# AMDANet: Attention-Driven Multi-Perspective Discrepancy Alignment for RGB-Infrared Image Fusion and Segmentation

Haifeng Zhong[1,2], Fan Tang[3], Zhuo Chen[4], Hyung Jin Chang[4], Yixing Gao[1,2,*]
[1] School of Artificial Intelligence, Jilin University,
[2] Engineering Research Center of Knowledge-Driven Human-Machine Intelligence,
Ministry of Education, China, [3] Institute of Computing Technology, Chinese Academy of Sciences,
[4] School of Computer Science, University of Birmingham
zhonghf23@mails.jlu.edu.cn, gaoyixing@jlu.edu.cn, tangfan@ict.ac.cn,
zxc417@student.bham.ac.uk, H.J.Chang@bham.ac.uk

## Abstract

*The challenge of multimodal semantic segmentation lies in establishing semantically consistent and segmentable multimodal fusion features under conditions of significant visual feature discrepancies. Existing methods commonly construct cross-modal self-attention fusion frameworks or introduce additional multimodal fusion loss functions to establish fusion features. However, these approaches often overlook the challenge caused by feature discrepancies between modalities during the fusion process. To achieve precise segmentation, we propose an **A**ttention-Driven **M**ultimodal **D**iscrepancy **A**lignment **Net**work (**AMDANet**). **AMDANet** reallocates weights to reduce the saliency of discrepant features and utilizes low-weight features as cues to mitigate discrepancies between modalities, thereby achieving multimodal feature alignment. Furthermore, to simplify the feature alignment process, a semantic consistency inference mechanism is introduced to reveal the network's inherent bias toward specific modalities, thereby compressing cross-modal feature discrepancies from the foundational level. Extensive experiments on the FMB, MFNet, and PST900 datasets demonstrate that **AMDANet** achieves mIoU improvements of 3.6%, 3.0%, and 1.6%, respectively, significantly outperforming state-of-the-art methods. The code is available at https://github.com/Zhonghaifeng6/AMDANet*

## 1. Introduction

Current semantic segmentation methods [5, 13, 15, 55] mostly depend on visible light sensors for scene understanding. In certain challenging environments, such as nighttime,

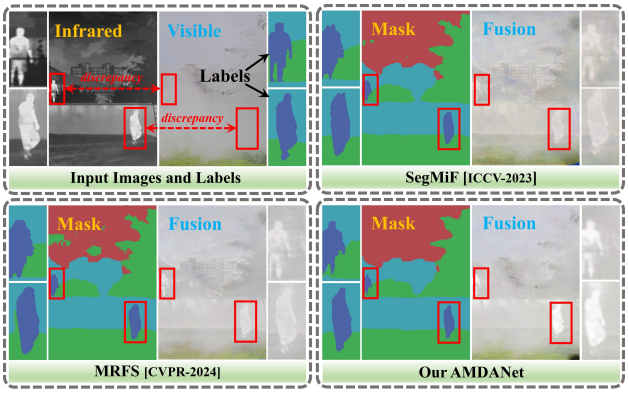


Figure 1. Comparison with advanced SegMiF [60] and MRFS [25] in image fusion and semantic segmentation is presented. As shown in the figure, when there is a significant difference between the infrared and visible images, it is difficult for other methods to align cross-modal features, leading to suboptimal results. In contrast, our method not only compresses the difference features between cross-modality by weight adjustment but also achieves semantic consistency alignment, thus achieving better results.

incomplete scene descriptions can hinder the accuracy of semantic understanding. Some studies [25, 38, 41, 48, 49, 60, 68] have leveraged the unique thermal imaging capabilities of infrared sensors, incorporating infrared images as a complementary modality in segmentation to enable more effective perception for complex environments.

Existing methods [8, 25, 51, 60, 68] predominantly focus on leveraging semantic consistency among multimodal features to construct segmentable feature representations, with the key challenge lying in identifying and integrating semantically consistent features across modalities. Current research addresses this from two perspectives: designing multimodal fusion loss functions [45, 65] to jointly optimize fusion and segmentation tasks, or employing cross-modal self-attention mechanisms [25, 60, 68] to align semantic

*Corresponding author.

features. However, these works primarily emphasize aggregating semantically consistent multimodal features while neglecting the detrimental impact of inter-modal discrepant features on the fusion process. As illustrated in Fig. 1, feature ambiguities arising from visual disparities—such as divergent contours, shapes, and textures across modalities—severely hinder the establishment of consistent semantic representations. For fusion loss-driven methods [45, 65], minimizing such discrepancies often results in biased dominance of features from one modality, whereas cross-modal attention mechanisms [25, 60, 68] risk discarding critical details from either modality during feature matching. Thus, the critical step for multimodal semantic segmentation lies in compressing and aligning inter-modal feature discrepancies during fusion to accurately construct semantically consistent and segmentable features.

To systematically address the obstacles posed by inter-modal feature discrepancies in building unified fusion features, we propose an **A**ttention-Driven **M**ultimodal **D**iscrepancy **A**lignment Network (AMDANet). We categorize the discrepancies into two types: **visual discrepancies** caused by modality-specific appearances, and **empirical feature biases** introduced by the encoder's inherent preferences for specific modalities due to factors like regularization and nonlinear activation [35, 43, 47]. For **visual discrepancies**, we design a feature discrepancy alignment module (FDAM). FDAM adopts a divide-and-conquer strategy, leveraging self-enhanced single-modality features as cues to eliminate mismatched information from both local and global perspectives, thereby strengthening the coupling between multimodal features. To address **empirical feature biases**, we propose a semantic consistency inference (SCI) mechanism. SCI inhibits encoder-induced modality preferences at their source by leveraging the semantic similarity between features of different modalities, thereby preventing the progressive accumulation of discrepant features across modalities in deeper network layers. Finally, to enhance multimodal fusion, we introduce a mutual feature mask learning (MFML) strategy. MFML employs pixel-level feature masking to promote the learning of complementary and meaningful cross-modal representations, ensuring robustness against partial modality degradation.

We conduct extensive experiments and ablations on FMB [25], MFNet [10], and PST900 [38] datasets, to measure the performance of our method. The experimental results demonstrate that our method achieves superior accuracy compared to state-of-the-art (SOTA) methods.

Our contributions can be summarized as follows:

- We propose a multimodal semantic segmentation network, AMDANet, which aligns feature discrepancies across modalities from a multi-perspective to establish easily segmentable feature representations.
- To align discrepant features that affect cross-modal feature fusion, we propose the feature discrepancy alignment module (FDAM) and the semantic consistency inference (SCI). FDAM and SCI adopt a divide-and-conquer approach to align cross-modal features from local and global perspectives, respectively.
- We evaluate our method on the FMB, MFNet, and PST900 datasets. The results demonstrate the superiority of our method. Comprehensive ablation studies validate the effectiveness of the proposed modules and strategies.

## 2. Related work

**Multi-modal image fusion.** With the development of deep learning [3, 28, 66], multimodal image fusion methods [14, 33, 57, 65, 71] have made significant progress. Existing methods can be broadly categorized into four types: based on generative adversarial network [29, 30], based on autoencoder [18, 20, 26], based on unified model [16, 53, 54, 59], and based on algorithm unrolling model [4, 7, 64]. These methods focus on improving the fusion effect of multimodal images by designing novel networks or loss functions, but often neglect whether the fusion results are well-suited to downstream tasks. Recently, some studies [21, 24, 51] have considered how to integrate the fusion process with downstream recognition tasks. For instance, Liu et al. [24] combine image fusion and detection tasks by designing a dual adversarial learning fusion network. Similarly, Wu et al. [51] propose a fusion framework that generates fusion features beneficial to downstream tasks by minimizing cross-modal semantics. However, existing works lack consideration of how to establish a unified feature paradigm from multimodal images based on task attributes. In contrast, our method focuses on fully eliminating discrepant features between multimodal images to establish fusion features conducive to downstream tasks.

**Multimodal semantic segmentation.** Multimodal semantic segmentation [1, 25, 60, 68] involves constructing an end-to-end network to simultaneously achieve image fusion and semantic segmentation. Existing methods [8, 25, 44, 60, 68, 69] primarily implement semantic segmentation at the feature level using strategies such as weighted averaging [61], summation [40, 67], and concatenation [10, 38]. For example, Shivakumar et al. [38] propose a dual-stream architecture that maps feature streams from different modalities. Zhao et al. [68] achieve complementary and fused features across multiple levels of different modalities by weighting the extracted multi-scale and multi-level features. Although current methods have made great strides in accuracy, they ignore the barriers of discrepant features to establishing consistent features. In contrast to existing methods, our method emphasizes resolving the obstacles that semantic discrepancies pose to the establishment of consistent features from multiple perspectives, thereby facilitating the alignment of multimodal features.

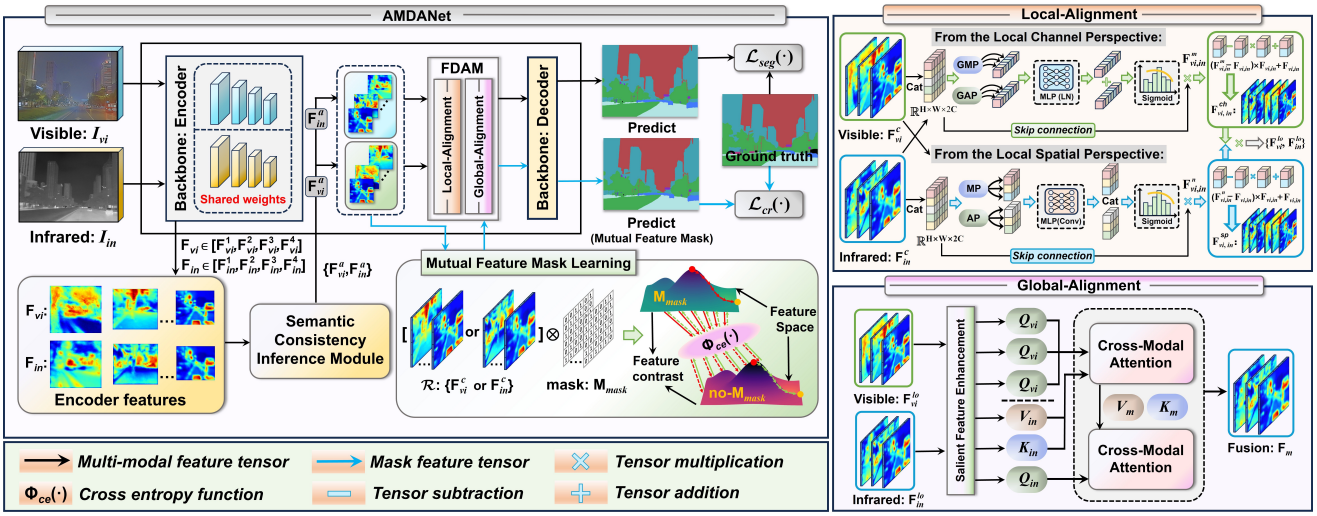


Figure 2. Overview of our AMDANet. The left part describes the workflow of AMDANet and mutual feature mask learning. The right part provides a detailed description of the FDAM, which suppresses the discrepant features from both local and global perspectives.

## 3. Method

**Preliminaries:** Our goal is to eliminate the discrepancies between different modalities that hinder the establishment of semantic consistency features and couple effective multimodal features into a unified framework. Given the multimodal inputs be a pair of visible light and infrared images, denoted as $I_{vi} \in \langle I^1_{vi}, I^2_{vi}, \ldots, I^n_{vi} \rangle$ and $I_{in} \in \langle I^1_{in}, I^2_{in}, \ldots, I^n_{in} \rangle$. Our method employs an encoder [52] to perform feature extraction across four levels, generating the foundational visible features $\mathbf{F}_{vi} \in \langle \mathbf{F}^1_{vi}, \mathbf{F}^2_{vi}, \mathbf{F}^3_{vi}, \mathbf{F}^4_{vi} \rangle$ and infrared features $\mathbf{F}_{in} \in \langle \mathbf{F}^1_{in}, \mathbf{F}^2_{in}, \mathbf{F}^3_{in}, \mathbf{F}^4_{in} \rangle$ required to construct consistent representations.

**Method overview:** The overview of our AMDANet is outlined in Fig. 2. The AMDANet consists of three key components: the semantic consistency inference (SCI) (Sec. 3.1), the feature discrepancy alignment (FDAM) (Sec. 3.2), and the mutual feature masking learning (MFML) (Sec. 3.3). First, to address inherent feature biases generated by the encoder, we leverage the SCI to evaluate the network's biased performance across different modalities. Based on the evaluation results, biases are excluded from $\mathbf{F}_{vi}$ and $\mathbf{F}_{in}$, hereby easing the alignment of multimodal features. Second, to achieve semantic alignment between multimodal features, we employ the FDAM to remove invalid features that are prone to misjudgment from both local channels and global spatial dimensions. Finally, we utilize the MFML to achieve fusion of multimodal features by randomly applying mask perturbations [37] to specific modality features.

### 3.1. Semantic Consistency Inference

Under the influence of nonlinear factors such as regularization and activation functions, the encoder in the model often exhibits empirical feature bias toward specific modalities [9, 35, 43, 47]. The feature bias exacerbates the divergence of feature representations across modalities, hindering the model's ability to establish semantically consistent multimodal fusion features. To address this, as shown in Fig. 3, we propose Semantic Consistency Inference (SCI). The core of SCI lies in enforcing the encoder to produce consistent semantic representations for identical semantic content across different modalities, thereby suppressing discrepant features caused by feature bias.

For the four hierarchical features $\mathbf{F}_{in}$ and $\mathbf{F}_{vi}$ extracted by the encoder from infrared and visible images, respectively, we first employ cosine similarity to compute cross-modal semantic similarity as a bias indicator:

$$\mathcal{S}_m = \frac{\langle \mathbf{F}_{in} \cdot \mathbf{F}_{vi} \rangle}{\|\mathbf{F}_{in}\| \cdot \|\mathbf{F}_{vi}\|} \tag{1}$$

where $\mathcal{S}_m$ is the bias indicator. We utilize a threshold $\tau = 0.4$ (*the analysis of $\tau$ is provided in the supplementary material.*), where $\mathbf{F}_{in}$ and $\mathbf{F}_{vi}$ are identified as being interfered by encoder bias when $\mathcal{S}_m < \tau$. For the features in $\mathbf{F}in$ and $\mathbf{F}vi$ that are affected by encoder bias, we compute their difference features as follows:

$$\begin{aligned} \mathbf{P}_{in} &= \mathbf{F}_{in} \odot (1 - \mathbf{M}_{sc}) + \mathbf{F}_{vi} \odot \mathbf{M}_{sc}, \\ \mathbf{P}_{vi} &= \mathbf{F}_{vi} \odot (1 - \mathbf{M}_{sc}) + \mathbf{F}_{in} \odot \mathbf{M}_{sc}, \end{aligned} \tag{2}$$

where $\odot$ is matrix multiplication and $\mathbf{M}_{sc}$ is an ambiguity mask generated by a multilayer perceptron (denoted as $\ell_\omega$):

$$\mathbf{M}_{sc} = f_N\left(\ell_\omega\left(\mathcal{CAT}\left(\mathbf{F}_{in}, \mathbf{F}_{vi}\right)\right)\right), \tag{3}$$

where $\mathcal{CAT}$ is the concatenation and $f_N$ is the Sigmoid. The role of $\mathbf{M}_{sc}$ is to discriminate bias-affected discrepant features by leveraging similar semantic content across modalities. Based on the $\mathbf{P}_{in}$ and $\mathbf{P}_{vi}$, we calculate the bias components of the encoder toward different modalities by contrasting them with the original features $\mathbf{F}_{in}, \mathbf{F}_{vi}$:

$$\Delta\mathbf{P}_{vi} = \mathbf{F}_{vi} - \mathbf{P}_{vi}, \Delta\mathbf{P}_{in} = \mathbf{F}_{in} - \mathbf{P}_{in}. \tag{4}$$

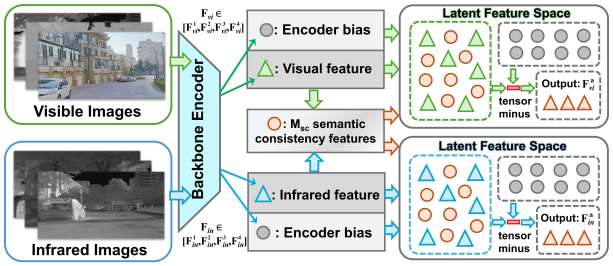


Figure 3. Structure of the Semantic Consistency Inference.

Finally, by introducing a learnable parameter $\lambda$, we suppress the discrepant features influenced by the feature bias from the original features using the bias components:

$$\mathbf{F}_{vi}^{a} = \mathbf{F}_{vi} - \lambda \Delta \mathbf{P}_{vi}, \mathbf{F}_{in}^{a} = \mathbf{F}_{in} - \lambda \Delta \mathbf{P}_{in}. \quad (5)$$

After discrepant features compression, $\mathbf{F}_{vi}^{a}$ and $\mathbf{F}_{in}^{a}$ can be more effectively aligned in subsequent modules, thereby simplifying the modeling complexity of fusion features.

## 3.2. Feature Discrepancy Alignment Module

To address the visual discrepancies in multimodal images caused by their distinct appearances, we design the Feature Discrepancy Alignment Module (FDAM). The FDAM consists of local-alignment and global-alignment.

### 3.2.1. Local-Alignment

The role of local-alignment is to align features between modalities from the perspective of fine-grained visual features by leveraging local attention mechanisms [23, 56]. As shown in Fig. 2, for outputs $\{\mathbf{F}_{vi}^{a}, \mathbf{F}_{in}^{a}\}$ of SCI, we apply global max and average pooling along the channel dimension to capture effective response characteristics. We then process these feature responses with an MLP to generate attention weights $\mathcal{A}_{vi,in}^{ch}$ for the effective and ineffective features. Based on the $\mathcal{A}_{vi,in}^{ch}$, we multiply them back into $\mathbf{F}_{vi,in}^{a}$ ($\mathcal{A}_{vi,in}^{ch} \odot \mathbf{F}_{vi,in}^{a}$) to generate the feature cues $\mathbf{F}_{vi,in}^{m}$:

$$\mathcal{A}_{vi,in}^{ch} = f_N\left(\ell_\omega\left(\vartheta_a\left(\mathbf{F}_{vi,in}\right)\right) + \ell_\omega\left(\vartheta_m\left(\mathbf{F}_{vi,in}\right)\right)\right), \quad (6)$$

where $\vartheta_a$ and $\vartheta_m$ represent the global average and max pooling, respectively. Traditional squeeze-excitation methods [6, 11, 12] focus on enhancing effective features but are challenged in suppressing visual discrepancies. To address this issue, we calculate the difference between the feature cues $\mathbf{F}_{vi,in}^{m}$ and the initial features $\mathbf{F}_{vi,in}$ to eliminate the discrepancies between modalities. We then use the sigmoid to add the reallocated weight results to the effective features, further compressing the discrepant features:

$$\mathbf{F}_{vi,in}^{ch} = f_N\left(\mathbf{F}_{vi,in}^{m} - \mathbf{F}_{vi,in}\right) \odot \mathbf{F}_{vi,in} + \mathbf{F}_{vi,in}. \quad (7)$$

From the local spatial perspective, we apply max pooling and average pooling operations to $\{\mathbf{F}_{vi}, \mathbf{F}_{in}\}$ to capture

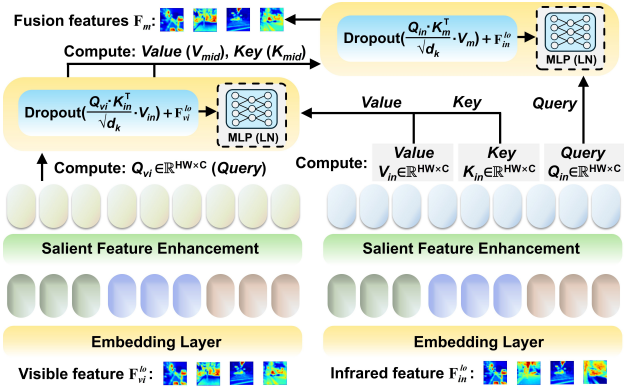


Figure 4. Structure of the Saliency Cross-Modal Attention.

their pixel-level response characteristics in the local spatial dimension. Then, we use a convolution kernel to map the effective local correlation features and apply a sigmoid to generate spatial attention weights $\mathcal{A}_{vi,in}^{sp}$. Then, we multiply $\mathcal{A}_{vi,in}^{sp}$ with $\mathbf{F}_{vi,in}^{a}$ ($\mathcal{A}_{vi,in}^{sp} \odot \mathbf{F}_{vi,in}^{a}$) to generate the feature clues $\mathbf{F}_{vi,in}^{n}$ of spatial dimension:

$$\mathcal{A}_{vi,in}^{sp} = f_N\left(Conv\left(\mathcal{P}_a\left(\mathbf{F}_{vi,in}\right), \mathcal{P}_m\left(\mathbf{F}_{vi,in}\right)\right)\right), \quad (8)$$

where $\mathcal{P}_a$ and $\mathcal{P}_m$ represent average and max pooling, respectively. Similarly, we use the sigmoid to process the difference between the feature cues $\mathbf{F}_{vi,in}^{n}$ and the initial features $\mathbf{F}_{vi,in}$, eliminating visual discrepancy features within the local spatial dimension:

$$\mathbf{F}_{vi,in}^{sp} = f_N\left(\mathbf{F}_{vi,in}^{n} - \mathbf{F}_{vi,in}\right) \odot \mathbf{F}_{vi,in} + \mathbf{F}_{vi,in}. \quad (9)$$

We add the features $\{\mathbf{F}_{vi}^{ch} + \mathbf{F}_{vi}^{ch}, \mathbf{F}_{in}^{sp} + \mathbf{F}_{in}^{sp}\}$ to obtain local alignment results $\{\mathbf{F}_{vi}^{lo}, \mathbf{F}_{in}^{lo}\}$.

### 3.2.2. Global-Alignment

Local alignment focuses on aligning visual discrepancies from a local perspective of feature maps but lacks the capacity to address such discrepancies from a global perspective [2, 27]. Previous works [25, 60] have demonstrated that cross-modal long-range context modeling can facilitate multimodal feature alignment. However, non-critical features across modalities may lead to feature misjudgment during cross-modal matching, causing the model to discard detailed features of one modality. To address this, as shown in Fig. 4, we propose a saliency cross-modal attention in global-alignment. Our method aligns features based on the salient characteristics of each modality, effectively avoiding feature misjudgment caused by non-critical features.

First, we employ salient feature enhancement [32] to perform self-enhancement on the effective contextual features within $\{\mathbf{F}_{vi}^{lo}, \mathbf{F}_{in}^{lo}\}$. Then, we apply a linear layer to the enhanced $\mathbf{F}_{vi}^{lo}$ to compute $Q_{vi} \in \mathbb{R}^{HW \times C}$, and use a linear layer on the enhanced $\mathbf{F}_{in}^{lo}$ to compute $K_{in} \in \mathbb{R}^{HW \times C}$ and

Table 1. Quantitative comparison results on the MFNet. **Best** and second-best are highlighted in **bold** and underline (mIoU).

| Method | Car | Person | Bike | Curve | Car stop | Cone | Bump | Guard. | mIoU |
|---|---|---|---|---|---|---|---|---|---|
| EGFNet [69] | 87.6 | 69.8 | 58.8 | 42.8 | 33.8 | 48.3 | 47.1 | 7.2 | 54.8 |
| SeAFusion [44] | 84.2 | 71.1 | 58.7 | 33.1 | 20.1 | 40.4 | 33.9 | 0.1 | 48.8 |
| LASNet [17] | 84.2 | 67.1 | 56.9 | 41.1 | 39.6 | 48.8 | 40.1 | **18.9** | 54.9 |
| EAEFNet [22] | 87.6 | **77.6** | 63.8 | 48.6 | 35.3 | 52.4 | 58.3 | 14.2 | 58.9 |
| MDRNet+ [63] | 87.1 | 69.8 | 60.9 | 47.8 | 34.2 | 50.2 | 55.1 | 8.2 | 56.8 |
| SGFNet [50] | 88.4 | 76.6 | 64.3 | 45.8 | 31.2 | 57.1 | 55.4 | 6.1 | 57.6 |
| MMSMC [70] | 89.2 | 69.1 | 63.5 | 46.4 | 41.9 | 48.8 | 57.6 | 8.8 | 58.1 |
| SegMiF [25] | 87.8 | 71.4 | 63.2 | 47.5 | 31.1 | 48.9 | 50.3 | 0.1 | 56.1 |
| MRFS [60] | 89.4 | 75.4 | **64.9** | 49.2 | 37.2 | 53.1 | **58.8** | 5.4 | 59.1 |
| OpenRss [62] | 87.4 | 68.9 | 63.2 | 46.8 | 43.2 | 49.5 | 56.6 | 7.4 | 58.5 |
| **Ours** | **90.9** | 76.9 | 62.1 | **49.4** | **51.7** | **57.3** | 58.4 | 14.6 | **62.1**$^{\uparrow 3\%}$ |

Table 2. Quantitative comparison results on the FMB (mIoU).

| Method | Car | Person | Truck | T-Lamp | T-Sign | Build. | Vegeta. | Pole | mIoU |
|---|---|---|---|---|---|---|---|---|---|
| SeAFusion [44] | 76.2 | 59.6 | 15.1 | 34.4 | 68.1 | 80.1 | 83.5 | 38.4 | 51.9 |
| LASNet [17] | 73.2 | 58.3 | 33.1 | 32.6 | 68.5 | 80.8 | 83.4 | 41.1 | 55.7 |
| EAEFNet [22] | 79.7 | 61.6 | 22.5 | 34.3 | 74.6 | 82.3 | 86.6 | 46.2 | 58.1 |
| MDRNet+ [63] | 75.4 | 67.1 | 27.1 | 41.4 | 68.4 | 79.8 | 82.7 | 45.3 | 55.5 |
| SGFNet [50] | 75.2 | 67.2 | 34.6 | 45.8 | 71.4 | 78.2 | 82.7 | 42.8 | 56.1 |
| SegMiF [25] | 78.7 | 65.5 | 42.4 | 35.6 | 71.7 | 80.1 | 85.1 | 35.7 | 58.5 |
| MRFS [60] | 76.2 | 71.3 | 34.4 | **50.1** | **75.8** | 85.4 | 87.1 | **53.6** | 61.2 |
| OpenRss [62] | 74.5 | 63.5 | 41.5 | 35.3 | 71.5 | 77.1 | 84.3 | 36.2 | 59.3 |
| **Ours** | **83.6** | **71.5** | **57.7** | 37.4 | 73.8 | **85.9** | **87.3** | 52.8 | **64.8**$^{\uparrow 3.6\%}$ |

$V_{in} \in \mathbb{R}^{HW \times C}$. $Q_{vi}$ is used to query $K_{in}$ to obtain cross-modal matching scores. These scores highlight the different focuses on effective and ineffective information across the multimodal features. Then, we refine the consistent representation of both multimodal features by modulating $V_{in}$:

$$\mathcal{M}_{cross}^{vi} = \textbf{Dropout}\left(\left(Q_{vi} \cdot K_{in}^{\mathrm{T}}\right)/\sqrt{d_k}\right) \cdot V_{in}, \tag{10}$$

where $d_k$ is the number of heads and $\mathcal{M}_{cross}^{vi}$ is the long-range refined value based on visible features. Then, we use $\mathcal{M}_{cross}^{vi}$ to enhance $\mathbf{F}_{vi}^{lo}$, resulting in an improved features $\mathbf{F}_{vi}^{gv}$ based on the global view of infrared features:

$$\mathbf{F}_{mid} = \ell_{\omega}\left(\mathcal{M}_{cross}^{vi} + \mathbf{F}_{vi}^{lo}\right). \tag{11}$$

The same approach is applied to use $Q_{in}$, generated from $\mathbf{F}_{in}^{lo}$, to match $K_{mid}$ and $V_{mid}$, generated from $\mathbf{F}_{mid}$. Then, we compute the long-range refined value $\mathcal{M}_{cross}^{in}$ based on infrared features. We use $\mathcal{M}_{cross}^{in}$ to modulate and refine the features $\mathbf{F}_{in}^{lo}$ from the global perspective of visible features, finally outputting multimodal fusion features $\mathbf{F}_m$. Our cross-modal attention mechanism seamlessly integrates effective information from both infrared and visible-light modalities into consistent semantic features that are easy to fuse and segment from a global perspective.

### 3.3. Mutual Feature Mask Learning

The different feature distributions in multimodal images often result in varying degrees of contribution from each modality's features to the prediction. In such cases, it becomes challenging for the network to learn complementary cross-modal features. To address this, as shown in Fig. 2, we propose a mutual feature mask learning (MFML) strategy to promote the complementarity and fusion of features across modalities. Unlike directly applying masks to images [37], the innovation of MFML lies in performing pixel-level masking directly on the feature maps, thus preventing the backbone from falsely reconstructing the image masks. For the input $\{\mathbf{F}_{vi}^{a}, \mathbf{F}_{in}^{a}\}$, we apply masks along the channel dimension of the feature map, randomly masking the features of one modality to generate mask features $\{\mathbf{F}_{in}^{m}, \mathbf{F}_{vi}^{m}\}$:

$$\begin{aligned} \mathbf{F}_{in}^{m}, \mathbf{F}_{vi}^{m} &= \mathcal{R}\left\{\mathbf{F}_{in}^{c} \textbf{ or } \mathbf{F}_{vi}^{c}\right\} \odot \mathbf{M}_{mask}, \\ \mathbf{M}_{mask} &= \sum_{c}^{C} \cdot \sum_{h,w}^{H,W} \{0 \textbf{ or } 1\}, \end{aligned} \tag{12}$$

where $\mathcal{R}$ represents the random selection of either an infrared or visible-light feature map for the feature masking operation. $\mathbf{M}_{mask}$ is a mask matrix with the same dimensions as the selected feature map, where each pixel value is either 0 or 1. Based on $\{\mathbf{F}_{vi}^{m}, \mathbf{F}_{in}^{m}\}$, we use a consistency regularization loss $\mathcal{L}_{cr}$ to measure the consistency between the predictions of the masked and unmasked features:

$$\mathcal{L}_{cr} = \Phi_{ce}\left(\mathcal{D}\left\{\mathbf{F}_{in}^{c}, \mathbf{F}_{vi}^{c}\right\}, \mathcal{D}\left\{\mathbf{F}_{in}^{m}, \mathbf{F}_{vi}^{m}\right\}\right), \tag{13}$$

where $\mathcal{D}(\cdot)$ denotes the decoder [60]. $\Phi_{ce}(\cdot)$ represents the cross-entropy loss. Our objective in consistency regularization prediction [31, 39, 58] is to minimize $\mathcal{L}_{cr}$, thereby promoting complementarity between features from different modalities through masked prompts.

Table 3. Quantitative comparison results on the PST900 (mIoU).

| Method | Hand-Drill | Back-Pack | Fire-Exti. | Survivor | mIoU |
|---|---|---|---|---|---|
| EGFNet [69] | 64.7 | 83.1 | 71.3 | 74.3 | 78.8 |
| SeAFsuion [44] | 65.6 | 59.6 | 41.1 | 29.5 | 58.9 |
| LASNet [17] | 77.8 | 86.5 | 82.8 | 75.5 | 84.4 |
| EAEFNet [22] | 80.4 | 87.7 | 84.1 | 76.2 | 85.6 |
| MDRNet+ [63] | 66.3 | 81.4 | 76.3 | 71.3 | 79.7 |
| SGFNet [50] | 62.4 | **89.2** | 73.3 | 72.7 | 79.8 |
| MMSMC [70] | **82.8** | 75.8 | 79.9 | 74.7 | 82.1 |
| SegMiF [25] | 63.2 | 76.3 | 63.5 | 75.5 | 74.6 |
| MRFS [60] | 79.7 | 87.4 | 88.2 | 79.6 | 86.9 |
| OpenRss [62] | 78.3 | 84.2 | 83.7 | 72.1 | 84.2 |
| **Ours** | 81.2 | 87.7 | **88.6** | **80.8** | **88.5**$^{\uparrow 1.6\%}$ |

Table 4. Quantitative comparison results (mAP).

| Method | MFNet-dataset | FMB-dataset | PST-dataset |
|---|---|---|---|
| SeAFsuion [44] | 67.7 | 68.8 | 76.9 |
| LASNet [17] | 75.2 | 74.6 | 82.5 |
| EAEFNet [22] | 70.5 | 71.8 | 81.4 |
| MDRNet+ [63] | 68.9 | 69.5 | 78.8 |
| SGFNet [50] | 73.5 | 72.4 | 85.2 |
| SegMiF [25] | 74.2 | 76.2 | 88.7 |
| MRFS [60] | 75.6 | 75.6 | 90.2 |
| OpenRss [62] | 72.7 | 75.3 | 87.8 |
| **Ours** | **77.4**$^{\uparrow 1.8\%}$ | **77.1**$^{\uparrow 0.9\%}$ | **91.8**$^{\uparrow 1.6\%}$ |

### 3.4. Loss Function

The total loss function is composed of the image fusion loss $\mathcal{L}_{fus}$, the semantic segmentation loss $\mathcal{L}_{seg}$, and the mask consistency regularization loss $\mathcal{L}_{cr}$. We use the cross-entropy loss as the semantic segmentation loss:

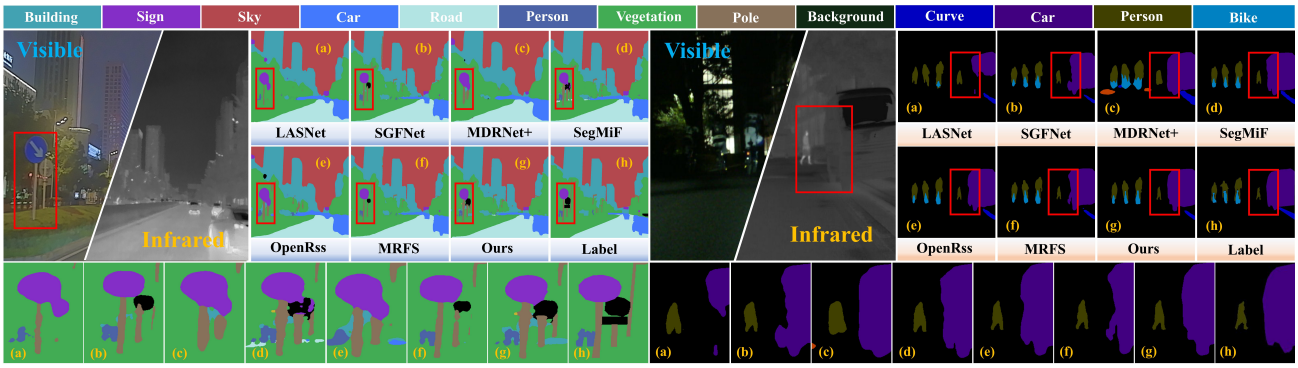


Figure 5. Qualitative comparison results on the FMB and MFNet datasets, respectively.

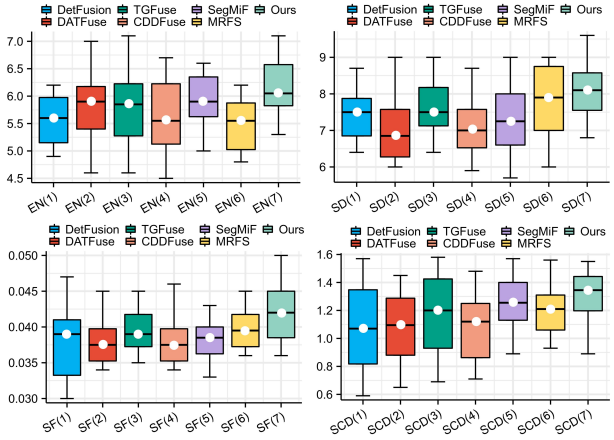


Figure 6. Quantitative comparisons of image fusion with SOTA methods on MFNet, in which the circles indicate mean values.

$$\mathcal{L}_{seg}\left(p_s, g\right) = -\sum g \times \log\left(p_s\right), \tag{14}$$

where $p_s$ and $g$ represent the prediction and the ground truth, respectively. Following prior work [60], we use salient information loss $\mathcal{L}_{si}$ and consistency color loss $\mathcal{L}_{co}$ to define the image fusion loss $\mathcal{L}_{fus} = \mathcal{L}_{si} + \mathcal{L}_{co}$:

$$\begin{aligned} \mathcal{L}_{si} &= \left\| \nabla\left(p_f - \max\left(I_{vi}, I_{in}\right)\right)\right\|, \\ \mathcal{L}_{co} &= \left\| p_f|_{cr+cb} - \mathcal{S}\left(I_{vi}|_{cr+cb}\right)\right\|, \end{aligned} \tag{15}$$

where the $\nabla$ and $||$ of $\mathcal{L}_{si}$ denote the gradient operator and the mean absolute error (MAE), while $p_f$ represents the fused image. Specifically, we use the fusion head of [60] to process the output of FDAM to generate the fused image. In $\mathcal{L}_{co}$, $cr$ and $cb$ refer to the red and blue chrominance components obtained after converting the image to the YCrCb color space [19]. $\mathcal{S}\left(\cdot\right)$ denotes data augmentation. Overall, the total loss $\mathcal{L}_{total}$ is defined as:

$$\mathcal{L}_{total} = \alpha_1\mathcal{L}_{seg} + \alpha_2\mathcal{L}_{fus} + \alpha_3\mathcal{L}_{cr}, \tag{16}$$

where $\alpha_1$, $\alpha_2$ and $\alpha_3$ are hyperparameters, the specific value is 1, 0.5, and 0.5, respectively. *The analysis of $\alpha_1$, $\alpha_2$ and $\alpha_3$ is provided in the supplementary material.*

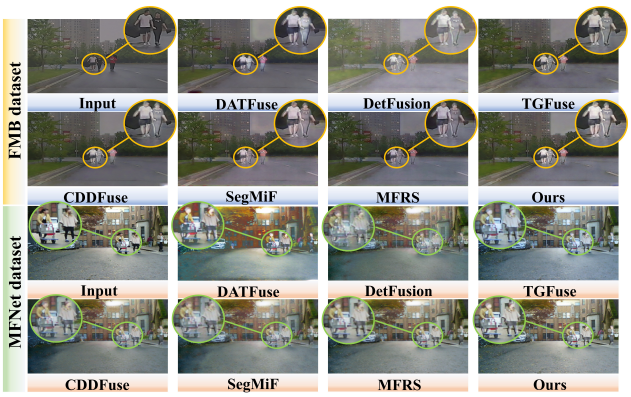


Figure 7. Visual comparison results of different image fusion methods on the FMB and MFNet datasets, respectively.

## 4. Experiments

**Datasets.** We conduct experiments on three representative multimodal semantic segmentation datasets: FMB [60], MFNet [10], and PST900 [38].

**Implementation details.** Our AMDANet is trained and evaluated using an NVIDIA Tesla A40 (45G) GPU. We employ the adaptive moment estimation (Adam) optimizer with an initial learning rate of $6\times10^{-5}$ and adopt a warm-up strategy for learning rate scheduling. We use the mean Intersection over Union (mIoU) and mean Average Precision (mAP) to evaluate segmentation performance.

### 4.1. Results of semantic segmentation

We show the superiority of our method in multimodal semantic segmentation by qualitatively and quantitatively comparing it with ten advanced methods: EGFNet [69], SeAFusion [44], LASNet [17], EAEFNet [22], MDRNet+ [63], SGFNet [50], MMSMC [70], SegMiF [25], MFRS [60], and OpenRss [62].

Tab. 1, 2, 3, and 4 present the quantitative comparison results on the FMB, MFNet, and PST 900 datasets, respectively. The results show that our method outperforms all comparable methods across all datasets, achieving the highest mIoU and mAP scores. The sparse image quality [60]

Table 5. Ablation studies of different components on MFNet.

| Dataset | SCI | MFML | FDAM | mIoU | mAP |
|---|---|---|---|---|---|
| MFNet | ✗ | ✗ | ✗ | 53.1 | 61.7 |
| | ✗ | ✔ | ✔ | 58.1 | 69.2 |
| | ✔ | ✗ | ✔ | 59.7 | 73.5 |
| | ✔ | ✔ | ✗ | 58.6 | 71.2 |
| | ✔ | ✔ | ✔ | **62.1** | **77.1** |

Table 6. Ablation studies of different loss functions on FMB.

| Dataset | $\mathcal{L}_{cr}$ | $\mathcal{L}_{co}$ | $\mathcal{L}_{si}$ | mIoU | mAP |
|---|---|---|---|---|---|
| FMB | ✗ | ✔ | ✔ | 61.6 | 75.2 |
| | ✔ | ✗ | ✔ | 62.3 | 75.8 |
| | ✔ | ✔ | ✗ | 60.5 | 73.5 |
| | ✔ | ✔ | ✔ | **64.8** | **77.4** |

of the PST900 interferes with the boundary between valid and invalid features, resulting in a smaller improvement for our method compared to its gains on the FMB and MFNet. Despite this, our method achieves satisfactory results across the majority of categories in the FMB and MFNet datasets. Compared to the advanced MSRF, our method improves the mIoU by 3.6%, 3.0%, and 1.6% on FMB, MFNet, and PST 900, respectively. This improvement validates the effectiveness of our method in aligning feature discrepancies between modalities from multiple perspectives. Fig. 5 illustrates qualitative comparisons on the FMB and MFNet datasets across different methods. The results show that our method can produce more accurate results. For example, our method precisely delineates the contours of pedestrians, while other methods only approximate the area.

### 4.2. Results of image fusion

We further demonstrate the superiority of our method in multimodal image fusion through qualitative and quantitative comparisons with six advanced methods: DetFusion [42], DATFuse [46], TGFuse [34], CDDFuse [65], SegMiF [25], and MRFS [60]. Following the work [25], we use entropy (EN), standard deviation (SD), spatial frequency (SF), and structural correlation difference (SCD) to evaluate the quantitative results on the FMB dataset across different methods. As shown in Fig. 6, our method achieves the highest EN and SCD, indicating that our method's joint enhancement of multimodal features leads to improved visual quality. Additionally, our method achieves higher SF and SD scores, reflecting its ability to suppress discrepant features and produce outputs with rich textural details. Fig. 7 presents a qualitative comparison of the FMB and MFNet datasets. Our method excels in enhancing subtle texture information and improving realistic visual quality. For instance, as shown in Fig. 7, our fused images improve pedestrian visibility in low-light conditions. Furthermore, our method effectively aligns semantic ambiguities across multimodal features from multiple perspectives, producing

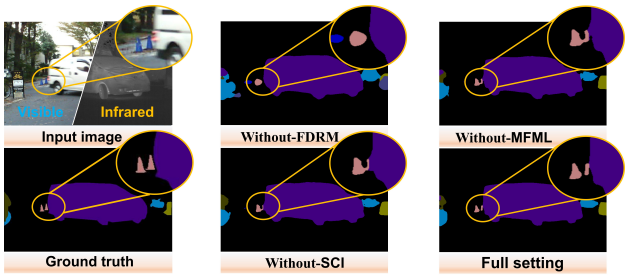


Figure 8. Visual ablation results of different components.

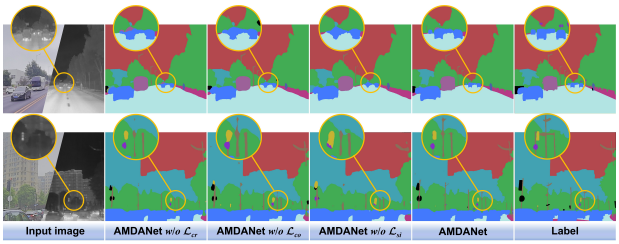


Figure 9. Visual ablation results of different loss functions.

semantically coherent fusion results. For example, in Fig. 7, our method successfully suppresses artifacts typically introduced by infrared images.

### 4.3. Ablation studies

We conduct ablation studies to evaluate the effectiveness of the modules in AMDANet, including: removing all modules, removing SCI, removing MFML, removing FDAM, and without any modifications. For removing Global-Eliminating, we fuse the infrared and visible features output by Local-Eliminating to generate the fusion features.

Tab. 5 shows the performance of these different variants on the MFNet dataset. As seen in Tab. 5, all variants exhibit varying degrees of performance decline compared to the AMDANet, demonstrating the effectiveness of each module design. The variant without FDAM shows the most significant drop in performance, indicating that the absence of inter-modal ambiguity suppression hinders the formation of consistent features that facilitate semantic segmentation. Fig. 8 presents visual comparison results. Examining the barricade in Fig. 8 reveals that while AMDANet without MFML can approximate the barricade's general location, it struggles to accurately depict its structural shape. This is because the different feature distributions between modalities increase the difficulty of modeling fusion features. Similar issues appear in AMDANet without SCI, suggesting that feature fusion is complicated by feature bias of the encoder and feature discrepancies between modalities.

To validate the effectiveness of $\mathcal{L}_{co}$, $\mathcal{L}_{si}$, and $\mathcal{L}_{cr}$, we conduct corresponding ablation studies. As shown in Tab. 6, removing the $\mathcal{L}_{co}$ leads to mIoU and mAP decreases of 2.5% and 1.6%, respectively, while removing the $\mathcal{L}_{si}$ resulted in mIoU and mAP reductions of 4.3% and 3.9%. These findings indicate that $\mathcal{L}_{co}$ and $\mathcal{L}_{si}$ are effective in

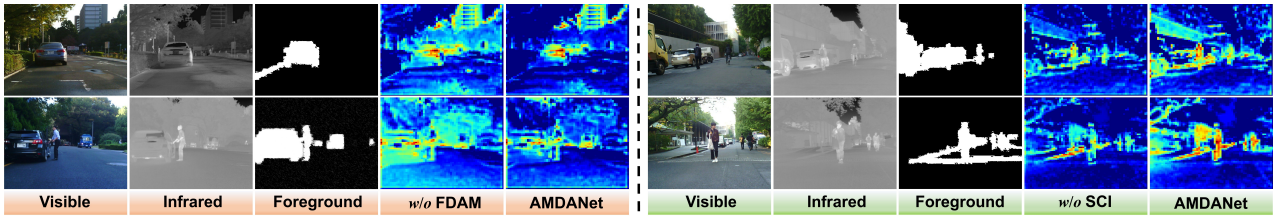


Figure 10. Visual comparison of AMDANet output features in different settings.

Table 7. Quantitative comparisons of Parameters and FLOPs

| Method | #Params (M) | FLOPs (G) | mIoU |
|---|---|---|---|
| EAEFNet [22] | 316.49 | 147.21 | 58.1 |
| SeAFusion [44] | **102.53** | **13.06** | 51.9 |
| SegMiF [25] | 526.21 | 45.61 | 58.5 |
| MFRS [60] | 219.16 | 134.97 | 61.2 |
| Ours | 197.85 | 119.02 | **64.5** |

guiding the model to establish multimodal features suitable for both visual and segmentation tasks. Similarly, removing the $\mathcal{L}_{cr}$ leads to mIoU and mAP drops of 3.2% and 2.2%, demonstrating that $\mathcal{L}_{cr}$ effectively promotes the network to fuse multimodal features. The second and third columns in Fig. 9 illustrate that, without $\mathcal{L}_{co}$ and $\mathcal{L}_{si}$, the model struggles to establish accurate multimodal information, leading to poor coherence in the edge. In the fourth column of Fig. 9, the absence of $\mathcal{L}_{cr}$ introduces ambiguous information into the results, reducing the model's ability to distinguish similar regions and causing color blending artifacts. In contrast, the complete AMDANet captures image details and boundary information more accurately.

### 4.4. Discussion about SCI and FDAM

To validate the effectiveness of FDAM and SCI, we apply Grad-CAM [36] to the intermediate features of the decoder on the MFNet dataset, comparing the performance of AMDANet without ($w/o$) FDAM and SCI. As shown in the left part of Fig. 10, the CAM map generated by AMDANet without FDAM is more chaotic compared to the full AMDANet, with excessive attention allocated to irrelevant regions. This is due to significant distribution differences between modalities, causing the fused features to retain ambiguous characteristics from both modalities. Under the interference of these feature discrepancies, the network mislocalizes critical information of the image. In contrast, FDAM effectively compresses the ambiguous features between modalities from both local and global views, facilitating the establishment of a more accurate and consistent feature representation. As shown in the right part of Fig. 10, compared to AMDANet with SCI, AMDANet without SCI generates CAM maps that show insufficient attention to critical information. For example, in the second row of Fig. 10, while the network accurately locates the pedestrian, it assigns a lower heat value to the pedestrian area. This is because the encoder's biased feature tendencies toward different modality images diffuse some of the network's attention, hindering the establishment of consistent features. In contrast, with SCI enabled, the network can focus more effectively on establishing semantically consistent features.

### 4.5. Complexity analysis

We compare the Parameters and FLOPs of different methods on the FMB dataset to evaluate the complexity of each method. In Tab. 7, while our method does not achieve the best performance in terms of Params and FLOPs, its complexity remains reasonable and even surpasses some comparable methods. For instance, compared to EAEFNet, our method achieves a reduction in both Params and FLOPs while improving the mIoU by 6.4%. Although our method does not perform as well as SeAFusion in terms of Params and FLOPs, we achieve a significant improvement of 12.6% in mIoU by sacrificing some lightweight capabilities.

## 5. Conclusion and limitation

In this paper, we propose AMDANet, a multimodal semantic segmentation method that constructs segmentable fusion features by aligning modality-discrepant features from a multi-level perspective. We introduce the Feature Discrepancy Attention Module (FDAM) to suppress irrelevant features and propose Semantic Consistency Inference (SCI) to alleviate feature bias caused by the encoder. Extensive experiments on the MFNet, FMB, and PST900 datasets validate the superior segmentation performance of AMDANet. Despite its high accuracy, the method's computational complexity limits its deployment on resource-constrained platforms. As part of future work, we plan to adopt model compression strategies, such as knowledge distillation, to reduce complexity while preserving segmentation accuracy.

## 6. Acknowledgements

This research was supported by the National Natural Science Foundation of China under Grant Nos. 62203184 and W2421093, the International Cooperation Project of Jilin Province under Grant No. 20250205079GH and the Beijing Natural Science Foundation under No. L221013. This research was also supported by an IITP grant (No. RS-2022-II220608) funded by the MSIT, Republic of Korea.

## References


[1] Zhaochong An, Guolei Sun, Yun Liu, Runjia Li, Min Wu, Ming-Ming Cheng, Ender Konukoglu, and Serge Belongie. Multimodality helps few-shot 3d point cloud semantic segmentation, 2025. 2

[2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12299–12310, 2021. 4

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[4] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3333–3348, 2021. 2

[5] Xin Fan, Xiaolin Wang, Jiaxin Gao, Jia Wang, Zhongxuan Luo, and Risheng Liu. Bi-level learning of task-specific decoders for joint registration and one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11726–11735, 2024. 1

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[7] Fangyuan Gao, Xin Deng, Mai Xu, Jingyi Xu, and Pier Luigi Dragotti. Multi-modal convolutional dictionary learning. *IEEE Transactions on Image Processing*, 31:1325–1339, 2022. 2

[8] Suining Gao, Xiubin Yang, Li Jiang, Zongqiang Fu, and Jiamin Du. Global feature-based multimodal semantic segmentation. *Pattern Recognition*, 151:110340, 2024. 1, 2

[9] Sander Greenland, Judea Pearl, and James M. Robins. Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29 – 46, 1999. 3

[10] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017. 2, 6

[11] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13713–13722, 2021. 4

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[13] Yubin Hu, Yuze He, Yanghao Li, Jisheng Li, Yuxing Han, Jiangtao Wen, and Yong-Jin Liu. Efficient semantic segmentation by altering resolutions for compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22627–22637, 2023. 1

[14] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *Computer Vision – ECCV 2022*, pages 539–555, Cham, 2022. Springer Nature Switzerland. 2

[15] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masked transformers for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 752–761, 2023. 1

[16] Hyungjoo Jung, Youngjung Kim, Hyunsung Jang, Namkoo Ha, and Kwanghoon Sohn. Unsupervised deep image fusion with structure tensor representations. *IEEE Transactions on Image Processing*, 29:3845–3858, 2020. 2

[17] Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1223–1235, 2023. 5, 6

[18] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. 2

[19] Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024. 6

[20] Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021. 2

[21] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11040–11052, 2023. 2

[22] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fusion for rgb-thermal perception tasks. *IEEE Robotics and Automation Letters*, 8(7):4060–4067, 2023. 5, 6, 8

[23] Shuyuan Lin, Xiao Chen, Guobao Xiao, Hanzi Wang, Feiran Huang, and Jian Weng. Multi-stage network with geometric semantic attention for two-view correspondence learning. *IEEE Transactions on Image Processing*, 33:3031–3046, 2024. 4

[24] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5811, 2022. 2

[25] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. 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)*, pages 8115–8124, 2023. 1, 2, 4, 5, 6, 7, 8

[26] Risheng Liu, Zhu Liu, Jinyuan Liu, and Xin Fan. Searching a hierarchically aggregated fusion architecture for fast multi-modality image fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 1600–1608, New York, NY, USA, 2021. Association for Computing Machinery. 2

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 4

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[29] Jiayi Ma, Pengwei Liang, Wei Yu, Chen Chen, Xiaojie Guo, Jia Wu, and Junjun Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020. 2

[30] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2021. 2

[31] Huayu Mai, Rui Sun, Tianzhu Zhang, and Feng Wu. Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3391–3401, 2024. 5

[32] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2022. 4

[33] Fangzhou Meng, Xiangyu Chen, Hongying Tang, Chaoyi Wang, and Guanjun Tong. B2mfuse: A bi-branch multiscale infrared and visible image fusion network based on joint semantics injection. *IEEE Transactions on Instrumentation and Measurement*, 73:1–17, 2024. 2

[34] Dongyu Rao, Tianyang Xu, and Xiao-Jun Wu. Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Transactions on Image Processing*, pages 1–1, 2023. 7

[35] Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*, 42(5): 1511–1519, 2013. 2, 3

[36] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 8

[37] Ukcheol Shin, Kyunghyun Lee, In So Kweon, and Jean Oh. Complementary random masking for rgb-thermal semantic segmentation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11110–11117, 2024. 3, 5

[38] Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9441–9447, 2020. 1, 2, 6

[39] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3107, 2024. 5

[40] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019. 2

[41] Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion. *IEEE Transactions on Automation Science and Engineering*, 18(3):1000–1011, 2021. 1

[42] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 4003–4011. Association for Computing Machinery, 2022. 7

[43] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems*, pages 1513–1524. Curran Associates, Inc., 2020. 2, 3

[44] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022. 2, 5, 6, 8

[45] Wei Tang, Fazhi He, and Yu Liu. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Transactions on Multimedia*, 25:5413–5428, 2023. 1, 2

[46] Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023. 7

[47] Xudong Wang, Long Lian, and Stella X. Yu. Unsupervised visual attention and invariance for reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6677–6687, 2021. 2, 3

[48] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. In *Advances in Neural Information Processing Systems*, pages 4835–4845. Curran Associates, Inc., 2020. 1

[49] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12186–12195, 2022. 1

[50] Yike Wang, Gongyang Li, and Zhi Liu. Sgfnet: Semantic-guided fusion network for rgb-thermal semantic segmenta-

tion. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7737–7748, 2023. 5, 6

[51] Zongwei Wu, Jingjing Wang, Zhuyun Zhou, Zhaochong An, Qiuping Jiang, Cédric Demonceaux, Guolei Sun, and Radu Timofte. Object segmentation by mining cross-modal semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 3455–3464, New York, NY, USA, 2023. Association for Computing Machinery. 1, 2

[52] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, pages 12077–12090. Curran Associates, Inc., 2021. 3

[53] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12484–12491, 2020. 2

[54] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. 2

[55] Jiacong Xu, Zixiang Xiong, and Shankar P. Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19529–19539, 2023. 1

[56] Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9893–9903, 2023. 4

[57] Bin Yang, Yuxuan Hu, Xiaowen Liu, and Jing Li. Cefusion: An infrared and visible image fusion network based on cross-modal multi-granularity information interaction and edge guidance. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–16, 2024. 2

[58] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7236–7246, 2023. 5

[59] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12797–12804, 2020. 2

[60] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26974–26983, 2024. 1, 2, 4, 5, 6, 7, 8

[61] Qiang Zhang, Tonglin Xiao, Nianchang Huang, Dingwen Zhang, and Jungong Han. Revisiting feature fusion for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1804–1818, 2021. 2

[62] Guoqiang Zhao, Junjie Huang, Xiaoyun Yan, Zhaojing Wang, Junwei Tang, Yangjun Ou, Xinrong Hu, and Tao Peng. Open-vocabulary rgb-thermal semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 5, 6

[63] Shenlu Zhao, Yichen Liu, Qiang Jiao, Qiang Zhang, and Jungong Han. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9380–9394, 2024. 5, 6

[64] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1186–1196, 2022. 2

[65] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5906–5916, 2023. 1, 2, 7

[66] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890, 2021. 2

[67] Man Zhou, Jie Huang, Xueyang Fu, Feng Zhao, and Danfeng Hong. Effective pan-sharpening by multiscale invertible neural network and heterogeneous task distilling. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 2

[68] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802, 2021. 1, 2

[69] Wujie Zhou, Shaohua Dong, Caie Xu, and Yaguan Qian. Edge-aware guidance fusion network for rgb–thermal scene parsing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3571–3579, 2022. 2, 5, 6

[70] Wujie Zhou, Han Zhang, Weiqing Yan, and Weisi Lin. Mmsmcnet: Modal memory sharing and morphological complementary networks for rgb-t urban scene semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7096–7108, 2023. 5, 6

[71] Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7099–7108, 2024. 2