tir}}|^2)\right)), \tag{5}$$

where σ denotes the Sigmoid function, $\mathcal{M}_{\text{thres}}$ denotes the multi-layer perception responsible for threshold prediction.

The decomposition of each frequency-domain feature into low-frequency and high-frequency components is achieved through a differentiable soft separation strategy to support backpropagation during training. Given a frequency-domain feature $\tilde{F} \in \mathbb{R}^{H\times W\times C}$, we first compute the distance matrix $\|\omega\| \in \mathbb{R}^{H\times W\times 1}$ from each frequency location (u, v) to the zero-frequency center. Subsequently, the modality-specific cut-off frequency $\tau \in \mathbb{R}^{C\times 1\times 1}$, predicted by a lightweight MLP, is expanded to a tensor $\tau' \in \mathbb{R}^{C\times H\times W}$ with the same shape as the feature. Finally, a differentiable low-pass mask is generated using the Sigmoid function:

$$M^{\text{low}} = \sigma\left(\frac{\tau' - \|\omega\|'}{\delta}\right), \tag{6}$$

where $\delta > 0$ is a temperature coefficient that controls the smoothness of the mask, and $\|\omega\|'$ is broadcast along the channel dimension of $\|\omega\|$ to match the shape of τ'. The high-pass mask is then defined as: $M^{\text{high}} = 1 - M^{\text{low}}$. Using these masks, the low- and high-frequency components are separated via element-wise multiplication:

$$\tilde{F}^{\text{low}}_{rgb/tir} = M^{\text{low}} \odot \tilde{F}_{rgb/tir}, \tilde{F}^{\text{high}}_{rgb/tir} = M^{\text{high}} \odot \tilde{F}_{rgb/tir}, \tag{7}$$

This soft filtering mechanism remains continuously differentiable throughout the forward pass, enabling end-to-end training while preserving the physical interpretability of frequency-domain disentanglement.

Low-frequency components primarily encode the global structure, contours, and semantic content of an image, exhibiting strong cross-modal consistency. To effectively fuse the shared semantic information between RGB and TIR modalities while preserving the smoothness of frequency-domain features, we adopt a residual formulation for low-frequency fusion:

$$\tilde{F}_{fuse}^{\text{low}} = \left(\tilde{F}_{rgb}^{low} + \tilde{F}_{tir}^{low}\right) + \mathcal{F}_{mlp}\left(\tilde{F}_{rgb}^{low} + \tilde{F}_{tir}^{low}\right), \tag{8}$$

where $\mathcal{F}_{\text{mlp}}(\cdot)$ consists of three linear layers, batch normalization, and ReLU activation functions. In this residual formulation, the first term directly preserves the original DCT coefficients with their full sign information intact, thereby retaining the important spatial phase and structural information encoded in the coefficient signs. The second term serves only as a learned residual refinement that adaptively adjusts the fused representation through position-wise channel transformations, providing a complementary non-negative correction signal that enhances the fused features.

In contrast, high-frequency components contain rich local details, edges, and texture information, but often exhibit significant distribution shifts due to differences in imaging mechanisms between modalities. To achieve fine-grained alignment and enhance cross-modal complementarity, we introduce a cross-attention-based fusion module. Given the high-frequency components $\tilde{F}_{\text{rgb}}^{\text{high}} \in \mathbb{R}^{H\times W\times C}$ and $\tilde{F}_{\text{tir}}^{\text{high}} \in \mathbb{R}^{H\times W\times C}$, we first reshape their spatial dimensions from $H \times W \times C$ to $HW \times C$. The bidirectional cross-attention is then formulated as: $\tilde{F}_{\text{rgb}}^{\text{high}} \in \mathbb{R}^{H\times W\times C}$ and $\tilde{F}_{\text{tir}}^{\text{high}} \in \mathbb{R}^{H\times W\times C}$, we first reshape their spatial dimensions from $H \times W \times C$ to $HW \times C$. The bidirectional cross-attention is then formulated as:

$$\begin{aligned}
\bar{F}_{\text{tir}}^{\text{high}} &= \text{LN}\left(\sigma\left(\frac{(\tilde{F}_{\text{rgb}}^{\text{high}} W_{\text{tir}}^{Q})(\tilde{F}_{\text{tir}}^{\text{high}} W_{\text{tir}}^{K})^{\top}}{\sqrt{d}}\right)\tilde{F}_{\text{tir}}^{\text{high}} W_{\text{tir}}^{V}\right),\\
\bar{F}_{\text{rgb}}^{\text{high}} &= \text{LN}\left(\sigma\left(\frac{(\tilde{F}_{\text{tir}}^{\text{high}} W_{\text{rgb}}^{Q})(\tilde{F}_{\text{rgb}}^{\text{high}} W_{\text{rgb}}^{K})^{\top}}{\sqrt{d}}\right)\tilde{F}_{\text{rgb}}^{\text{high}} W_{\text{rgb}}^{V}\right),\\
\tilde{F}_{\text{fuse}}^{\text{high}} &= \bar{F}_{\text{tir}}^{\text{high}} + \bar{F}_{\text{rgb}}^{\text{high}},
\end{aligned} \tag{9}$$

where $W^{Q}, W^{K}, W^{V} \in \mathbb{R}^{C\times d}$ denote the learnable query, key, and value projections, d is the projection dimension, and $\text{LN}(\cdot)$ is layer normalization. We adopt the Sigmoid function instead of the conventional Softmax for attention weight computation. Softmax enforces a competition among all elements in the attention matrix, forcing the weights to sum to 1. Consequently, strong responses tend to dominate the attention distribution, while weak but discriminative high-frequency responses, such as edges, textures, and small objects, are relatively suppressed. In contrast, Sigmoid assigns attention weights independently. Therefore, weak high-frequency responses are not suppressed by dominant responses, which helps preserve fine details in the high-frequency fusion branch.

It is worth noting that the $HW \times HW$ attention matrix in Equation (9) introduces cross-position mixing in the frequency domain, i.e., the coefficients at a given output frequency position may aggregate information from multiple other frequency positions. We believe that such cross-position interaction is not only acceptable but necessary for high-frequency fusion. On the one hand, high-frequency components encode edges, textures, and fine details. In the DCT domain, a single spatial edge or texture pattern is represented by energy distributed across multiple frequency positions. Therefore, cross-position interaction is not only acceptable but necessary for high-frequency fusion, as it allows the model to aggregate related frequency coefficients that collectively represent a coherent spatial structure. On the other hand, the attention is computed between the two modalities. A given high-frequency position in RGB may need to attend to a different position in TIR to find the most relevant cross-modal correspondence, especially when the two modalities exhibit different noise patterns or contrast distributions.

Finally, the fused low- and high-frequency features are added in the frequency domain and reconstructed:

$$F_{fused} = \mathcal{J}_{\text{IDCT}}\left(\tilde{F}_{fuse}^{\text{low}} + \tilde{F}_{fuse}^{\text{high}}\right), \tag{10}$$

where $\mathcal{J}_{\text{IDCT}}$ denotes the 2D Inverse Discrete Cosine Transform (2D-IDCT), which converts the fused frequency-domain representation back to the spatial domain. This "divide-and-fuse" strategy fully leverages the interpretability of the frequency domain.

4. Experiments

4.1. Experimental Setup

During training, we employ widely adopted data augmentation strategies for semantic segmentation on RGB-T image pairs, including resizing, random cropping, and random horizontal flipping. To further improve robustness to illumination variations, we apply random perturbations to the brightness, contrast, saturation, and hue of the RGB images.

We perform end-to-end training using the AdamW optimizer [35] with a polynomial learning rate decay schedule. The batch size is set to 2, the initial learning rate to 1×10^{-4}, and the weight decay to 5×10^{-2}. Since the backbone network is pre-trained on large-scale datasets, we apply a learning rate multiplier of 0.1 to the backbone, which results in an effective learning rate of 1×10^{-5} to avoid disrupting its well-learned general-purpose features. All newly introduced modules retain the base learning rate of 1×10^{-4}.

4.2. Datasets

To validate the effectiveness of the proposed method on RGB-T semantic segmentation, we conduct extensive experiments on three widely used RGB-T semantic segmentation benchmark datasets: MFNet, PST900 and FMB.

MFNet [9] contains 1569 aligned pairs of RGB and thermal images captured in urban environments, annotated with nine semantic classes. The dataset includes 820 daytime image pairs and 749 nighttime image pairs, each with a resolution of 480 × 640. It is split into training, validation, and test subsets following a ratio of 2:1:1.

PST900 [21] consists of 894 aligned RGB-T image pairs labeled across five semantic categories. Compared to MFNet, PST900 provides higher-resolution images at 720 × 1280. The dataset is divided into a training set of 597 image pairs and a test set of 288 image pairs.

FMB [22] contains 1500 aligned visible-infrared image pairs captured in diverse real-world scenes, with 14 semantic categories meticulously annotated. All image pairs have a resolution of 800 × 600, and the dataset is split into a training set of 1220 image pairs and a test set of 280 image pairs.

4.3. Ablation Study

To verify the effectiveness of each component in the proposed method, we conduct systematic ablation studies on the MFNet dataset, using mIoU as the primary evaluation metric.

First, we evaluate the contributions of the two core modules in our overall architecture: the Spatial-Channel Enhancement Module (SCEM) and the Dynamic Frequency Separation Fusion Module (DFSFM). As shown in Table 2, the baseline two-stream network achieves an mIoU of 60.8%. Adding only the SCEM improves performance to 61.5%, demonstrating its positive role in enhancing local details and channel-wise responses for both RGB and thermal features. Incorporating only the DFSFM leads to a significant gain, with mIoU rising to 61.7%, which validates the importance of explicitly modeling frequency-domain information for cross-modal fusion. The full model (SCEM + DFSFM) achieves the best result of 62.9% mIoU, confirming that the two modules work synergistically to improve segmentation accuracy.

Table 2. Ablation study on the effectiveness of SCEM and DFSFM. The **bold** value indicates the best performance.

Model Configuration	SCEM	DFSFM	mIoU (%)
Base			60.8
+SCEM	✓		61.5
+DFSFM		✓	61.7
Full Model (Ours)	✓	✓	**62.9**

Second, we delve into the design rationale behind the differentiated fusion strategy within the Dynamic Frequency Separation Fusion Module (DFSFM). As shown in Table 3, we systematically evaluate various combinations of fusion mechanisms for the high- and low-frequency branches. When both branches employ a lightweight multilayer perception (MLP), the mIoU is 62.3%, indicating that MLP lacks sufficient capacity to model high-frequency texture and edge details. In contrast, applying cross-attention (CA) uniformly to both branches degrades performance to 62.1%. This is because low-frequency components primarily encode global structure and semantic information, which do not require complex global interactions; forcing CA into this branch introduces redundant computation and increases sensitivity to noise, thereby harming fusion stability. More critically, when the fusion strategies are incorrectly assigned, specifically by using CA in the low-frequency branch and MLP in the high-frequency branch (High-MLP / Low-CA), performance drops significantly to 61.0%. This clearly demonstrates that high-frequency details are highly sensitive to the choice of cross-modal alignment mechanism, whereas low-frequency semantics benefit from a simple and robust fusion approach. By comparison, our proposed configuration, which uses MLP in the low-frequency branch and CA in the high-frequency branch (High-CA/Low-MLP), achieves the best result of 62.9% mIoU. This validates our core design principle: the fusion mechanism should align with the intrinsic characteristics of each frequency component. Specifically, introducing cross-attention in the high-frequency branch effectively aligns edge and texture details across RGB and thermal modalities, while employing a lightweight MLP in the low-frequency branch preserves structural consistency and enables efficient, robust semantic fusion. This frequency-aware, differentiated fusion strategy successfully strikes an optimal balance between detail fidelity and semantic stability.

Table 3. Ablation study on low- and high-frequency fusion strategies. "MLP" denotes lightweight multilayer perception; "CA" denotes cross-attention fusion. The **bold** value indicates the best performance.

Method	Low-Freq	High-Freq	mIoU (%)
Same (MLP)	MLP	MLP	62.3
Same (CA)	CA	CA	62.1
High-MLP/Low-CA	CA	MLP	61.0
High-CA/Low-MLP	MLP	CA	**62.9**

Third, to validate the design choice of the activation function in the cross-attention module, we conduct ablation experiments comparing the adoption of Softmax, normalized Sigmoid and Sigmoid in Equation (9). As shown in Table 4, normalized Sigmoid degrades performance to 61.9%, even slightly below Softmax (62.1%), because the normalization again suppresses weak high-frequency responses in a manner similar to Softmax. In contrast, our Sigmoid with layer-wise scale calibration via layer normalization achieves the best result of 62.9%, preserving high-frequency details while maintaining stable feature scales across layers.

Table 4. Ablation on the activation function for the cross-attention-based high-frequency fusion module. The **bold** value indicates the best performance.

	Softmax	Normalized Sigmoid	Sigmoid
mIoU (%)	62.1	61.9	**62.9**

Fourth, we conduct additional spectral analysis on the features of stage 2 and 3 extracted by the ConvNeXt encoder on the MFNet dataset. As shown in Table 5, the overall spectral trend is preserved across all stages. At both Stage 2 and Stage 4, the energy distribution follows the same pattern as the image level: the DC component dominates, followed by a sharp drop at low frequencies and a gradual decay at higher frequencies. The above observations confirm that the separation of low-frequency and high-frequency remains structurally meaningful even in deep feature representations.

Table 5. Average radial DCT energy distributions versus normalized frequency radius on the MFNet training set. The table reports the channel-averaged radial DCT energy for the raw RGB and TIR images and for the ConvNeXt encoder features at Stage 2 and Stage 4. The energy magnitude is represented on a logarithmic scale, and the normalized frequency radius ranges from 0 (DC component) to 1 (highest frequency). At all levels, the RGB and TIR modalities exhibit similar energy in the ultra-low-frequency region, whereas their energy distributions diverge substantially in the mid- to high-frequency region.

Source	0	0.2	0.4	0.6	0.8	1.0
RGB image	1.6×10^{2}	1.0×10^{-2}	5.9×10^{-4}	7.5×10^{-5}	3.4×10^{-5}	4.1×10^{-6}
TIR image	2.8×10^{2}	1.3×10^{-3}	1.3×10^{-4}	2.1×10^{-5}	8.6×10^{-6}	2.5×10^{-6}
RGB feature (Stage 2)	2.4×10^{3}	3.5×10^{-1}	1.2×10^{-2}	2.8×10^{-3}	9.4×10^{-4}	1.6×10^{-4}
TIR feature (Stage 2)	4.1×10^{3}	2.7×10^{-2}	3.6×10^{-3}	5.1×10^{-4}	1.5×10^{-4}	5.2×10^{-5}
RGB feature (Stage 4)	6.8×10^{3}	2.1×10^{-1}	5.3×10^{-3}	6.7×10^{-3}	2.1×10^{-3}	3.2×10^{-5}
TIR feature (Stage 4)	9.4×10^{3}	1.2×10^{-2}	1.8×10^{-3}	1.4×10^{-4}	4.8×10^{-4}	1.5×10^{-6}

Fifth, we investigate the design of the frequency descriptor used for threshold prediction. In Equation (5), the GAP operation is applied over the spatial dimensions ($H \times W$) of the DCT coefficient map $\tilde{F} \in \mathbb{R}^{H \times W \times C}$, producing a channel-wise descriptor $\mathbf{z} \in \mathbb{R}^{C}$. This is essentially equivalent to a frequency-binned descriptor with a single bin. we further compare against two alternative descriptors with 4 and 8 frequency bins, respectively. As depicted in Table 6, the 4-bin variant achieves 0.2% improvement over GAP, which is within the range of training variance and not statistically significant. However, it introduces more parameters in the predictor MLP due to the increased dimension of the input, while yielding negligible gains. Therefore, the above results confirm that the channel-wise global descriptor already captures sufficient information for threshold prediction, and explicit frequency binning does not provide meaningful additional benefit.

Table 6. Ablation on the number of frequency-binned descriptors used for threshold prediction in Equation (5).

	1 Bin	4 Bins	8 Bins
mIoU (%)	62.9	63.1	62.8

Sixth, we provide a branch-wise ablation to quantitatively demonstrate the effectiveness and complementarity of the heterogeneous low- and high-frequency fusion. Specifically, we further train two degraded models with solely low-frequency branch and high-frequency branch, respectively. Then, we compare their segmentation performance with full model that adopts both branches for heterogeneous fusion. As exhibited in Table 7, removing the low-frequency branch reduces mIoU from 62.9% to 60.7%, indicating that the low-

frequency branch plays a major role in preserving global structure and cross-modal semantic consistency. Removing the high-frequency branch reduces mIoU to 61.4%, indicating that the high-frequency branch is also essential for preserving edges, textures, and fine details. More importantly, the full model outperforms both single-branch variants, which confirms that the two branches are complementary rather than redundant.

Table 7. Ablation on the contribution of each frequency branch on MFNet.

Low-Freq Branch	High-Freq Branch	mIoU (%)
×	✓	60.7
✓	×	61.4
✓	✓	62.9

Finally, we evaluate the impact of different threshold design strategies for low- and high-frequency separation on segmentation performance. As shown in Table 8, we examine fixed thresholds ranging from 0.3 to 0.7. Among them, $\tau = 0.5$ achieves the best mIoU of 61.9%. When τ is too small (e.g., 0.3), too few coefficients are retained in the low-frequency branch, leading to insufficient fusion of global structure and semantic information, while more mid- and high-frequency components are pushed into the high-frequency branch, increasing sensitivity to modality-specific noise. When τ is too large (e.g., 0.7), excessive mid- and high-frequency coefficients are included in the low-frequency branch, where the lightweight MLP fusion is not designed to align fine details. Replacing the fixed threshold with learnable parameters improves performance to 62.5%, demonstrating that end-to-end optimization of frequency band boundaries provides measurable gains. Building upon this, our lightweight MLP-based approach that dynamically predicts thresholds conditioned on input content further increases mIoU to 62.9%. This consistent improvement indicates that content-aware adaptive frequency partitioning is more effective than either static thresholds or unconstrained learnable parameters in balancing the complementary low-frequency structural cues and high-frequency detail responses between RGB and thermal infrared modalities.

Table 8. Ablation on threshold design for low- and high-frequency separation. The **bold** value indicates the best performance.

	$\tau = 0.3$	$\tau = 0.4$	$\tau = 0.5$	$\tau = 0.6$	$\tau = 0.7$	Learnable τ	MLP τ
mIoU (%)	61.2	61.6	61.9	61.4	61.0	62.5	**62.9**

4.4. Efficiency Analysis

Our method offers two variants, Ours(S) and Ours(B), which use ConvNeXtV2-B and ConvNeXtV2-L as their network backbone, respectively. To substantiate the efficiency of our approach, we report FLOPs, parameter count, model size, and inference speed (FPS) for all compared methods at the matched input resolution of 480 × 640 on the MFNet dataset. As shown in Table 9, although the DCT/IDCT operations bring extra computation, the overall efficiency of our model remains competitive. The lightweight variant Ours(S) achieves 61.2% mIoU with 155.2 GFLOPs, 151.8 M parameters, 378.4 MB model size, and 18.3 FPS. Compared with CMX, Ours(S) obtains 2.0 higher mIoU and higher FPS (18.3 vs. 11.7), while using only moderately more FLOPs and parameters. Compared with MiLNet, Ours(S) achieves 1.1% higher mIoU with comparable model size, although its FPS is lower. The larger variant Ours(B) achieves the best mIoU of 62.9% among all compared methods, with 272.6 GFLOPs and 12.5 FPS. Its FLOPs are lower than RTFNet (337.0 G), and its model size is smaller than RTFNet (766.2 MB vs. 1024.0 MB), while its mIoU is substantially higher.

These results indicate that the additional frequency transform cost is acceptable within the overall accuracy efficiency trade-off.

Table 9. Comparison with advanced methods on the MFNet dataset from both segmentation accuracy and efficiency perspectives.

Method	FLOPs (G)	Param (M)	Size (MB)	FPS	mIoU (%)
CMX [23]	134.3	139.9	533.6	11.7	59.2
EGFNet [36]	201.3	62.8	239.6	8.6	54.8
LASNet [14]	233.8	93.6	357.0	24.9	54.9
RTFNet [10]	337.0	254.5	10240	26.2	53.2
MiLNet [37]	136.3	92.3	352.1	24.6	60.1
Ours(S)	155.2	151.8	378.4	18.3	61.2
Ours(B)	272.6	253.7	766.2	12.5	62.9

4.5. Comparison with Advanced Methods

We comprehensively evaluate the proposed method on three representative RGB-T semantic segmentation benchmarks, covering indoor search-and-rescue (PST900), road understanding for autonomous driving (MFNet), and high-precision urban scene parsing (FMB). As shown in Tables 10–12, our method achieves new state-of-the-art performance across all datasets, fully validating the universality and robustness of the proposed dynamic frequency-band decoupling and frequency-specific fusion mechanism. Notably, Open-RGBT, RoadFormer+ and KTB only report the overall average performance (mIoU and mAcc) in their original publications and do not provide per-category results. Therefore, we do not include their per-category comparisons.

Table 10. Comparisons with Advanced Methods on the PST900 Dataset. "-" indicates that the corresponding metric is not reported in the original paper. The **bold** value indicates the best performance.

Methods	Venue	Class-Wise Accuracy (Acc, %)					
		Background	Hand-Drill	Backpack	Fire-Extinguisher	Survivor	mAcc
MDRNet+ [15]	TNNLS'2024	99.30	77.10	79.10	77.90	69.80	81.30
CAINet [38]	TMM'2024	99.66	**95.87**	**96.09**	88.38	**91.35**	94.27
MiLNet [37]	TIP'2025	99.79	95.37	93.52	94.92	81.32	92.99
Open-RGBT [39]	ICRA'2025	-	-	-	-	-	93.75
IQSeg-CL [40]	PR'2026	**99.90**	93.87	82.83	92.40	73.80	88.55
SCRNet [41]	PR'2025	99.68	83.96	89.61	88.63	85.04	89.38
LLESeg [42]	TIM'2025	99.83	85.25	84.62	85.74	69.54	85.00
Ours	–	99.87	94.19	94.40	**97.24**	89.20	**94.98**
		Class-Wise IoU (mIoU, %)					
		Background	**Hand-Drill**	**Backpack**	**Fire-Extinguisher**	**Survivor**	**mIoU**
MDRNet+ [15]	TNNLS'2024	98.90	40.60	71.10	52.40	62.30	68.70
CAINet [38]	TMM'2024	99.50	80.30	88.02	77.21	78.69	84.47
MiLNet [37]	TIP'2025	99.54	76.25	89.64	82.15	77.97	85.11
Open-RGBT [39]	ICRA'2025	-	-	-	-	-	89.76
IQSeg-CL [40]	PR'2026	99.41	78.92	81.29	88.38	69.74	83.55
SCRNet [41]	PR'2025	99.37	70.71	82.37	69.64	75.74	79.57
KTB [43]	CVPR'2025	-	-	-	-	-	88.70
LLESeg [42]	TIM'2025	99.31	75.60	79.00	82.00	67.61	80.70
Ours	–	**99.69**	**83.58**	**91.92**	**89.82**	**85.63**	**90.13**

Table 11. Comparisons with Advanced Methods on the MFNet Dataset. "-" indicates that the corresponding metric is not reported in the original paper. The **bold** value indicates the best performance.

Methods	Venue	Car	Person	Bike	Curve	Car_stop	Guardrail	Color_cone	Bump	mIoU
MDRNet+ [15]	TNNLS'2024	87.1	69.8	60.9	47.8	34.2	8.2	50.2	55.0	56.8
RoadFormer+ [44]	TIV'2025	-	-	-	-	-	-	-	-	62.7
CAINet [38]	TMM'2024	88.5	66.3	**68.7**	**55.4**	31.5	9.0	48.9	60.7	58.6
MiLNet [37]	TIP'2025	87.9	72.3	64.5	46.2	45.8	10.1	55.8	60.1	60.1
Open-RGBT [39]	ICRA'2025	-	-	-	-	-	-	-	-	61.6
SCRNet [41]	PR'2025	87.9	74.1	62.7	47.4	36.2	7.5	53.2	54.4	58.0
KTB [43]	CVPR'2025	-	-	-	-	-	-	-	-	59.9
LLESeg [42]	TIM'2025	88.6	73.2	64.8	46.8	30.0	8.8	52.5	62.4	58.4
Ours	-	**91.3**	**74.7**	68.2	52.2	**47.6**	**11.8**	**57.2**	**65.3**	**62.9**

Table 12. Comparisons with Advanced Methods on the FMB Dataset. * denotes reproduced using official code. "-" indicates that the corresponding metric is not reported in the original paper. The **bold** value indicates the best performance.

Methods	Venue	Class-Wise Accuracy (Acc, %)								
		Car	Person	Truck	T-Lamp	T-Sign	Building	Vegetation	Pole	Mean
EGFNet [36]	AAAI'2022	83.6	79.5	33.5	58.6	82.6	88.5	89.3	63.8	63.0
SegMiF [22]	ICCV'2023	85.3	78.3	74.4	**86.4**	86.1	90.0	91.6	**72.5**	74.5
Roadformer+ * [44]	TIV'2024	**94.5**	**83.3**	47.2	55.0	88.1	**93.7**	94.4	68.6	77.8
IQSeg-CL [40]	PR'2026	92.7	71.5	38.6	-	85.1	91.1	92.7	59.4	67.6
AGFNet [45]	TITS'2025	95.3	80.0	46.7	56.8	85.2	91.6	91.4	65.5	70.6
MiLNet * [37]	TIP'2025	94.2	81.2	55.8	56.6	**88.9**	90.9	93.8	70.2	76.9
Ours	–	94.7	78.2	68.0	48.5	88.6	92.2	**94.3**	69.3	**80.0**
		Class-Wise IoU (IoU, %)								
		Car	Person	Truck	T-Lamp	T-Sign	Building	Vegetation	Pole	Mean
EGFNet [36]	AAAI'2022	77.4	63.0	17.1	25.2	66.6	77.2	83.5	41.5	47.3
SegMiF [22]	ICCV'2023	78.3	65.4	47.3	43.1	74.8	82.0	85.0	49.8	54.8
Roadformer+ * [44]	TIV'2024	84.6	76.1	41.3	49.9	83.9	**86.9**	**89.6**	57.0	71.6
IQSeg-CL [40]	PR'2026	81.5	60.6	36.3	-	75.5	81.6	86.0	49.3	60.5
AGFNet [45]	TITS'2025	78.9	62.1	36.9	46.3	72.3	81.0	85.4	**63.0**	60.5
MiLNet * [37]	TIP'2025	83.5	64.2	65.2	**51.0**	82.9	84.3	86.7	58.9	70.6
Ours	–	**86.8**	**72.5**	**57.9**	44.4	**83.5**	85.5	88.2	56.0	**73.3**

On the PST900 indoor search-and-rescue dataset, our method sets a new state-of-the-art with 90.13% mIoU, achieving the highest IoU on all semantic classes. Notably, it significantly outperforms existing approaches on safety-critical small objects such as Survivor (85.63%), demonstrating high sensitivity to life-critical targets.

On the more challenging MFNet autonomous driving dataset, our approach attains an mIoU of 62.9% (mean ± std over five checkpoints sampled from the converged training phase: 62.9 ± 0.14%), consistently surpassing the previous best method RoadFormer+ (62.7%). This comparison is conducted under fair conditions: both Ours(B) and RoadFormer+ use the ConvNeXt-L pretrained backbone and fine-tune the backbone with a learning rate of 0.1. Moreover, our method achieves top performance in 6 out of 8 classes, including easily missed small objects such as car_stop, guardrail, and bump, confirming its strong capability in preserving fine details under low-light and cluttered background conditions.

On the high-annotation-density FMB urban dataset, we report per-class results on these 8 classes, while the mean metrics are computed over all 14 classes. Our method again achieves the best overall performance with 80.0% mean accuracy and 73.3% mean IoU. It leads notably in structured categories such as Car, Truck, and T-Sign, as well as in small or ambiguous classes like Person, showcasing its strong modeling capacity for complex semantic layouts and cross-modal consistency.

In summary, whether in safety-critical scenarios (PST900), adverse lighting conditions (MFNet), or highly complex urban environments (FMB), our method consistently delivers superior segmentation performance by explicitly decoupling high-frequency components such as edges and textures from low-frequency components such as global structure and semantics, and applying robust fusion to low frequencies and refined fusion to high frequencies. This frequency-aware heterogeneous fusion paradigm enables efficient and adaptive integration of complementary multi-modal information across diverse real-world conditions.

4.6. Qualitative Comparison

To qualitatively assess the segmentation quality, we compare our method with existing RGB-T approaches. As shown in Figure 5, our predictions consistently exhibit cleaner object boundaries, fewer false positives, and better preservation of semantic coherence across diverse scenes. These observations align with our quantitative results, confirming that our approach produces more accurate and reliable segmentation outputs overall.

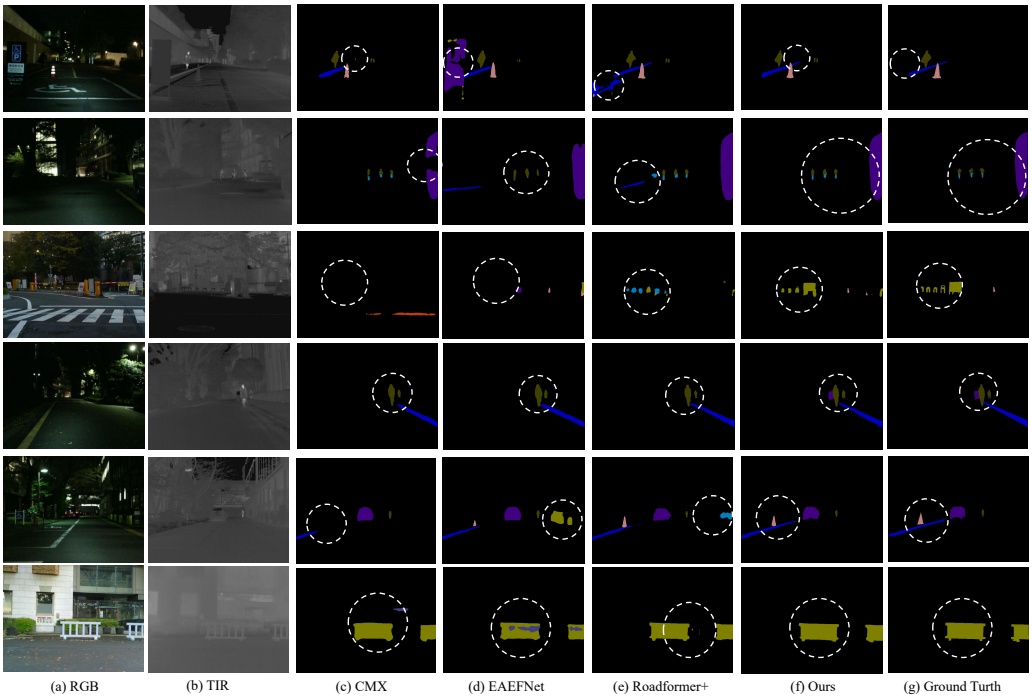


Figure 5. Visualisations on MFNet. From left to right: (**a**) RGB input, (**b**) TIR input, (**c**) prediction from CMX, (**d**) prediction from EAEFNet, (**e**) prediction from Roadformer+, (**f**) prediction from our method, and (**g**) ground truth.

5. Conclusions

In this work, a novel frequency-aware fusion framework is proposed for RGB-T semantic segmentation, with its superiority demonstrated quantitatively and qualitatively. Specifically, dynamic frequency separation and heterogeneous fusion pathways are introduced, highlighting the importance of adaptively separating high-frequency details from low-frequency semantics based on input content and treating them differently, emphasizing cross-modal detail alignment in the high-frequency stream and semantic consistency in the low-frequency stream. Extensive experiments on three standard benchmark datasets validate the effectiveness of the proposed design, which holds promise for advancing RGB-T segmentation toward more principled and frequency-aware model architectures.

Author Contributions: Conceptualization, S.Z., H.Y. and Y.Z.; methodology, S.Z., X.W. and H.Y.; software, S.Z., X.W., H.Y. and J.J.; validation, S.Z. and B.L.; formal analysis, X.W., H.Y., S.Z. and B.L.; investigation, S.Z., B.L. and Y.Z.; resources, B.L. and Y.Z.; data curation, S.Z., J.J., X.W. and H.Y.; writing—original draft preparation, S.Z., X.W. and H.Y.; writing—review and editing, B.L., J.J. and Y.Z.; visualization, X.W., H.Y., S.Z. and B.L.; supervision, B.L. and Y.Z.; project administration, B.L. and S.Z.; funding acquisition, S.Z. and H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62602669), the Fundamental Research Funds for the Central Universities (2025QN1162), the "Taihu Light" Science and Technology Project of Wuxi (K20231010).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, Q.; Zhao, S.; Luo, Y.; Zhang, D.; Huang, N.; Han, J. 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 (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 2633–2642.
2. Liang, M.; Hu, J.; Bao, C.; Feng, H.; Deng, F.; Lam, T.L. Explicit attention-enhanced fusion for RGB-thermal perception tasks. *IEEE Robot. Autom. Lett.* **2023**, *8*, 4060–4067. [CrossRef]
3. Li, X.; Yu, L.; Chen, H.; Fu, C.W.; Xing, L.; Heng, P.A. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 523–534. [CrossRef]
4. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
6. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.
7. Tang, Z.; Xu, T.; Feng, Z.; Zhu, X.; Wang, H.; Shao, P.; Cheng, C.; Wu, X.J.; Awais, M.; Atito, S.; et al. Revisiting RGBT Tracking Benchmarks from the Perspective of Modality Validity: A New Benchmark, Problem, and Method. *arXiv* **2024**, arXiv:2405.00168.
8. Tang, Z.; Xu, T.; Wu, X.J.; Kittler, J. Multi-Level Fusion for Robust RGBT Tracking via Enhanced Thermal Representation. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 1–24. [CrossRef]
9. Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; Harada, T. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: Piscataway, NJ, USA, 2017; pp. 5108–5115.
10. Sun, Y.; Zuo, W.; Liu, M. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robot. Autom. Lett.* **2019**, *4*, 2576–2583. [CrossRef]
11. Wang, M.; Zhu, Z.; Wang, Y.; Tu, R.; Weng, J.; Yu, X. Edge-Supervised Attention-Aware Fusion Network for RGB-T Semantic Segmentation. *Electronics* **2025**, *14*, 1489.
12. Hwang, J.Y.; Lee, Y.J.; Jung, H.G.; Suhr, J.K. Pruning-Friendly RGB-T Semantic Segmentation for Real-Time Processing on Edge Devices. *Electronics* **2025**, *14*, 3408.
13. Kafka, O.; Rankl, C.; Moser, D. Beyond RGB: Early Stage Fusion of Thermal and Visual Modalities for Robust Maritime Perception. *Electronics* **2025**, *14*, 4746. [CrossRef]
14. Li, G.; Wang, Y.; Liu, Z.; Zhang, X.; Zeng, D. RGB-T semantic segmentation with location, activation, and sharpening. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 1223–1235. [CrossRef]
15. Zhao, S.; Liu, Y.; Jiao, Q.; Zhang, Q.; Han, J. Mitigating Modality Discrepancies for RGB-T Semantic Segmentation. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 9380–9394. [CrossRef]
16. Zhao, S.; Wang, J.; Zhang, Q.; Han, J. Towards efficient RGB-T semantic segmentation via feature generative distillation strategy. *Inf. Fusion* **2025**, *123*, 103282. [CrossRef]
17. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FcaNet: Frequency Channel Attention Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 783–792.

18. Ma, R.; Zhang, Y.; Zhang, B.; Fang, L.; Huang, D.; Qi, L. Learning Attention in the Frequency Domain for Flexible Real Photograph Denoising. *IEEE Trans. Image Process.* **2024**, *33*, 3707–3721. [CrossRef]
19. Su, B.; Liu, J.; Su, X.; Luo, B.; Wang, Q. CFCANet: A Complete Frequency Channel Attention Network for SAR Image Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11750–11763. [CrossRef]
20. Zhang, F.; Panahi, A.; Gao, G. FsaNet: Frequency Self-Attention for Semantic Segmentation. *IEEE Trans. Image Process.* **2023**, *32*, 4757–4772. [CrossRef]
21. Shivakumar, S.S.; Rodrigues, N.; Zhou, A.; Miller, I.D.; Kumar, V.; Taylor, C.J. PST900: RGB-thermal calibration, dataset and segmentation network. In *Proceedings of the 2020 IEEE international conference on robotics and automation (ICRA)*; IEEE: Piscataway, NJ, USA, 2020; pp. 9441–9447.
22. Liu, J.; Liu, Z.; Wu, G.; Ma, L.; Liu, R.; Zhong, W.; Luo, Z.; Fan, X. Multi-interactive Feature Learning and a Full-time Multi-modality Benchmark for Image Fusion and Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 8115–8124.
23. Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; Stiefelhagen, R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 14679–14694. [CrossRef]
24. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2223–2232.
25. Acharya, T.; Ray, A.K. *Image Processing: Principles and Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2005.
26. Lee, J.H.; Heo, M.; Kim, K.R.; Kim, C.S. Single-Image Depth Estimation Based on Fourier Domain Analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 330–339.
27. Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; Heng, P.A. FedDG: Federated Domain Generalization on Medical Image Segmentation via Episodic Learning in Continuous Frequency Space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1013–1023.
28. Xu, Z.Q.J.; Zhang, Y.; Luo, T. Overview frequency principle/spectral bias in deep learning. *Commun. Appl. Math. Comput.* **2025**, *7*, 827–864. [CrossRef]
29. Gueguen, L.; Sergeev, A.; Kadlec, B.; Liu, R.; Yosinski, J. Faster Neural Networks Straight from JPEG. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Volume 31.
30. Li, X.; Xu, F.; Yu, A.; Lyu, X.; Gao, H.; Zhou, J. A Frequency Decoupling Network for Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–21. [CrossRef]
31. Azad, R.; Kazerouni, A.; Sulaiman, A.; Bozorgpour, A.; Aghdam, E.K.; Jose, A.; Merhof, D. Unlocking Fine-Grained Details with Wavelet-Based High-Frequency Enhancement in Transformers. In Proceedings of the Machine Learning in Medical Imaging, Marrakesh, Morocco, 6–10 October 2024; pp. 207–216.
32. Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.K.; Ren, F. Learning in the Frequency Domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020.
33. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 16133–16142.
34. Jiang, X.; Li, Y.; Jiang, T.; Xie, J.; Wu, Y.; Cai, Q.; Jiang, J.; Xu, J.; Zhang, H. RoadFormer: Pyramidal deformable vision transformers for road network extraction with remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102987. [CrossRef]
35. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
36. Zhou, W.; Dong, S.; Xu, C.; Qian, Y. Edge-aware guidance fusion network for rgb-thermal scene parsing. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 3571–3579.
37. Liu, J.; Liu, H.; Li, X.; Ren, J.; Xu, X. MiLNet: Multiplex Interactive Learning Network for RGB-T Semantic Segmentation. *IEEE Trans. Image Process.* **2025**, *34*, 1686–1699. [CrossRef]
38. Lv, Y.; Liu, Z.; Li, G. Context-Aware Interaction Network for RGB-T Semantic Segmentation. *IEEE Trans. Multimed.* **2024**, *26*, 6348–6360. [CrossRef]
39. Yu, M.; Yue, Y.; Yang, L.; He, X.; Yang, Y.; Fu, M. Open-RGBT: Open-Vocabulary RGB-T Zero-Shot Semantic Segmentation in Open-World Environments. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA), Atlanta, GA, USA, 19–23 May 2025; pp. 1–7.
40. Liu, C.; Liu, H.; Zhuo, J.; Zou, B.; Chen, J.; Zhao, Q.; Ma, H. Implicit alignment and query refinement for RGB-T semantic segmentation. *Pattern Recognit.* **2026**, *169*, 111951. [CrossRef]
41. Zhao, S.; Jin, Z.; Jiao, Q.; Zhang, Q.; Han, J. Resolving semantic conflicts in RGB-T semantic segmentation. *Pattern Recognit.* **2025**, *162*, 111398. [CrossRef]

42. Guo, X.; Liu, Y.; Xue, W.; Zhang, Z.; Zhuang, Y. Low-Light Enhancement and Global-Local Feature Interaction for RGB-T Semantic Segmentation. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 1–13. [CrossRef]
43. Cai, J.; Su, J.; Li, Q.; Yang, W.; Wang, S.; Zhao, T.; He, S.; Liu, W. Keep the Balance: A Parameter-Efficient Symmetrical Framework for RGB+X Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 10–17 June 2025; pp. 10587–10598.
44. Huang, J.; Li, J.; Jia, N.; Sun, Y.; Liu, C.; Chen, Q.; Fan, R. RoadFormer+: Delivering RGB-X Scene Parsing Through Scale-Aware Information Decoupling and Advanced Heterogeneous Feature Fusion. *IEEE Trans. Intell. Veh.* **2025**, *10*, 3156–3165. [CrossRef]
45. Zhou, X.; Wu, X.; Bao, L.; Yin, H.; Jiang, Q.; Zhang, J. AGFNet: Adaptive Gated Fusion Network for RGB-T Semantic Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 6477–6492. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.