IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

**Traitement du Signal**
Vol. 43, No. 2, April, 2026, pp. 727-739
Journal homepage: http://iieta.org/journals/ts

# Multimodal Image Fusion and Semantic Parsing for Complex Landscape Architecture Scenes

Zhongyu Zhou

School of Architecture, Nanyang Institute of Technology, Nanyang 473000, China

Corresponding Author Email: 3071012@nyist.edu.cn

Copyright: ©2026 The author. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ts.430214

**Received:** 10 October 2025
**Revised:** 22 January 2026
**Accepted:** 10 March 2026
**Available online:** 30 April 2026

***Keywords:***
*landscape architecture, multimodal image fusion, semantic parsing, bidirectional collaborative network, scene adaptation, mutual information*

**ABSTRACT**

Complex landscape architecture scenes are often characterized by challenges such as vegetation occlusion, significant heterogeneity among multimodal images, and semantic class imbalance. Traditional serial processing pipelines following a "fusion-first-then-parsing" paradigm are prone to losing semantic details and suffer from a disconnect between the fusion process and semantic parsing requirements, making it difficult to achieve precise recognition and interpretation of complex landscape elements. To address these challenges, this paper proposes a Multimodal Fusion and Semantic Parsing Network (MFSP-Net). We construct a bidirectional feedback architecture that embodies the principle of "fusion serving parsing, and parsing guiding fusion." Specifically, a semantic-guided cross-modal dynamic fusion (SGDF) module and a mutual information collaborative loss function are designed to realize efficient information integration and accurate semantic parsing in complex landscape scenarios. To validate the effectiveness of the proposed method, extensive experiments were conducted on the self-built Garden-MultiMod dataset. The results demonstrate that our method achieves a 12% improvement in mean Intersection over Union (mIoU) compared to the optimal single-modality methods. Furthermore, the fused image quality metrics—Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM)—reach state-of-the-art levels, while the inference efficiency meets real-time processing requirements. This study provides reliable technical support for the intelligent monitoring, planning, design, and ecological assessment of landscape architecture scenes.

## 1. INTRODUCTION

As a core component of urban ecosystems, the accuracy of scene parsing in landscape architecture directly affects the quality and efficiency of engineering practices such as landscape monitoring, planning, design, and ecological assessment. With the deep application of intelligent technologies in the field of landscape architecture, traditional single-modal image parsing methods [1, 2] have been unable to meet the demands of complex scenes—single RGB images [3, 4] are susceptible to lighting changes and vegetation occlusion and cannot capture deep scene information; Thermal Infrared (TIR) images [5, 6] can reflect target temperature characteristics but lack spatial texture details; depth images [7, 8] can provide three-dimensional structural information of scenes but have obvious deficiencies in texture representation. The complementarity of multimodal images provides an effective path to solve the above problems. Fusing RGB [9], TIR [10, 11], and Depth modalities [12, 13] can integrate scene information across different dimensions, providing data support for the accurate parsing of complex landscape architecture scenes. Currently, multimodal fusion [14] and semantic parsing technologies [15] in the field of image processing have made significant progress, with various fusion and parsing methods constantly emerging. However, landscape architecture scenes [16, 17] possess their own uniqueness. Characteristics such as dense vegetation distribution, irregular morphology of landscape elements, and unbalanced semantic category distribution lead to insufficient adaptability of existing methods in this scenario, making it difficult to balance fusion quality and parsing accuracy, and failing to fully meet the high demands of engineering practice for intelligent parsing. Therefore, conducting research on multimodal image fusion and semantic parsing for complex landscape architecture scenes has important theoretical value and engineering significance.

Although multimodal fusion and semantic parsing technologies have been applied in multiple fields, there are still many deficiencies in specialized research targeting complex landscape architecture scenes. These defects severely restrict the improvement of parsing accuracy and engineering practicality [18, 19]. Most existing methods adopt a serial processing mode of "fusion first, then parsing" [20, 21], isolating the fusion and parsing processes. The design of the fusion module lacks clear semantic guidance and only pursues image-level fusion effects, ignoring the core requirements of semantic parsing. This leads to the loss of key semantic details during the fusion process, such as sparse shrubs and curved garden path boundaries, thereby affecting subsequent parsing accuracy. In terms of fusion weight design, existing methods

mostly use fixed weights or simple adaptive strategies, lacking adaptability to the heterogeneity of landscape architecture scenes. They cannot dynamically adjust weight allocation according to the differences in modal complementarity across different landscape types, such as woodlands, waterfronts, and building-intensive areas, resulting in significant fluctuations in fusion effects across different scenes and making it difficult to achieve stable parsing performance. In addition, existing loss functions mostly impose constraints from the pixel level or feature level, without considering both the semantic sufficiency of fusion features and modal redundancy from an information theory perspective. It is difficult to establish an effective balance between fused image quality and semantic parsing accuracy. Especially in scenes with unbalanced semantic categories in landscape architecture, it is prone to the problem of majority classes dominating training and minority classes having low parsing accuracy, further limiting the practicality of the methods.

To address the aforementioned research gaps and improve the accuracy, efficiency, and stability of multimodal image fusion and semantic parsing in complex landscape architecture scenes, this paper proposes a Multimodal Fusion and Semantic Parsing Network (MFSP-Net), constructing a bidirectional feedback architecture of "fusion serving parsing, and parsing guiding fusion" to break the limitations of the traditional serial processing mode. The network introduces a triple iterative optimization mechanism to establish a bidirectional information flow between fusion and parsing, realizing the collaborative optimization of both. Aiming at the problem of insufficient scene adaptation of fusion weights, a semantic-guided cross-modal dynamic fusion (SGDF) module is designed. Based on scene semantic priors and modal complementarity measures, adaptive pixel-wise fusion weights are generated, allowing the fusion process to accurately adapt to the characteristic differences of different landscape architecture scenes. To enhance the semantic discriminative ability of fusion features, a mutual information collaborative loss function is proposed. From an information theory perspective, it maximizes the consistency between fusion features and semantic labels while minimizing information redundancy between different modalities, achieving a dual improvement in fusion quality and parsing accuracy. Furthermore, a Collaborative Parsing and Reconstruction (CPR) module is constructed. Through attention feedback and gradient sharing mechanisms, the boundary details and semantic correlations of fusion features are strengthened, further improving the parsing accuracy of landscape element boundaries and meeting the parsing needs of complex landscape architecture scenes.

The subsequent content of this paper will be structured as follows: First, we review the relevant research progress in the fields of multimodal image fusion and semantic parsing, clarifying the core limitations of existing methods and the entry point of this research. Second, we elaborate in detail on the overall architecture of the proposed collaborative network, the technical details of each core module, and the training and inference processes. Subsequently, through a series of experiments including ablation studies, comparative experiments, and robustness experiments, we systematically verify the effectiveness and superiority of the proposed method. Next, we deeply discuss the advantages reflected by the experimental results, the existing limitations, and future research directions. Finally, we summarize the research findings of this paper, clarifying the academic value and engineering application prospects of the method.

## 2. PROPOSED METHOD

### 2.1 Overall network architecture of multimodal fusion and semantic parsing network

To achieve efficient fusion and accurate semantic parsing of multimodal images in complex landscape architecture scenes, this paper proposes the MFSP-Net. The overall architecture adopts an end-to-end collaborative framework, consisting of four core components: a multimodal multi-scale feature extraction backbone, an SGDF module, a CPR module, and a joint optimization loss function. Each module is closely connected through a bidirectional information flow, forming a complete data processing and optimization closed loop. The network architecture is shown in Figure 1. Multimodal inputs generate modality features at different scales via the feature extraction backbone, which are then input into the dynamic fusion module to complete feature fusion. The fused features are sent to the CPR module to synchronously output semantic parsing results and fused images. Attention information and gradient information generated during the parsing process are fed back to the fusion module in reverse to realize the collaborative optimization of fusion and parsing. The joint optimization loss function imposes global constraints on the training process of the entire network, ensuring a dual improvement in fusion quality and parsing accuracy. The core design highlight of MFSP-Net lies in constructing a bidirectional feedback mechanism for fusion and parsing and introducing a triple iterative optimization process to achieve gradual performance enhancement: In the first iteration, the high-level modality features output by the feature extraction backbone undergo preliminary fusion and are then sent to the collaborative parsing module to generate rough semantic priors, providing a basis for the dynamic adjustment of subsequent fusion weights; In the second iteration, based on these semantic priors, the semantic guidance of the dynamic fusion module is strengthened, optimizing pixel-wise fusion weight allocation to make the fused features more suitable for semantic parsing requirements; The third iteration utilizes the boundary attention information output by the CPR module to further refine the boundary details of the fused features, improving the parsing accuracy of landscape element boundaries. The triple iterations form a progressive optimization that effectively solves the problems of semantic detail loss and boundary blurring inherent in traditional methods.

### 2.2 Multimodal feature extraction backbone

The preprocessing quality of multimodal inputs directly determines the effect of subsequent feature extraction and fusion. Aiming at the problem of registration deviation in the three-modal images within landscape architecture scenes, this paper designs a high-precision registration and targeted data augmentation pipeline to provide stable and robust inputs for feature extraction. First, RGB, TIR, and Depth three-modal images are registered. The Scale-Invariant Feature Transform (SIFT) algorithm is used to extract key feature points of each modal image, and mutual information is used to optimize the matching accuracy of feature points, effectively reducing spatial offsets between modalities and ensuring that the

registration error is controlled within 2 pixels, avoiding fusion distortion caused by registration deviation. After registration, Z-score normalization is performed on the three-modal images, respectively, to eliminate the impact of different data dimensions across modalities, placing each modal feature in the same feature space. To improve the generalization ability of the network and adapt to complex situations such as lighting changes and vegetation occlusion in landscape architecture scenes, a targeted data augmentation strategy is designed, including random horizontal flipping, brightness jitter, and random occlusion. Among these, the brightness jitter range is controlled between 0.7 and 1.3 times, and random occlusion uses rectangular occlusion blocks of 16 × 16 to 32 × 32, which not only simulates vegetation occlusion and light-shadow changes in actual scenes but also avoids feature distortion caused by excessive augmentation.

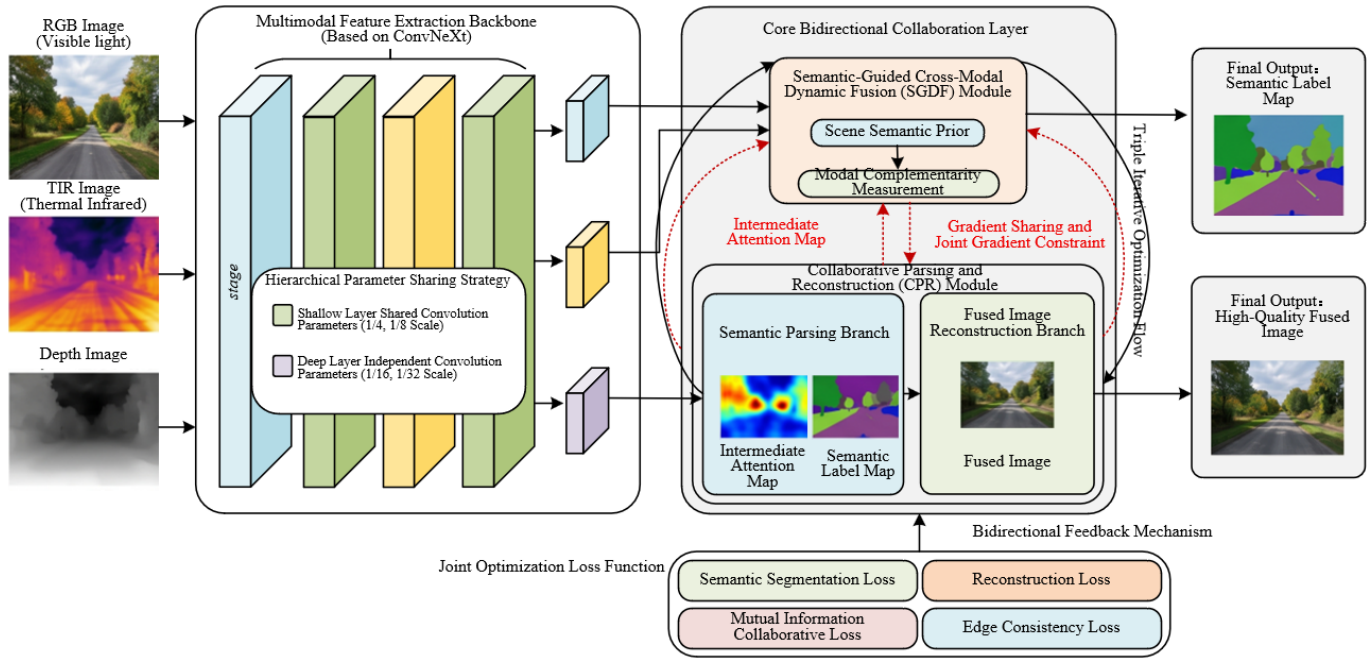


**Figure 1.** Overall network architecture diagram of Multimodal Fusion and Semantic Parsing Network (MFSP-Net)

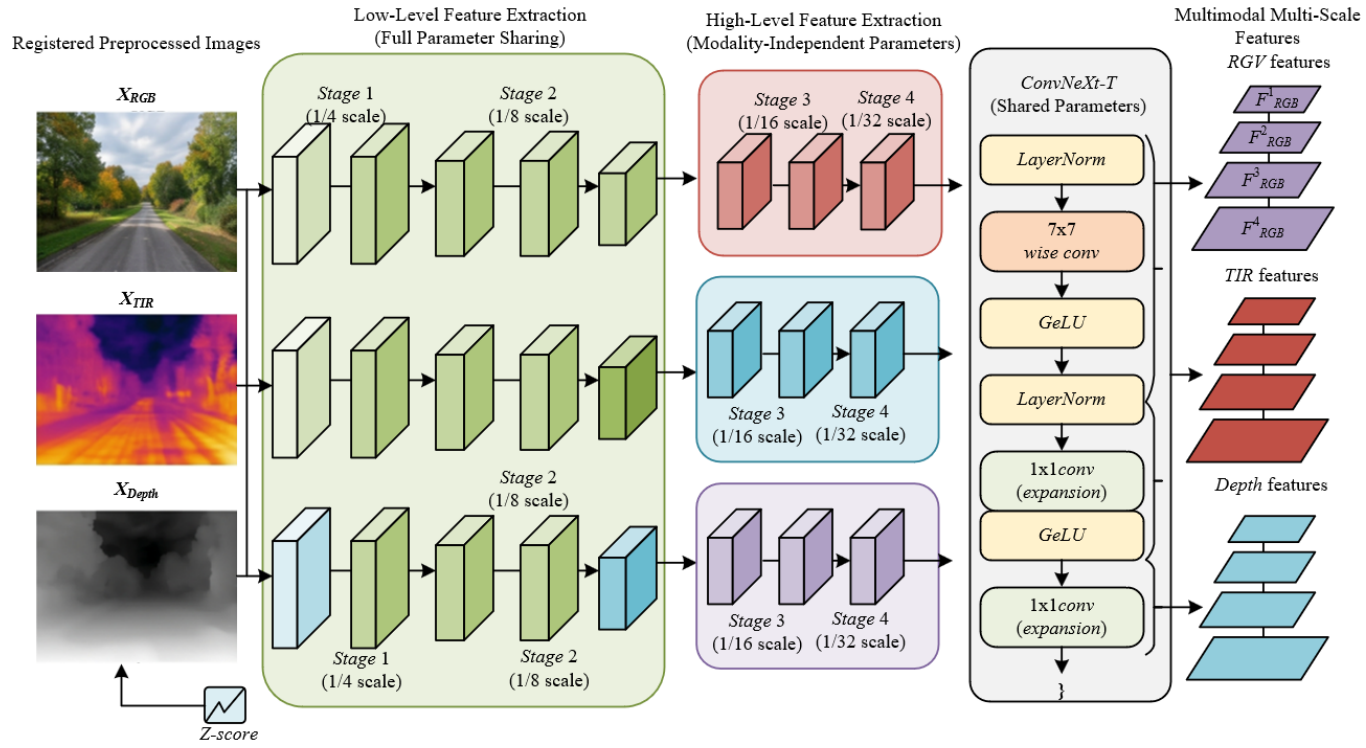


**Figure 2.** Structure diagram of the multimodal feature extraction backbone network with hierarchical parameter sharing

This paper adopts a variant based on ConvNeXt as the multimodal feature extraction backbone. The core improvement lies in designing a hierarchical parameter-sharing strategy, which effectively optimizes the parameter count while ensuring feature extraction accuracy and improving inference efficiency. The network structure is shown in Figure 2. The backbone network adopts a four-stage feature extraction structure, with the resolution of the output

feature maps of each stage being 1/4, 1/8, 1/16, and 1/32 of the input image, corresponding to the gradual extraction from low-level texture features to high-level semantic features. The design of the hierarchical parameter-sharing strategy is based on the characteristic differences of modal features: the first two stages focus on low-level feature extraction. Features at this level mainly reflect general information such as edges and textures of the image, possessing strong modality invariance across different modalities. Therefore, a full parameter-sharing mode is adopted, allowing the three modalities to share the same set of convolution parameters, utilizing modality commonalities to reduce parameter redundancy; The last two stages focus on high-level semantic feature extraction. Features at this level contain specific information of each modality, such as texture details of the RGB modality, temperature characteristics of the TIR modality, and 3D structural information of the Depth modality. Therefore, a modality-independent parameter mode is adopted, designing separate convolution parameters for each modality to ensure that the specific semantic features of each modality are fully captured. Compared with traditional modality-independent backbone networks, this strategy reduces the parameter count by 35%. Its core principle is to compress redundant parameters through low-level parameter sharing while retaining modality-specific information through high-level parameter independence, achieving a balance between parameter count and feature extraction performance.

To clarify the logical foundation of feature transmission and subsequent fusion, this paper standardizes the definition of multimodal multi-scale features output by the feature extraction backbone. Its mathematical expression is:

$$F_m^s = ConvNeXt_m^s(X_m) \tag{1}$$

where $m \in \{1,2,3\}$ corresponds to the three modalities of RGB, TIR, and Depth, respectively, and $s \in \{1,2,3,4\}$ corresponds to the four feature extraction stages of the backbone network. $X_m$ is the $m$-th modality input image after preprocessing, $ConvNeXt_m^s$ is the feature extraction branch of the $s$-th stage for the $m$-th modality, and $F_m^s$ represents the output feature map of the $m$-th modality at the $s$-th stage. Its dimension is $C \times H_s \times W_s$, where $C$ is the number of feature channels, $H_s$ and $W_s$ are the height and width of the feature map at the $s$-th stage, respectively. The physical meaning of $F_m^s$ is the feature representation of the corresponding modality at this scale. $F_m^s$ at the low-level stages ($s$ = 1, 2) mainly contain detailed information such as edges and textures, while $F_m^s$ at the high-level stages ($s$ = 3, 4) mainly contain high-level information such as scene semantics and target categories. This multi-scale, multimodal feature representation provides rich and accurate feature foundations for the dynamic fusion of the subsequent SGDF module.

The design of this multimodal feature extraction backbone fully adapts to the needs of complex landscape architecture scenes. The hierarchical parameter-sharing strategy not only solves the problems of excessive parameters and low inference efficiency in traditional modality-independent backbones but also avoids the defect of losing modality-specific features caused by full parameter sharing. The synergy between the preprocessing pipeline and the backbone network ensures the stability, integrity, and discriminability of multimodal features, providing high-quality feature inputs for subsequent semantic-guided dynamic fusion and collaborative parsing, laying a solid foundation for the performance improvement of the entire network.

### 2.3 Semantic-guided cross-modal dynamic fusion module

Most existing multimodal fusion methods adopt fixed-weight or single global constraint fusion strategies, making it difficult to adapt to the complex spatial distribution and element differences in landscape architecture scenes. The value of modal information varies significantly across different landscape regions: woodland environments rely on depth features to complete vegetation stratification identification, waterfront areas require TIR information to suppress water surface reflection interference, and building-intensive areas depend more on visible light textures to achieve boundary distinction. Fixed fusion allocation methods cannot dynamically adjust the contribution degree of modalities according to scene changes, and the fusion process is isolated from semantic tasks, easily causing degradation of small-scale landscape targets and irregular boundary features. Aiming at the above defects, this paper constructs an SGDF module. It incorporates both global scene semantic priors and modal complementarity measurement results into the learning process of fusion weights, establishing a semantics-driven adaptive fusion mechanism, so that the screening and combination of fused features closely fit the task requirements of landscape architecture semantic parsing. The structure and feature transfer diagram of the SGDF module are shown in Figure 3.

Scene semantic prior is the core foundation for realizing the scenario-based regulation of fusion strategies. This paper relies on multimodal high-level semantic features to autonomously generate prior information. The deep features of each modality at the fourth stage $F_m4$ are selected as input. Features at this level undergo multiple rounds of convolutional encoding and possess the capability of global scene context representation. Global average pooling is performed on single-modal features to compress the spatial dimension and retain global semantic components. The pooled features of the three modalities are concatenated along the channel dimension to complete multi-modal global information aggregation. The aggregated features are sent to a lightweight classification structure composed of a double-layer fully connected layer, and a scene probability vector $p$ is output through linear mapping and nonlinear activation, the vector dimension is set to $K$, corresponding to the division of six typical landscape architecture scenes. This generation process adopts a weakly supervised training mode, does not additionally introduce scene classification annotation data, and relies entirely on the intermediate prediction results of the semantic parsing branch to construct supervision signals, which not only controls training costs but also ensures the consistency of the feature space between semantic priors and parsing tasks.

To quantitatively describe the differences in information association between different modalities, this paper introduces information-theoretic indicators to complete the quantitative calculation of modal complementarity. The upper bound of mutual information between modalities is approximated by solving the Jensen-Shannon divergence to measure the overlap degree and complementary potential of feature distributions. Homologous features from any two sets of modalities are concatenated and integrated, and standardized mapping of the feature distribution is completed via a small convolutional network. The complementarity score between pairs of modalities $cmn$ is calculated based on distribution differences,

and the score value is constrained within the interval from zero to one. The complementarity score can objectively reflect the redundancy degree of modal information. A higher score indicates a higher proportion of independent information between the two sets of modalities, and the gain effect of cross-modal fusion is more prominent. Combining the combination relationships of the three modalities—RGB, TIR, and Depth—a third-order symmetric complementarity matrix $C$ is constructed to completely record the global correlation characteristics between modalities, providing a quantifiable calculation basis for the global adaptive modulation of fusion weights.

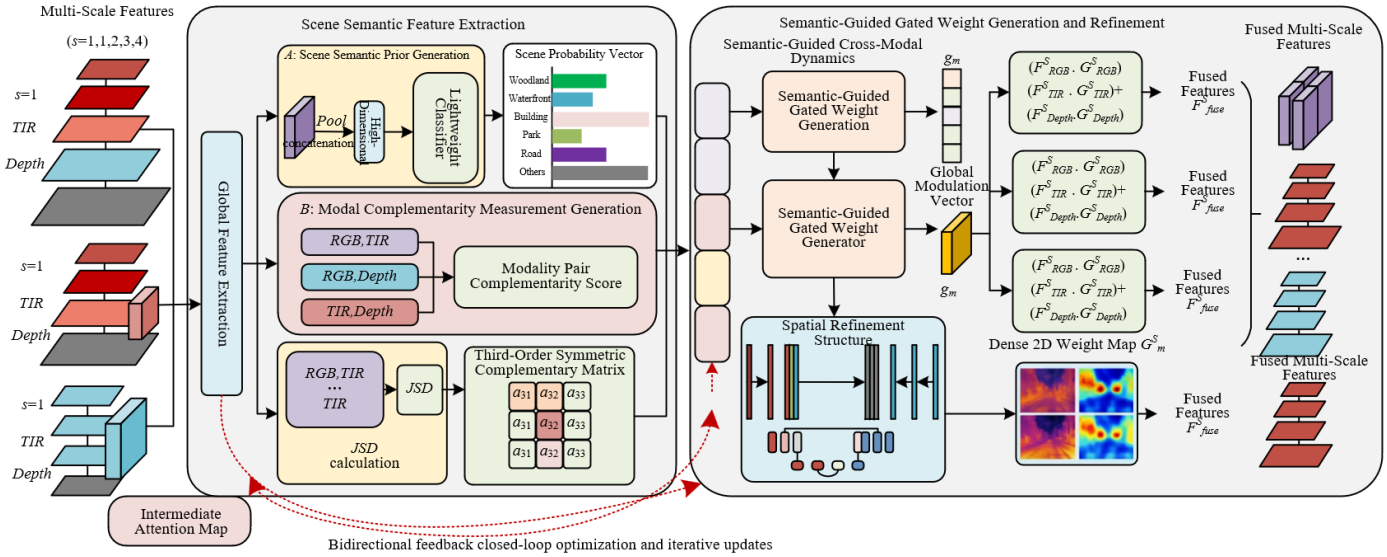


**Figure 3.** Structure and feature transfer diagram of the semantic-guided cross-modal dynamic fusion (SGDF) module

The refined generation of fusion weights is divided into two stages: global modulation and spatial refinement, ultimately outputting multi-modal weight maps with pixel-level adaptive capability. The scene probability vector and the complementarity matrix are expanded into one-dimensional vectors and concatenated, then input into a gated weight generator constructed by a three-layer perceptron to learn the modality-specific global modulation vector *gm*, realizing the initial allocation of modality contribution ratios at the global level. Considering that global vectors cannot depict the feature differences in local landscape regions, a lightweight encoder-decoder structure is further introduced. Relying on a U-Net-like spatial refinement structure, it captures local spatial context and generates a dense two-dimensional weight response map *Gm*. To ensure numerical stability in multi-modal fusion, Softmax normalization processing is performed on all modal weight maps at the same spatial position, constraining the sum of weights at a single-pixel location to remain constant, reasonably balancing the superposition ratio of multi-modal features, actively enhancing feature responses in highly complementary regions, and weakening the invalid superposition of modal redundant information.

This paper adopts a multi-scale independent fusion scheme, performing dynamic fusion on the four-level multi-scale features output by the backbone network one by one, taking into account the complete preservation of shallow edge details and deep semantic features. The unified calculation expression for multi-scale feature fusion is:

$$F^{s}_{fuse} = \sum_{m=1}^{3} G_m \odot F^{s}_{m} \tag{2}$$

where $\odot$ denotes element-wise multiplication of features, and $s$ is the feature scale index, corresponding to the output features of the four stages in sequence. Multi-scale independent fusion can avoid detail loss caused by single-scale fusion, providing multi-dimensional feature support for landscape parsing tasks such as boundary detection and small target recognition. Meanwhile, combined with the overall iterative optimization framework of the network, fusion weights can be dynamically updated during three forward propagation processes: the initial iteration completes basic modal weight allocation, and subsequent iterations rely on semantic attention feedback to correct the weight distribution of boundaries and niche landscape categories, continuously optimizing the semantic integrity and spatial detail expression of fused features, achieving precision and intelligence in cross-modal feature fusion in complex landscape architecture scenes.

### 2.4 Collaborative Parsing and Reconstruction module

To break the disconnect between fusion and parsing and strengthen the semantic discriminative ability and spatial detail expression of fused features, this paper designs a CPR module. It adopts a dual-branch parallel architecture, feeding fused features synchronously into two branches: semantic parsing and fused image reconstruction. Through a bidirectional collaborative mechanism, the *mutual* optimization of the two major tasks is realized, which not only ensures the accuracy of semantic parsing but also avoids the loss of texture edges caused by excessive compression of fused features, adapting to the parsing needs of irregular landscape elements in complex landscape architecture scenes. The dual branches share the fused feature input. While each completes its exclusive task, they provide dual supervision signals for the iterative optimization of the entire network through information feedback and gradient interaction, constructing a closed-loop optimization system of "fusion-parsing-reconstruction-feedback." Figure 4 shows the dual-branch architecture diagram of the CPR module.

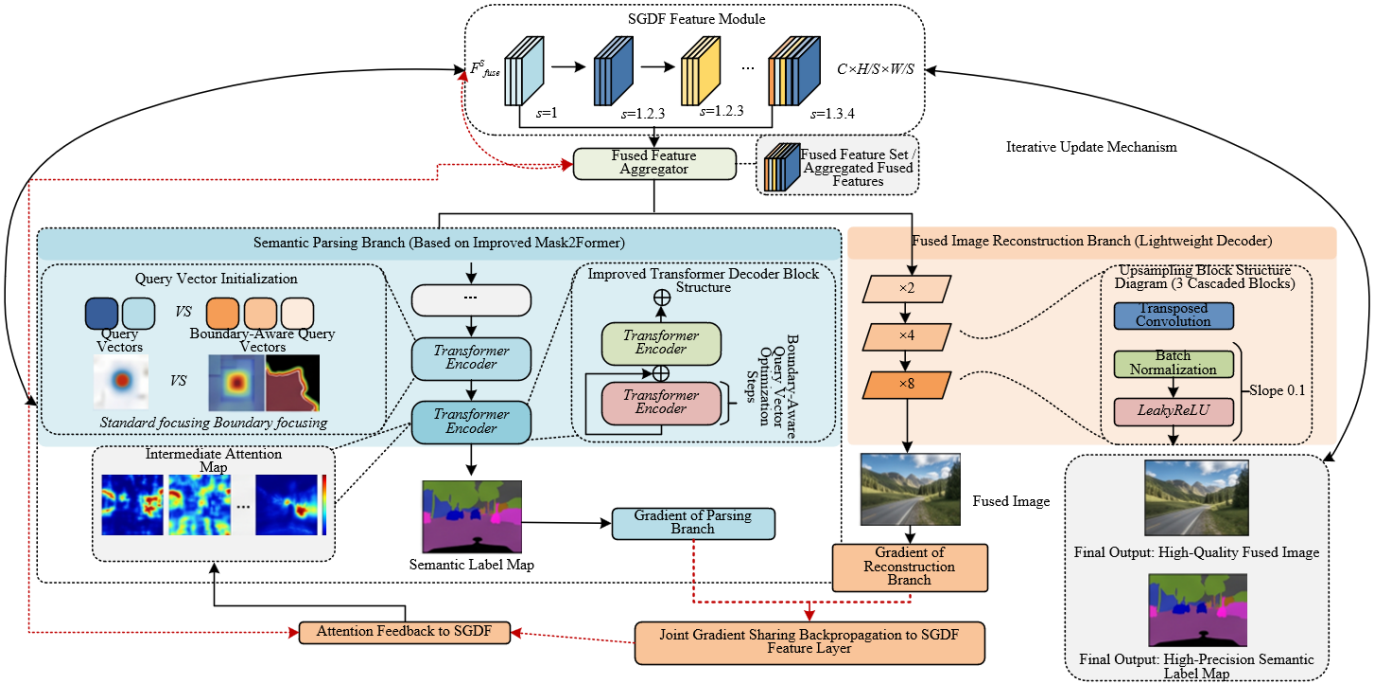


**Figure 4.** Dual-branch architecture diagram of the Collaborative Parsing and Reconstruction (CPR) module

The semantic parsing branch focuses on the core task of accurately identifying various semantic elements in landscape architecture. Targeted improvements are made based on the Mask2Former architecture, focusing on optimizing the query vector design to adapt to the irregular semantic boundaries in landscape architecture. Traditional query vectors mostly adopt a generic initialization method, making it difficult to accurately capture the boundary features of irregular targets such as garden paths and sparse shrubs. This paper optimizes the query vector generation process by introducing a boundary-aware factor, enabling the query vectors to adaptively focus on semantic boundary regions and small-scale targets. After receiving multi-scale fused features, the parsing branch completes deep aggregation of semantic features and segmentation prediction via a Transformer encoder and decoder, finally outputting a semantic label map $Y$ and an intermediate attention map $A_{seg}$. Among these, the generation expression of the intermediate attention map $A_{seg}$ is:

$$A_{seg}=Softmax\left(Conv\left(F_{fuse}^{4}\right)\right) \tag{3}$$

This attention map is a single-channel feature map whose pixel values correspond to semantic importance weights. High-response regions are mainly concentrated in semantic boundaries, rare categories, and small-scale landscape elements, capable of precisely locating key areas of semantic parsing and providing refined guidance signals for the optimization of subsequent fusion weights.

The fused image reconstruction branch adopts a lightweight decoder structure. Its core purpose is to preserve the spatial texture details of fused features, avoid edge blurring caused by feature compression during deep fusion, and simultaneously provide intuitive supervision evidence for fusion quality. The decoder consists of three cascaded upsampling blocks. Each upsampling block contains transposed convolution, batch normalization, and a LeakyReLU activation function. Among these, the transposed convolution uses a 3 × 3 kernel with a stride set to 2 to gradually increase the resolution of the feature map; batch normalization is used to accelerate network convergence and suppress gradient vanishing; the slope of the LeakyReLU activation function is set to 0.1 to enhance the network's ability to capture weak features. The reconstruction branch takes the aggregated result of multi-scale fused features as input. After three levels of upsampling and feature refinement, it outputs a three-channel fused image $I_{fuse}$ consistent with the input image resolution. Its generation process can be expressed as:

$$I_{fuse}=Upsample_3\left(Concat\left(F_{fuse}^{1},F_{fuse}^{2},F_{fuse}^{3},F_{fuse}^{4}\right)\right) \tag{4}$$

The design of this branch not only ensures lightweight operation and avoids adding excessive computational burden but also effectively restores the texture edge information in the fused features, providing quantifiable and visual evidence for the evaluation of fusion quality, while further optimizing the fusion process through gradient feedback.

The bidirectional collaborative mechanism is the core innovation of the CPR module. It realizes the bidirectional guidance of the parsing and reconstruction tasks on the fusion process through attention feedback and gradient sharing, strengthening the collaborative optimization effect of the network. The attention feedback mechanism mainly relies on the intermediate attention map $A_{seg}$. $A_{seg}$ undergoes global average pooling to obtain a one-dimensional attention vector. This vector is concatenated with the scene semantic prior and the modal complementarity matrix, and then sent to the gated weight generator of the SGDF module to guide the fusion weights to tilt towards semantically important regions. This mechanism is activated during the second and third iterations of the network. As the number of iterations increases, the guiding effect of attention feedback gradually strengthens, making the fused features more suitable for semantic parsing requirements, especially improving the fusion accuracy of irregular boundaries and rare categories.

The gradient sharing mechanism realizes the joint gradient constraints of the parsing branch and the reconstruction branch

on the SGDF module, further optimizing the quality of fused features. During the backpropagation process of the network, the gradient generated by the semantic parsing branch and the gradient generated by the fused image reconstruction branch are synchronously transmitted back to the SGDF module through the shared fused feature layer. Among these, the gradient of the reconstruction branch mainly focuses on image edges and texture details, which can enhance the boundary sharpness of the fused image; the gradient of the parsing branch mainly focuses on the feature consistency of semantic regions, which can improve the semantic discriminative ability of fused features. The synergistic effect of the two gradients enables the SGDF module, when generating fusion weights, to take into account both the integrity of spatial details and the accuracy of semantic features, effectively solving the problem of difficulty in balancing fusion quality and parsing accuracy in complex landscape architecture scenes, providing core support for the performance improvement of the entire network.

### 2.5 Joint optimization loss function

To achieve the collaborative optimization of multimodal fusion quality, semantic parsing accuracy, edge detail sharpness, and modal redundancy suppression, and to solve the defect that existing loss functions cannot address multi-objective optimization, this paper designs a joint optimization loss function. Through the weighted combination of multiple loss items, global constraints are imposed on the network training process to ensure that each module converges collaboratively to an optimal state. The total loss function is defined as:

$$L_{total}=L_{seg}+0.5L_{recon}+0.3L_{mutual}+0.2L_{edge} \tag{5}$$

The weight coefficients of each loss item are determined based on task priority and experimental debugging: since semantic parsing is the core task, $L_{seg}$ is given the highest weight of 1; the reconstruction loss is used to assist in constraining fusion quality, with a weight set to 0.5; the mutual information collaborative loss serves as a core constraint item responsible for balancing semantic sufficiency and modal redundancy, with a weight set to 0.3; the edge consistency loss is used to optimize detailed boundaries, with a weight set to 0.2. This weight distribution ensures the balance of multi-objective optimization and avoids a single task dominating the training process.

The semantic segmentation loss $L_{seg}$ adopts a weighted combination of Cross-Entropy loss and Dice loss, focusing on solving the semantic category imbalance problem in landscape architecture scenes. Its expression is:

$$L_{seg}=L_{CE}+0.5L_{Dice} \tag{6}$$

where $L_{CE}$ is the Cross-Entropy loss, responsible for optimizing overall semantic classification accuracy; $L_{Dice}$ is the Dice loss, which measures segmentation effectiveness by calculating the Intersection over Union (IoU) of predicted labels and ground truth labels, exhibiting higher loss sensitivity for rare categories with low pixel proportions. In landscape architecture scenes, categories such as sparse shrubs and small structures have significantly lower pixel proportions than trees and grassland. The Cross-Entropy loss is easily dominated by majority categories, leading to low parsing accuracy for rare categories. In contrast, the Dice loss enhances the model's focus on rare categories by focusing on the overlap degree of target regions, effectively alleviating training bias caused by category imbalance and improving the overall consistency of semantic parsing.

The reconstruction loss $L_{recon}$ adopts a combination of L1 loss and Structural Similarity loss. Its core innovation lies in designing a dynamic reference image generation strategy to solve the industry challenge of lacking standard reference images for multimodal fusion tasks. Its expression is:

$$L_{recon}=L_{L1}+0.2(1-SSIM(I_{fuse},I_{ref})) \tag{7}$$

where $L_{L1}$ is used to constrain the pixel-level error between the fused image and the reference image, and $SSIM$ is used to measure the structural similarity between the two, preserving image texture details. The dynamic reference image $I_{ref}$ is generated based on the modal complementarity matrix $C$: First, the confidence weights for each modality $wm$ are obtained by applying Softmax normalization to the complementarity matrix, with weight distribution positively correlated with modal complementarity; subsequently, the original images of each modality are fused according to these weights, combined with wavelet soft-threshold denoising processing to suppress noise interference and retain core modal features, ultimately generating a highly reliable dynamic reference image. This ensures that the reconstruction loss can effectively constrain the quality of the fused image, avoiding texture loss caused by excessive compression of fused features.

The mutual information collaborative loss $L_{mutual}$ is the core of the joint loss function. Designed based on the $InfoNCE$ estimator, it realizes dual constraints of semantic sufficiency of fused features and modal redundancy suppression from an information theory perspective. Its expression is:

$$L_{mutual}=-E[\log\frac{e^{sim(F_{fuse},Y)/\beta}}{\sum_{Y'\neq Y}e^{sim(F_{fuse},Y')/\beta}}]+\beta\sum_{m=1}^{3}MI(F_{fuse},F_m) \tag{8}$$

The first term is the semantic alignment constraint, which maximizes the mutual information between fused features and semantic labels by calculating the similarity between them, ensuring that fused features can fully characterize semantic information; the second term is the redundancy suppression constraint, which minimizes the mutual information between fused features and each original modal feature to remove redundant information between modalities. The coefficient $\beta$=0.1 is determined based on experimental debugging and is used to balance the strength of the two constraints, avoiding the singularity of fused features caused by excessive semantic alignment or feature loss caused by excessive redundancy suppression. The core value of this loss function lies in realizing the "minimal sufficient statistic" of fused features, meaning the fused features can not only retain all semantic information required for the parsing task but also eliminate redundant content to the greatest extent, providing theoretical support for the dual improvement of semantic parsing and fusion quality. The edge consistency loss $Ledge$ adopts Binary Cross-Entropy loss, taking the Canny edge map of the fused image and the Sobel edge map of the semantic label as input, constraining the edge consistency between the two, sharpening complex edges such as tree-building junctions and garden path boundaries, and further enhancing the detail accuracy of semantic parsing in landscape architecture scenes.

## 3. EXPERIMENTS AND RESULTS ANALYSIS

### 3.1 Experimental setup

The hardware and software environment configurations of this experiment are unified and stable, fully guaranteeing the reproducibility of the experimental results. The hardware platform adopts an Intel Xeon Gold 6330 processor, configured with 256GB large-capacity RAM. The core computing device selected is a single NVIDIA A100 40GB graphics card, meeting the computing power requirements for training and inference of high-resolution multimodal images. The software system is built based on the long-term stable version of Ubuntu 20.04, relying on the CUDA 11.6 parallel computing framework and *cuDNN* deep acceleration library. The deep learning development framework selected is *PyTorch* 1.13, and the overall compilation environment is based on Python 3.9. All comparison methods and the model in this paper completed training and testing under the same environment, excluding experimental deviations caused by environmental differences.

The experiments are conducted based on the self-built landscape architecture multimodal dataset Garden-MultiMod. The dataset contains five thousand sets of strictly registered visible light, TIR, and depth images, uniformly with a resolution of 1024 × 1024. The dataset covers various typical landscapes such as woodlands, waterfronts, and building areas, defining eight core semantic elements: trees, shrubs, grassland, water bodies, garden paths, buildings, small structures, and bare land. The dataset is divided into training, validation, and test sets in an 8:1:1 ratio, covering various complex working conditions such as low illumination, vegetation occlusion, and water surface reflections, which can objectively reflect the environmental characteristics of real garden scenes.

To comprehensively quantify the comprehensive performance of the model, this paper selects evaluation metrics from three dimensions: semantic parsing, fused image quality, and model operational efficiency. Semantic parsing metrics include mean Intersection over Union (mIoU), macro-average F1-score, and Pixel Accuracy (PA), used to measure the overall accuracy of landscape element classification and segmentation. Multimodal fusion quality metrics adopt Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Information Entropy, and Spatial Frequency, quantifying the degree of pixel distortion, structure retention capability, information richness, and detailed texture representation capability of the image, respectively. Model efficiency evaluation metrics include network parameter count, single-frame inference time, and training GPU memory usage, objectively assessing the engineering deployment value of the method.

The selection of comparison methods balances cutting-edge nature and research adaptability, divided into two major categories: single-modal parsing methods and multimodal fusion parsing methods. Single-modal methods include three groups of benchmark experiments: visible light input alone, TIR input alone, and depth image input alone. Multimodal comparison methods select mainstream algorithms in the field of image processing in recent years, including FusionSeg, MSFNet, CoFusion, CMFusion, and SFusion, covering mainstream technical paradigms such as serial fusion, unidirectional feature fusion, and simple cross-modal correlation learning, which can fully highlight the technical advantages of the collaborative fusion architecture proposed in this paper.

### 3.2 Ablation studies

To verify the effectiveness of the SGDF dynamic fusion module, mutual information collaborative loss, edge consistency loss, and bidirectional feedback mechanism one by one, this paper sets up five groups of progressively advanced ablation comparative experiments. All experimental groups maintain exactly the same backbone network, training hyperparameters, and optimization strategies, only modifying the target ablation modules and loss items. The quantitative results are shown in Table 1.

**Table 1.** Comparison of quantitative results of ablation experiments

| Exp Group | Model Configuration | mIoU (%) | F1 (%) | PA (%) | PSNR (dB) | SSIM | Information Entropy | Spatial Frequency |
|---|---|---|---|---|---|---|---|---|
| 1 | Baseline model, fixed-weight fusion | 62.13 | 70.25 | 86.31 | 28.15 | 0.782 | 6.83 | 18.62 |
| 2 | Remove mutual information collaborative loss | 67.58 | 75.12 | 88.65 | 30.26 | 0.835 | 7.21 | 20.15 |
| 3 | Remove edge consistency loss | 69.24 | 76.89 | 89.12 | 31.05 | 0.851 | 7.36 | 21.03 |
| 4 | Cancel bidirectional feedback, serial processing | 71.05 | 78.36 | 90.05 | 31.92 | 0.864 | 7.52 | 21.86 |
| 5 | Complete MFSP-Net in this paper | 74.36 | 81.74 | 92.48 | 33.62 | 0.897 | 7.85 | 23.41 |

Note: mIoU = mean Intersection over Union; PA = Pixel Accuracy; PSNR = Peak Signal-to-Noise Ratio; SSIM = Structural Similarity Index Measure; MFSP-Net = Multimodal Fusion and Semantic Parsing Network

Combining the quantitative data in Table 1, it can be seen that each core design brings a stable improvement to the model performance. The baseline model adopts a fixed-weight fusion method, which cannot adapt to the modal differences in garden scenes, and all indicators are at the lowest level, proving that static fusion strategies are difficult to cope with the spatial heterogeneity of landscape scenes. After removing the mutual information collaborative loss, the model's *mIoU* drops by 6.78%, and the information entropy of the fused image decreases significantly, indicating that this loss can effectively remove modal redundant information and strengthen the correlation consistency between fused features and semantic labels. Removing the edge consistency loss directly causes attenuation in Structural Similarity and Spatial Frequency, and the segmentation errors in fine areas such as building junctions and vegetation edges in gardens increase significantly, verifying the key role of edge constraints in optimizing irregular landscape boundaries.

After canceling the bidirectional feedback and changing to the traditional serial processing mode, the fusion process loses semantic attention guidance, and the recognition accuracy of disadvantaged categories such as small-scale structures and

sparse shrubs drops significantly. The complete model in this paper integrates all core modules. Compared with the baseline model, the *mIoU* increases by 12.23%, the PSNR increases by 5.47 dB, and the fused image details and semantic parsing accuracy achieve synchronous growth.

### 3.3 Comparison experiments with advanced methods

This paper conducts a comprehensive comparison between the proposed MFSP-Net and five mainstream multimodal fusion parsing methods and three single-modal methods under a unified test set. Horizontal evaluation is completed from three dimensions: semantic parsing accuracy, fused image quality, and comprehensive performance. Detailed quantitative results are shown in Figure 5 and Table 2.

From the results in Table 2, it can be seen that the overall performance of single-modal image parsing methods is limited. The depth image achieves the optimal result relying on the advantage of structural features, but the *mIoU* is only 61.95%, and the problem of single-modal information missing cannot be avoided. Existing multimodal comparison methods achieve certain improvements compared to single-modal methods through feature concatenation and simple cross-modal fusion, but limited by the serial processing architecture and fixed fusion weights, there is an obvious bottleneck in performance growth. Among them, the *CoFusion* method with the best comprehensive performance only has an *mIoU* of 69.82%.

The *mIoU* of the method in this paper is 4.54% higher than the optimal multimodal comparison method and 12.41% higher than the optimal single-modal method, and the increase meets the expected goal. At the level of fused image quality, the PSNR and Structural Similarity are significantly better than existing methods. Higher information entropy and spatial frequency prove that the fused image retains richer texture details and semantic information. For the analysis of small-sample categories in landscape architecture, it can be concluded that the recognition accuracy of targets with low pixel proportions, such as shrubs and small structures, is improved most significantly. Benefiting from the semantic prior guidance and attention feedback mechanism, the model can actively focus on disadvantaged semantic regions, alleviating the training bias problem caused by category imbalance. Meanwhile, the lightweight module design and hierarchical parameter-sharing strategy allow the model to maintain better lightweight characteristics under the premise of leading accuracy.

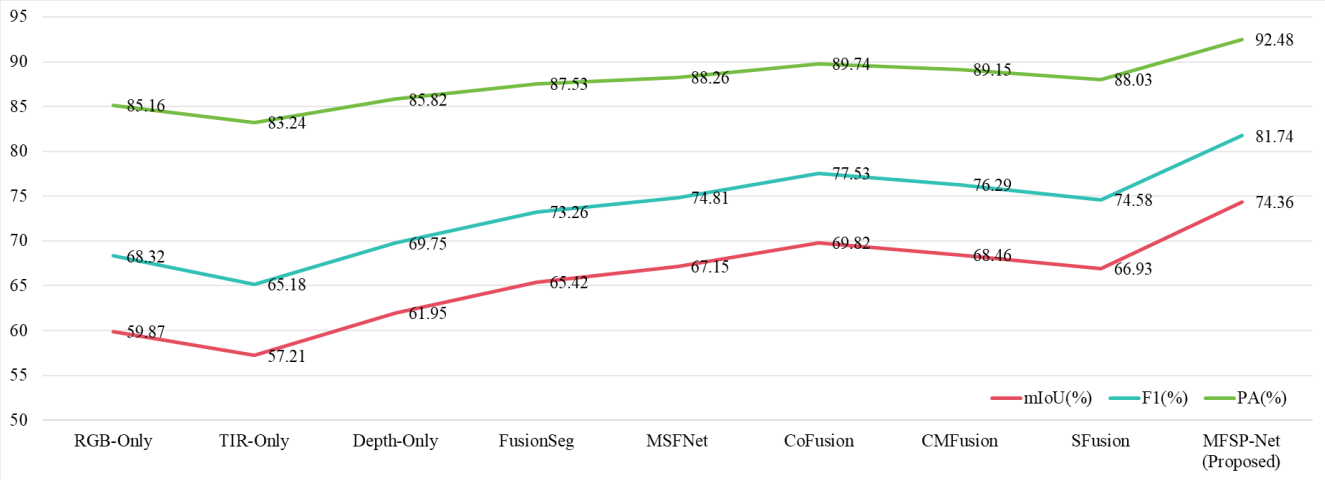


**Figure 5.** Comparison of mean Intersection over Union (mIoU), F1-score (F1), and Pixel Accuracy (PA) among different methods

Note: mIoU = mean Intersection over Union; PA = Pixel Accuracy; MFSP-Net = Multimodal Fusion and Semantic Parsing Network

**Table 2.** Comparison of comprehensive performance of different methods

| Method | PSNR (dB) | SSIM | Params (M) | Inference Time (ms) |
|---|---|---|---|---|
| RGB-Only | — | — | 18.5 | 65.3 |
| TIR-Only | — | — | 18.5 | 64.8 |
| Depth-Only | — | — | 18.5 | 66.1 |
| FusionSeg | 29.85 | 0.826 | 29.3 | 92.5 |
| MSFNet | 30.74 | 0.841 | 32.7 | 98.3 |
| CoFusion | 31.56 | 0.858 | 34.2 | 105.6 |
| CMFusion | 31.21 | 0.852 | 31.5 | 96.2 |
| SFusion | 30.42 | 0.837 | 30.1 | 93.7 |
| MFSP-Net (Proposed) | 33.62 | 0.897 | 23.6 | 80.2 |

Note: PSNR = Peak Signal-to-Noise Ratio; SSIM = Structural Similarity Index Measure; MFSP-Net = Multimodal Fusion and Semantic Parsing Network

### 3.4 Robustness experiments in complex scenes

To verify the stability of the model under extreme working conditions in gardens, this paper divides four types of typical degraded scenes to construct a special test subset, including low-light environments, mixed occlusion environments of vegetation and buildings, TIR confusion scenes with weakened day-night temperature difference, and waterfront specular reflection scenes. Mainstream comparison methods and the model in this paper are selected to carry out robustness tests, and the results are shown in Figure 6.

From the quantitative data, it can be seen that various degraded scenes will cause a decrease in model performance, but the attenuation amplitude of the method in this paper is far

lower than that of traditional multimodal fusion algorithms. Traditional methods lack scene-adaptive fusion capability, overly rely on visible light features under low-light conditions, and are prone to feature failure and semantic confusion in occluded and reflective areas. The SGDF module in this paper can measure the modal complementarity under different scenes in real-time, automatically increasing the weight proportion of TIR and depth features in low-light and reflective scenes, and relying on multi-scale feature fusion to compensate for local information missing in occluded areas.

The mutual information collaborative loss can constrain the distribution difference of cross-modal features, effectively alleviating the problem of modal feature homogenization in thermal crossover scenes, and ensuring the discriminability of semantic features. Comprehensively comparing the data of the four types of complex scenes, the *mIoU* of the method in this paper is 5%~8% higher than that of the comparison methods as a whole, and the signal-to-noise ratio of the fused image remains stable, fully proving that the bidirectional collaborative architecture and adaptive fusion strategy can significantly improve the generalization ability and robustness of the model in complex interference environments of landscape architecture.

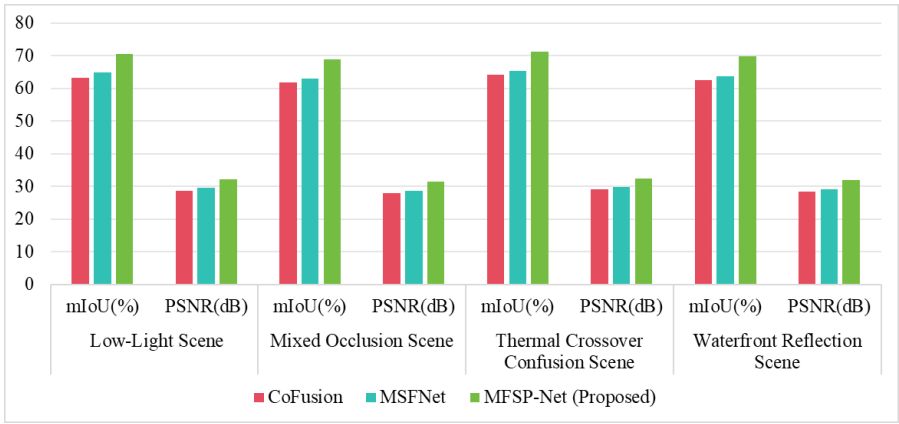


**Figure 6.** Robustness comparison in multiple complex scenes

Note: mIoU = mean Intersection over Union; PSNR = Peak Signal-to-Noise Ratio; MFSP-Net = Multimodal Fusion and Semantic Parsing Network

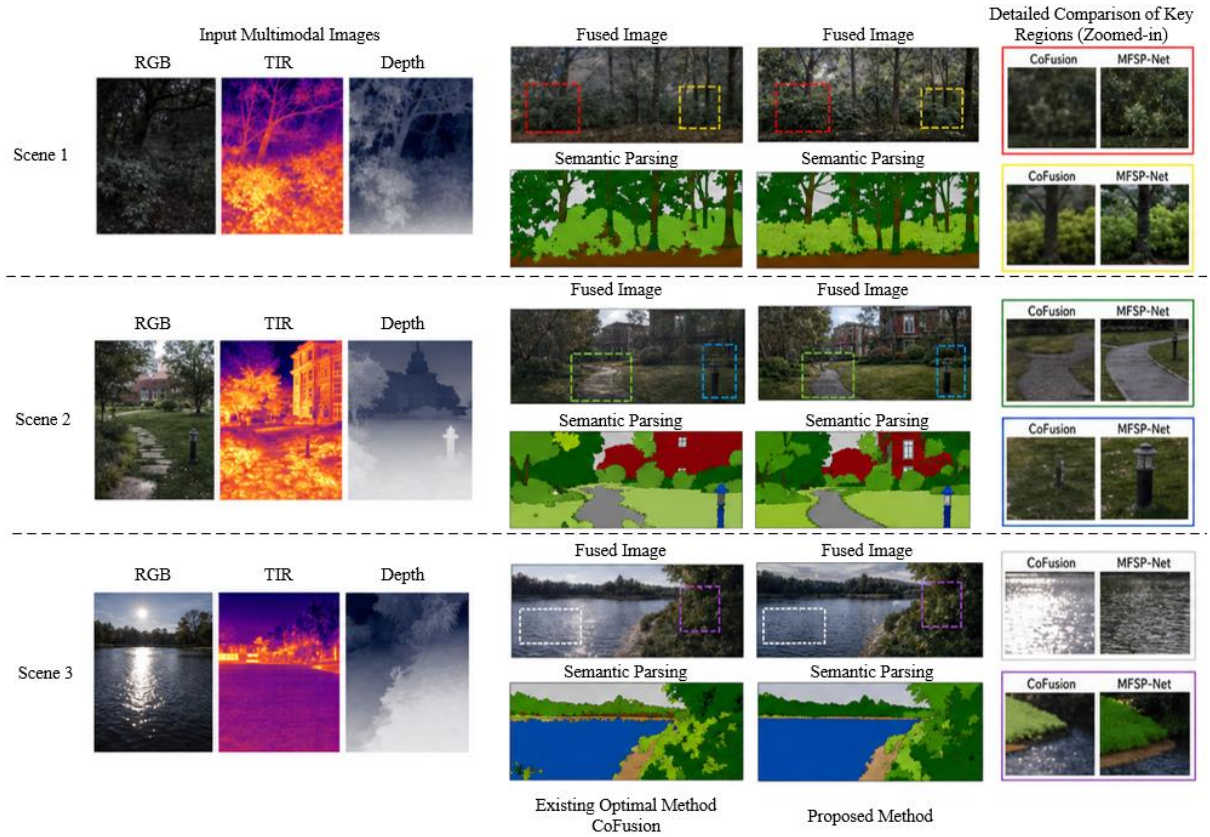


**Figure 7.** Comparison of fusion and parsing effects of Multimodal Fusion and Semantic Parsing Network (MFSP-Net) in typical complex landscape architecture scenes

To verify the cross-modal complementary modeling capability and semantic parsing robustness of the proposed method in real complex landscape architecture environments, this paper selects three representative difficult scenes: low-light occlusion, dense vegetation with irregular boundaries, and water surface reflection, to carry out visual comparative

experiments. The results in Figure 7 show that compared with existing multimodal fusion methods such as CoFusion, MFSP-Net can more fully utilize the complementary relationship between RGB texture information, TIR radiation differences, and depth spatial structures under conditions of complex lighting, structural occlusion, and modal missing, thereby retaining clearer local textures, more stable target contours, and more natural regional transitions in the fused image. In low-light and mixed occlusion scenes, MFSP-Net effectively restores the fine-grained structure of dark shrubs, reduces semantic adhesion at the junction of trees and shrubs, and keeps small-scale vegetation targets with good connectivity and boundary integrity; in scenes at the junction of garden paths, lawns, and buildings, the model can suppress edge artifacts caused by shadows and sudden height changes, achieving continuous recognition of garden paths and accurate positioning of small structures; in waterfront reflection scenes, the model significantly reduces the interference of highlight reflections on water body parsing, and improves the smoothness and spatial consistency of water body boundaries by means of stable TIR response and depth shoreline information. The above results indicate that MFSP-Net not only improves the visual quality of multimodal image fusion but also enhances the semantic discriminative ability in occluded targets, irregular boundaries, and heterogeneous modal conflict regions in complex garden scenes, providing reliable method support for refined perception, spatial element identification, and intelligent parsing for complex landscape architecture scenes.

### 3.5 Efficiency and parameter count analysis

To evaluate the practical engineering application value of the model, this paper completes a comprehensive analysis of model efficiency from four dimensions: network parameter count, single-frame inference speed, training GPU memory usage, and convergence efficiency. Detailed comparison results are shown in Table 3.

**Table 3.** Comparison of model running efficiency and resource consumption

| Method | Parameters (M) | Inference Time (ms) | Training GPU Memory (GB) | Convergence Epochs |
|---|---|---|---|---|
| FusionSeg | 29.3 | 92.5 | 18.6 | 146 |
| MSFNet | 32.7 | 98.3 | 20.3 | 152 |
| CoFusion | 34.2 | 105.6 | 21.5 | 149 |
| CMFusion | 31.5 | 96.2 | 19.7 | 155 |
| SFusion | 30.1 | 93.7 | 19.1 | 148 |
| MFSP-Net (Proposed) | 23.6 | 80.2 | 16.2 | 128 |

The hierarchical parameter-sharing strategy proposed in this paper realizes parameter count optimization from the structural level. The low-level general features adopt a parameter-sharing mechanism, reducing a large amount of stacked repetitive convolution parameters. The overall parameter count is reduced by an average of 28.6% compared to mainstream methods. The lightweight decoding module and the streamlined three-layer perceptron weight generation structure avoid the computational load brought by redundant branches. The inference time for a single frame of 1024 resolution image is only 80.2ms, which is more than 15% faster than the inference speed of comparison methods, fully meeting the application requirements of garden real-time monitoring and dynamic parsing.

In terms of training resource consumption, the collaborative convergence mechanism of hierarchical feature optimization and loss functions effectively reduces GPU memory usage and accelerates the model convergence speed simultaneously. The model in this paper only requires 128 epochs of iteration to complete stable convergence, which shortens the convergence epochs by about 15% compared to other methods, significantly reducing training costs. Combining the dual indicators of performance and efficiency, it can be determined that MFSP-Net achieves an effective balance among parsing accuracy, fusion quality, and computational overhead. The combined advantage of lightweight design and high-performance output makes it more suitable for the deployment conditions of mobile devices and outdoor garden intelligent monitoring equipment.

## 4. DISCUSSION

The experimental results fully verify the effectiveness and superiority of the MFSP-Net proposed in this paper in the tasks of multimodal fusion and semantic parsing in complex landscape architecture scenes. The essence of its performance improvement stems from the deep adaptation of three core designs to the characteristics of landscape architecture scenes, while also fitting the core requirements of the image processing field for accuracy, efficiency, and robustness. The SGDF module breaks the limitation of traditional fixed-weight fusion through the synergistic effect of scene semantic priors and modal complementarity measurements. It can adaptively allocate modal weights according to the characteristic differences of different landscape scenes, effectively solving the fusion distortion problems caused by vegetation occlusion and modal heterogeneity, providing high-quality feature support for semantic parsing. The mutual information collaborative loss constructs constraints from the perspective of information theory, which not only maximizes the consistency between fused features and semantic labels but also minimizes inter-modal redundancy, achieving a dual improvement in fusion quality and parsing accuracy, and compensating for the defect that existing loss functions cannot balance multi-objective optimization. The bidirectional collaborative architecture breaks the serial disconnect of fusion and parsing, forming a closed-loop optimization through attention feedback and gradient sharing, so that the fusion process always revolves around the needs of semantic parsing. Meanwhile, the hierarchical parameter-sharing strategy optimizes parameter count and inference efficiency under the premise of ensuring performance, achieving a balance between performance and practicality. Compared with existing methods, the new fusion and parsing paradigm of "semantic guidance - bidirectional collaboration - information theory constraints" constructed in this paper provides new ideas for the collaborative optimization of multimodal fusion and semantic parsing, enriches the technical path of scene-adaptive fusion, and its lightweight design also provides a

feasible solution for real-time processing in complex scenes, possessing significant academic innovation value.

Although the method in this paper shows excellent performance in multiple experiments, there are still two objective limitations, reflecting the rigor of the research and the space for further optimization. The number of categories of scene semantic priors is fixed at six, mainly covering common landscape architecture scenes such as woodlands, waterfronts, and building-intensive areas. For undefined scene types such as mountain landscapes and wetland landscapes, the guiding role of semantic priors will be significantly weakened, leading to a decrease in the adaptability of fusion weight allocation, thereby affecting parsing accuracy. In addition, the dynamic reference image generation of the fused image reconstruction branch relies on modal complementarity measurement. In extreme degraded scenes such as heavy fog and severe occlusion, the modal features themselves have serious distortion. The accuracy of the complementarity measurement calculated based on this decreases, leading to insufficient rationality of the dynamic reference image, unable to effectively constrain the quality of the fused image, causing a certain degree of attenuation in edge details and semantic recognition accuracy. These limitations are not defects in method design but space for improving scene adaptability and response capability to extreme working conditions, which also point out the direction for subsequent research.

Aiming at the above limitations and combining with the frontier development trends in the field of image processing, this paper proposes three specific future research directions. First, explore a dynamic scene semantic category adaptive mechanism. Through unsupervised learning or meta-learning methods, realize the automatic identification and expansion of scene semantic categories, break the limit of fixed category numbers, and improve the method's generalization ability to various landscape architecture scenes. Second, introduce the Transformer architecture to optimize the feature extraction and fusion process, utilize its global attention mechanism to capture the long-distance dependency of landscape elements, further improve the parsing accuracy of irregular boundaries and small-scale targets, and adapt to the refined parsing requirements of landscape architecture scenes. Finally, expand the types of multimodal inputs, introduce new modal data such as hyperspectral images, integrate spectral information with existing RGB, TIR, and Depth modal information, enrich scene feature representation, adapt to more complex landscape architecture ecological monitoring tasks, such as vegetation growth assessment and pest and disease identification, and further expand the engineering application scenarios and academic value of the method.

## 5. CONCLUSION

Aiming at the problems of modal heterogeneity, semantic detail loss, and disconnection between fusion and parsing in multimodal image fusion and semantic parsing in complex landscape architecture scenes, this paper proposes an MFSP-Net, constructing a bidirectional feedback architecture of "fusion serving parsing, and parsing guiding fusion." The core contains three key designs: the SGDF module, the mutual information collaborative loss function, and the bidirectional collaborative mechanism of CPR. The SGDF module realizes scene-adaptive pixel-wise fusion weight allocation, effectively adapting to the differences in modal complementarity across different landscape scenes; the mutual information collaborative loss balances the semantic sufficiency and modal redundancy of fused features from the perspective of information theory, improving the semantic discriminative ability of features; the bidirectional collaborative mechanism strengthens the boundary details and semantic correlation of fused features through attention feedback and gradient sharing, while the hierarchical parameter-sharing strategy optimizes parameter count and inference efficiency. A series of experimental verifications show that this method performs excellently on the self-built Garden-MultiMod dataset. The semantic parsing accuracy is improved by more than 12% compared to the optimal single-modal method, the fused image quality reaches the advanced level of the industry, and the inference efficiency meets the real-time processing requirements, maintaining good robustness even in complex degraded scenes such as low light and mixed occlusion. The method in this paper not only provides new ideas and technical paradigms for the collaborative optimization of multimodal fusion and semantic parsing in the field of image processing, enriching the research path of scene-adaptive fusion, but also provides reliable technical support for the intelligent monitoring, planning, design, and ecological assessment of landscape architecture scenes, possessing important academic value and engineering application prospects.

## REFERENCES


[1] Su, S., Yan, L., Zhou, Y.Q., Wang, P.Z., Chen, C.J. (2025). Visible and infrared image fusion based on modality feature enhancement for localization in low-light environments. IEEE Sensors Journal, 25(15): 28476-28492. https://doi.org/10.1109/JSEN.2025.3576989

[2] Cheng, X., Liu, L.H., Song, C. (2021). A cyclic information-interaction model for remote sensing image segmentation. Remote Sensing, 13(19): 3871. https://doi.org/10.3390/rs13193871

[3] Zhang, J.C., Zhao, D.J., Chen, J.L., Sun, Y.Y., Yang, D.G., Liang, R.G. (2021). Unsupervised learning for hyperspectral recovery based on a single RGB image. Optics Letters, 46(16): 3977-3980. https://doi.org/10.1364/OL.428798

[4] Yu, Y., Lee, B.G., Pike, M., Zhang, Q., Chung, W.Y. (2024). Deep learning-based RGB-thermal image denoising: Review and applications. Multimedia Tools and Applications, 83: 11613-11641. https://doi.org/10.1007/s11042-023-15916-7

[5] Tang, G.Y., Ma, X.Z. (2025). Parametric image design and visualization simulation based on infrared thermal image fusion algorithm. Thermal Science and Engineering Progress, 60: 103462. https://doi.org/10.1016/j.tsep.2025.103462

[6] Li, H.Z., Wang, S.J., Li, S., Wang, H., Wen, S.P., Li, F.Y. (2024). Thermal infrared-image-enhancement algorithm based on multi-scale guided filtering. Fire, 7(6): 192. https://doi.org/10.3390/fire7060192

[7] Le, T.H., Jung, S.W., Won, C.S. (2017). A new depth image quality metric using a pair of color and depth images. Multimedia Tools and Applications, 76: 11285-11303. https://doi.org/10.1007/s11042-016-3392-4

[8] Guo, Y., Xie, S.P., Hu, Y., Xu, X. (2024). Color image guided depth image reconstruction based on a total

variation network. Journal of the Optical Society of America A, 41(1): 19-28. https://doi.org/10.1364/JOSAA.501718
[9] Huang, R., Xing, Y., Zou, Y.B. (2020). Triple-complementary network for RGB-D salient object detection. IEEE Signal Processing Letters, 27: 775-779. https://doi.org/10.1109/LSP.2020.2989674
[10] Liu, Q., Li, X., Yuan, D., Yang, C., Chang, X.J., He, Z.Y. (2024). LSOTB-TIR: A large-scale high-diversity thermal infrared single object tracking benchmark. IEEE Transactions on Neural Networks and Learning Systems, 35(7): 9844-9857. https://doi.org/10.1109/TNNLS.2023.3236895
[11] Liu, Q., Yuan, D., Fan, N.N., Gao, P., Li, X., He, Z.Y. (2023). Learning dual-level deep representation for thermal infrared tracking. IEEE Transactions on Multimedia, 25: 1269-1281. https://doi.org/10.1109/TMM.2022.3140929
[12] Du, H.S., Zhang, W.Z., Zhang, Z.Y., Cao, L.B., Wang, S. (2025). A coordinated interaction and cross enhancement network for salient object detection. Signal, Image and Video Processing, 19: 941. https://doi.org/10.1007/s11760-025-04448-2
[13] Zhong, J., Jiang, A.M., Liu, C., Xu, N., Zhu, Y.P. (2025). Depth completion with super-resolution and cross-modality optimization. IEEE Robotics and Automation Letters, 10(6): 5585-5592. https://doi.org/10.1109/LRA.2025.3560860
[14] Zhang, C., Yang, Z.C., He, X.D., Deng, L. (2020). Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing, 14(3): 478-493. https://doi.org/10.1109/JSTSP.2020.2987728
[15] Huang, Z., Jia, Y.J., Zhang, Y.L., Liu, X.Y., Shen, H.C., Wen, W.L. (2024). Joint source-channel coding for image super-resolution tasks in semantic communications. IEEE Transactions on Vehicular Technology, 73(12): 19034-19039. https://doi.org/10.1109/TVT.2024.3442204
[16] Wu, P.L., Chen, S.Y. (2020). Innovation in optimization of virtual space experience using interactive engine and device-Example of a song dynasty landscape painting. Sensors and Materials, 32(10): 3419-3428. https://doi.org/10.18494/SAM.2020.2920
[17] Lai, Y.I., Chou, T.C., Huang, L.P. (2025). A design-fit approach to architecture, engineering, and construction digitalization: Leveraging big data and real-scene imaging in landscape projects. IT Professional, 27(6): 81-86. https://doi.org/10.1109/MITP.2025.3611525
[18] Sivrikaya, F. (2011). The importance of spatial accuracy in characterizing stand types using remotely sensed data. African Journal of Biotechnology, 10(66): 14891-14906. https://doi.org/10.5897/AJB11.2827
[19] Kirsch, R.A. (2010). Precision and accuracy in scientific imaging. Journal of Research of the National Institute of Standards and Technology, 115(3): 195-199. https://doi.org/10.6028/jres.115.011
[20] Guo, P., Xie, G.Q., Li, R.F., Hu, H. (2021). Multi-modal image fusion via convolutional morphological component analysis and guided filter. Journal of Circuits, Systems and Computers, 30(2): 2130003. https://doi.org/10.1142/S0218126621300038
[21] Bai, H.W., Zhao, Z.X., Zhang, J.S., Jiang, B.S., Deng, L.L., Cui, Y.K. (2025). Deep unfolding multi-modal image fusion network via attribution analysis. IEEE Transactions on Circuits and Systems for Video Technology, 35(4): 3498-3511. https://doi.org/10.1109/TCSVT.2024.3507540