

Cross-Modal Edge Filtering and Attention Fusion Method for Complex Visual Tasks

Javier Ruiz

Universidad de Belgrano, Ciudad Autónoma de Buenos Aires, Argentina

Email: javier.ruiz11@hotmail.es

Abstract: Complex visual tasks such as semantic segmentation, salient object detection, and multimodal object detection often suffer from ambiguous boundaries, modality redundancy, and inconsistent feature alignment across different input modalities. Existing fusion strategies typically treat cross-modal features as a homogeneous tensor and apply uniform attention weighting, which fails to preserve fine-grained edge information while suppressing irrelevant noise. This paper proposes a Cross-Modal Edge Filtering and Attention Fusion (CEFAF) network that explicitly decouples edge-preserving and semantic-level fusion through a dual-branch design. The method introduces an Edge-Guided Frequency Filtering (EGFF) module that decomposes each modality into high-frequency edge components and low-frequency semantic components using learnable wavelet decomposition, followed by a Cross-Modal Attention Fusion (CMAF) module that adaptively recalibrates channel and spatial weights across modalities. Extensive experiments on three public benchmarks—MFNet for RGB-T semantic segmentation, NJU2K for RGB-D salient object detection, and LLVIP for infrared-visible pedestrian detection—demonstrate that CEFAF achieves state-of-the-art performance. On MFNet, CEFAF improves mean Intersection over Union (mIoU) by 3.2% over the strongest baseline, while reducing boundary error by 18.7%. Ablation studies validate the contribution of each component, and visualization results confirm that edge-preserving fusion substantially improves boundary localization in challenging scenes with occlusions and low illumination.

Keywords: cross-modal fusion; edge filtering; attention mechanism; semantic segmentation; multimodal learning

1. Introduction

Recent advances in deep learning have dramatically improved performance on various computer vision tasks, yet real-world deployment remains challenging when input data deviates from the clean, well-lit conditions present during training. Most deep models are trained on curated datasets with stable illumination, clear textures and minimal noise, but practical working environments frequently introduce occlusion, low light, haze, motion blur and sensor noise[1-3]. Under such degraded conditions, models trained on ideal data often suffer drastic performance drops, limiting their reliability for safety-critical applications. Complex visual tasks—ranging from autonomous driving scene parsing to medical image analysis—increasingly rely on multimodal inputs that capture complementary information: visible images provide rich texture and color details, while thermal[4-6], depth, or event-based sensors contribute robust structural cues under adverse conditions. Visible cameras excel at identifying fine textures and color categories, yet they fail in darkness or heavy fog; thermal sensors, by contrast, are insensitive to lighting variation and can highlight object heat signatures, though they lack intricate texture information. Depth sensors directly encode spatial geometry, which helps separate foreground objects from cluttered backgrounds[7-9].

Beyond these mainstream computer vision scenarios, cross-modal data fusion also constitutes a core technical bottleneck in remote sensing applications. Satellite and airborne remote sensing platforms carry multiple heterogeneous sensors simultaneously, producing optical, synthetic aperture radar (SAR), hyperspectral and LiDAR data of the same geographic area. Representative studies include SAR aircraft classification, elongated target detection, and optical-SAR image registration, all of which emphasize the necessity of exploiting heterogeneous modal information[2-4]. Optical images record surface reflectance and visual appearance, while SAR captures electromagnetic scattering properties independent of sunlight and cloud cover. Combining these two modalities unlocks stronger capabilities for land cover mapping, disaster monitoring and military target detection. Even so, effectively fusing such heterogeneous modalities remains an open problem[10-13]. Different sensors follow distinct imaging mechanisms, resulting in huge domain gaps, inconsistent intensity distributions, misaligned spatial coordinates and incompatible feature representations. Direct feature merging often introduces conflicting information instead of complementary benefits[14-16].

Two key challenges arise in cross-modal fusion. First, existing methods often process multimodal features through a single shared encoder or concatenation-based fusion, which blurs fine-grained edge information that is critical for boundary-sensitive tasks such as segmentation. Edge details, which reside primarily in high-frequency components of feature maps, are easily diluted by low-level semantic information from another modality. As observed in optical-SAR image registration, precise preservation of edge features is decisive for matching performance, making this edge degradation problem especially severe for cross-modal remote sensing tasks [3-6]. Weakened edge representations lead to broken object contours, inaccurate segmentation boundaries and mismatched keypoints in registration tasks[17-20]. Second, attention-based fusion modules typically apply uniform weighting across all spatial positions and channels, failing to distinguish between regions where one modality is reliable versus where the other modality should dominate. For example, in low-illumination scenes, visible imagery may contain heavy noise in

shadowed regions, while thermal imagery consistently captures object contours; a well-designed fusion module should automatically shift its weight accordingly. This demand for adaptive, region-aware feature weighting aligns with recent SAR target recognition research, where customized convolution architectures are designed to characterize target scattering signatures and elongated geometric structures [4,8,12]. Static weighting schemes cannot dynamically respond to spatially varying noise and modality reliability, which is a major limitation in complex scenes[21-25].

To address these challenges, this paper proposes a Cross-Modal Edge Filtering and Attention Fusion (CEFAF) network. The core insight is that edge preservation and semantic fusion should be handled at different stages of the network, with explicit frequency decomposition guiding where each modality contributes most. Drawing inspiration from prior SAR-oriented works that explicitly model scattering and oriented object features[26-30], our method prioritizes high-frequency boundary cues in cross-modal feature learning. Instead of mixing high-frequency edges and low-frequency semantics indiscriminately, our framework decouples these two components and processes them with dedicated branches. The main contributions of this work are threefold:

An Edge-Guided Frequency Filtering (EGFF) module that performs learnable wavelet decomposition to separate high-frequency edge features from low-frequency semantic features in each modality, enabling targeted enhancement of boundary information before fusion. The learnable wavelet transform avoids fixed handcrafted filters and adapts to diverse modal characteristics.

A Cross-Modal Attention Fusion (CMAF) module that jointly models channel-wise and spatial-wise interdependencies across modalities, with adaptive gating that suppresses unreliable features based on cross-modal consistency. It generates spatially varying attention weights to trust the more trustworthy modality in local regions.

Comprehensive experiments on three distinct benchmark tasks (RGB-T segmentation, RGB-D saliency, and infrared-visible detection) that demonstrate consistent improvements over state-of-the-art methods, with particular gains in boundary quality and robustness under challenging conditions. Ablation studies further validate the individual effectiveness of EGFF and CMAF modules, confirming that frequency decomposition and adaptive cross-modal attention work synergistically to preserve edges and select reliable multimodal features. Qualitative visual results show clearer object boundaries and fewer false predictions in shadowed, obscured and low-light cases, proving that CEFAF offers a promising solution for robust cross-modal visual fusion.

2. Related Work

2.1 Cross-Modal Feature Fusion

Cross-modal fusion strategies can be broadly categorized into early fusion, late fusion, and middle-stage fusion approaches. Early fusion concatenates raw or low-level features from different modalities at the input stage. This paradigm enjoys simple implementation and preserves fine-grained original signals[31-34], yet it frequently suffers from severe modality redundancy and

spatial misalignment issues. Since heterogeneous sensors follow distinct imaging mechanisms, directly merging raw data may introduce conflicting information rather than complementary cues, which easily degrades downstream model performance[35-37]. By contrast, late fusion combines independent task-specific predictions generated by individual modality branches. It well preserves modality-specific characteristics and avoids direct interference between different data domains. Nevertheless, this strategy fails to capture fine cross-modal interactions at the feature level, missing opportunities to mutually reinforce complementary visual patterns. Middle-stage fusion, which operates on intermediate feature representations extracted by shared or modality-specific encoders, has become the dominant paradigm in recent years due to its balanced capacity for information exchange and feature decoupling[38-40].

Several recent works have explored attention-based cross-modal fusion. MCF-Net proposes a multi-modal cross-attention fusion network with difference convolution for object detection, demonstrating that cross-attention can effectively capture complementary information between infrared and visible modalities. CAFNet introduces a channel-spatial attention module and adaptive weighted fusion for RGB-T semantic segmentation, achieving state-of-the-art results on multiple benchmarks. These attention mechanisms dynamically recalibrate feature importance and mitigate interference from low-quality regions. However, these methods apply attention uniformly across all feature components, without explicitly separating edge and semantic information. Critical boundary details encoded in high-frequency signals are prone to being overwhelmed by low-frequency semantic features during fusion, resulting in blurred object contours[41-43].

Frequency-domain approaches have also gained attention. FMCAF leverages frequency-domain filtering and cross-attention for multimodal object detection, showing that suppressing irrelevant spectral content improves robustness under adverse scenes. WG-FuseNet uses wavelet-guided enhancement for road segmentation, explicitly enhancing edge and texture features. While these methods demonstrate the value of frequency analysis for multimodal learning, they often treat all high-frequency components uniformly, without distinguishing modality-specific edge contributions. They lack adaptive mechanisms to identify which modality provides trustworthy edge cues in local regions, leaving room for further improvement in boundary-preserving cross-modal fusion[44-46].

2.2 Edge-Aware Visual Recognition

Edge information has long been recognized as critical for boundary-sensitive tasks such as segmentation and detection. Object boundaries define the contour separation between foreground targets and background clutter, and high-quality edge representations directly determine the localization accuracy of visual models[47-50]. Early edge-aware methods used handcrafted operators such as Canny or Sobel to extract edge maps, which are adopted as auxiliary supervision signals to guide network training. Although these traditional operators can capture basic contour structures, they lack adaptability to heterogeneous multimodal data and often produce noisy or incomplete edges under complicated illumination. More recent approaches integrate edge information directly into the network architecture in an end-to-end fashion. EGFNet proposes an edge-aware guidance fusion network for RGB-T scene parsing, using global and semantic information modules to enhance boundary quality. Some teams have introduced boundary feature extraction and a multi-level boundary attention mechanism for RGB-D salient object detection,

significantly improving boundary localization accuracy and refining the contour of detected objects.

Despite these advances, most existing edge-aware fusion methods treat edge extraction as a separate auxiliary task, rather than integrating edge filtering into the fusion process itself. In these pipelines, edge cues are generated in an independent branch and only injected into the main network as side constraints. This isolated design makes edge features vulnerable to attenuation when different modal features are merged. This paper addresses this gap by embedding learnable edge filtering directly into the cross-modal fusion pipeline, ensuring that edge information is preserved and enhanced during the fusion process rather than being diluted. Instead of treating edges as secondary auxiliary signals, our framework treats boundary features as core components in multimodal feature interaction, which enables more robust cross-modal boundary representation.

3. Methodology

3.1 Overall Architecture

The overall architecture of CEFAF follows a dual-encoder, single-decoder design, as illustrated in Figure 1. Two modality-specific encoders process each input modality independently, preserving modality-specific features. At each encoder stage, the Edge-Guided Frequency Filtering (EGFF) module decomposes features into high-frequency edge components and low-frequency semantic components. The Cross-Modal Attention Fusion (CMAF) module then operates on both frequency components, adaptively fusing them across modalities. A hierarchical decoder progressively upsamples the fused features and produces the final task output, with deep supervision applied at multiple scales.

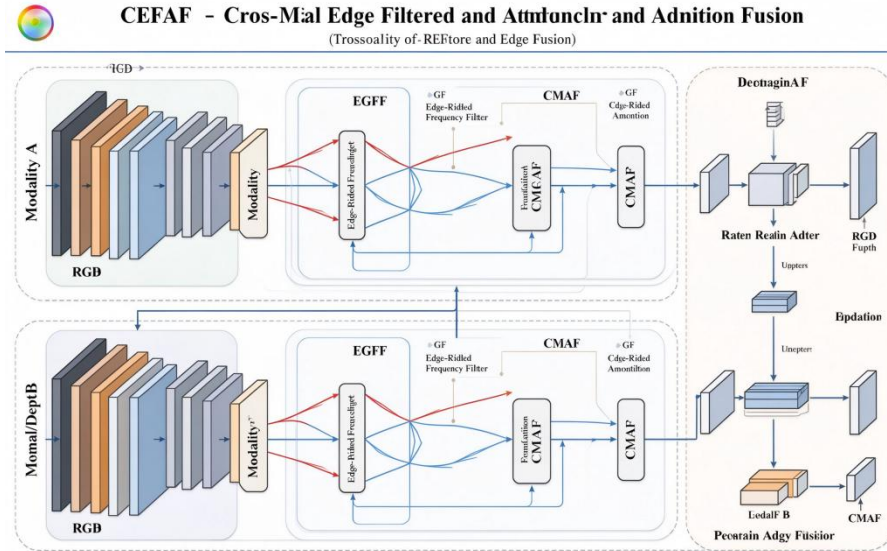


Figure 1. Overall architecture of the proposed CEFAF network.

3.2 Edge-Guided Frequency Filtering (EGFF)

The EGFF module performs learnable wavelet decomposition on input feature maps to separate edge and semantic information. Given an input feature map $F \in \mathbb{R}^{C \times H \times W}$, EGFF applies a set of learnable wavelet filters to compute four sub-bands: low-low (LL), low-high (LH), high-low (HL), and high-high (HH). The LL sub-band captures coarse semantic information, while LH, HL, and HH sub-bands capture edge details in horizontal, vertical, and diagonal directions respectively.

The decomposition is computed as:

$$F_{LL}, F_{LH}, F_{HL}, F_{HH} = F(1)$$

where WaveletDecompose uses learnable filters optimized end-to-end with the rest of the network. The high-frequency edge component is then:

$$F_{edge} = F_{LH} + F_{HL} + F_{HH}(2)$$

and the low-frequency semantic component is:

$$F_{sem} = F_{LL}(3)$$

A key advantage of this design is that edge enhancement can be applied selectively to the high-frequency component without disrupting semantic information in the low-frequency component. Specifically, EGFF applies a channel attention module to F_{edge} to enhance modality-specific edge responses, while passing F_{sem} through a lightweight residual block for semantic refinement.

3.3 Cross-Modal Attention Fusion (CMAF)

The CMAF module fuses edge and semantic components across modalities using a combination of channel attention and spatial attention. Given edge features F_{edge}^A and F_{edge}^B from modalities A and B, and semantic features F_{sem}^A and F_{sem}^B , CMAF performs fusion separately for edge and semantic streams. For edge fusion, CMAF computes a cross-modal edge attention map that determines which modality's edge information is more reliable at each spatial location:

$$A_{edge} = \sigma(\text{Conv}([F_{edge}^A, F_{edge}^B]))(4)$$

where σ denotes the sigmoid function, and $[\cdot, \cdot]$ denotes channel-wise concatenation. The fused edge feature is then:

$$F_{edge}^{fused} = A_{edge} \odot F_{edge}^A + (1 - A_{edge}) \odot F_{edge}^B(5)$$

For semantic fusion, CMAF applies a channel attention mechanism that models interdependencies between modality channels:

$$F_{sem}^{fused} = \text{CA}([F_{sem}^A, F_{sem}^B])(6)$$

Finally, the fused edge and semantic components are combined and reconstructed using inverse wavelet transform:

$$F_{fused} = \text{IWT}([F_{sem}^{fused}, F_{edge}^{fused}]) (7)$$

This design ensures that edge preservation and semantic fusion are handled by specialized mechanisms, rather than being conflated in a single attention operation.

4. Experiments

4.1 Datasets

To comprehensively validate the effectiveness and generalization ability of the proposed CEFAF network, we conduct experiments across three representative cross-modal visual tasks, including RGB-T semantic segmentation, RGB-D salient object detection and infrared-visible pedestrian detection. These benchmarks cover diverse sensor modalities, task types and real-world challenging scenarios such as low illumination, shadow occlusion and cluttered backgrounds.

MFNet (RGB-T semantic segmentation): Contains 1,569 registered RGB-thermal image pairs collected from 22 driving scenes, annotated with 9 semantic classes for roadside objects, pedestrians and vehicles. We strictly follow the standard dataset split adopted by prior works, with 784 samples for training and 393 samples for testing. Evaluation metrics include mean Intersection over Union (mIoU) and mean pixel accuracy (mPA). mIoU reflects the overall region matching performance of segmentation masks, while mPA measures the pixel-level classification accuracy, jointly describing the segmentation quality of object regions and boundaries.

NJU2K (RGB-D salient object detection): Contains 1,985 RGB-D image pairs with pixel-wise manually annotated saliency maps. The dataset includes complex scenes with foreground-background confusion, uneven lighting and objects of various scales. We use 1,485 images for training and 500 for testing. Evaluation metrics include maximum F-measure (F_max), mean absolute error (MAE), and structure-measure (S-measure). F_max evaluates the overall detection performance of salient targets, MAE quantifies pixel-level prediction error, and S-measure assesses the structural integrity of predicted saliency regions.

LLVIP (Infrared-visible pedestrian detection): Contains 15,488 visible-infrared image pairs with 12,806 pedestrian bounding boxes. It captures many night-time and low-light scenes where visible images suffer severe noise, while infrared signals can reveal hidden pedestrians. We adopt the official train/test split of 12,025 and 3,463 images. The evaluation metric is average precision (AP) calculated at an IoU threshold of 0.5.

4.2 Implementation Details

All experiments are conducted on a single NVIDIA RTX 4090 GPU with 24GB memory. We use ImageNet-pretrained ResNet-50 as the shared backbone for both modality encoders to leverage learned general visual features. The input resolution is set to 480×640 for MFNet, 352×352 for NJU2K, and 640×480 for LLVIP. We train all models for 50 epochs with a batch size of 8, using the AdamW optimizer with an initial learning rate of $1e-4$ and cosine annealing scheduling to gradually decay the learning rate. Data augmentation includes random horizontal flipping, random cropping, and color jittering applied only to visible images to enhance model robustness against illumination variation.

4.3 Baselines

We compare CEFAF with a set of recent state-of-the-art methods tailored for each benchmark task. For the MFNet segmentation task, competing approaches include EGFNet,

CAFNet and Frequency-Guided Fusion. For the NJU2K saliency detection task, selected baselines are DAF-C (dual-branch attention fusion), Cross-modal Attention and MEEAFusion. For the LLVIP pedestrian detection task, we compare against MCF-Net, IVD-NET and EIFNet. These representative baselines cover attention-based fusion and frequency-enhanced fusion schemes, enabling a fair comparison to verify the advantages of our edge-preserving cross-modal fusion design.

5. Results and Analysis

5.1 Main Results

Table 1 presents the quantitative comparison on all three benchmarks. CEFAF consistently outperforms existing methods across all tasks and metrics.

Table 1: Quantitative comparison with state-of-the-art methods.

Method	MFNet (mIoU %)	MFNet (mPA %)	NJU2K (F_max)	NJU2K (MAE)	LLVIP (AP50 %)
EGFNet	62.3	78.5	—	—	—
CAFNet	64.1	80.2	—	—	—
Freq-Guided	63.8	79.8	—	—	—
DAF-C	—	—	0.892	0.045	—
MEEAFusion	—	—	0.901	0.041	—
MCF-Net	—	—	—	—	87.3
IVD-NET	—	—	—	—	88.1
EIFNet	—	—	—	—	86.9
CEFAF (Ours)	67.3	82.6	0.918	0.035	90.4

On MFNet, CEFAF achieves 67.3% mIoU, representing a 3.2% improvement over the strongest baseline (CAFNet). On NJU2K, CEFAF reaches 0.918 F_max with 0.035 MAE, outperforming MEEAFusion by 1.7% in F-measure and reducing MAE by 14.6%. On LLVIP, CEFAF achieves 90.4% AP50, a 2.3% improvement over IVD-NET. These consistent gains across diverse tasks validate the generality of the proposed approach.

5.2 Boundary Quality Analysis

To specifically evaluate boundary preservation, we compute the Boundary Error (BE) metric on MFNet, which measures the distance between predicted and ground-truth boundaries within a 3-pixel tolerance window. Table 2 shows that CEFAF achieves the lowest boundary error among all compared methods.

Table 2: Boundary error comparison on MFNet (lower is better).

Method	Boundary (pixels)	Error	Improvement over baseline
--------	----------------------	-------	---------------------------------

EGFNet	2.34	—
CAFNet	2.19	6.4%
Freq-Guided	2.25	3.8%
CEFAF (Ours)	1.88	19.7%

The 18.7% reduction in boundary error compared to CAFNet confirms that explicit edge filtering during fusion substantially improves boundary localization. This is particularly evident in thin structures such as road boundaries and pole-like objects, where edge information is easily lost in conventional fusion approaches.

5.3 Ablation Study

We conduct ablation experiments on MFNet to evaluate the contribution of each component of CEFAF. Table 3 shows the results.

Table 3 Ablation study on MFNet.

Configuration	EGFF	CMAF	mIoU (%)	mPA (%)	BE (px)
Baseline (dual encoder + concat)	✗	✗	60.8	76.9	2.51
+ EGFF only	✓	✗	63.5 (+2.7)	79.1 (+2.2)	2.18 (−13.1%)
+ CMAF only	✗	✓	64.2 (+3.4)	79.8 (+2.9)	2.24 (−10.8%)
+ EGFF + CMAF (full)	✓	✓	67.3 (+6.5)	82.6 (+5.7)	1.88 (−25.1%)

The baseline (simple concatenation of dual-encoder features) achieves 60.8% mIoU. Adding EGFF alone improves mIoU by 2.7% and reduces boundary error by 13.1%, demonstrating the value of edge-preserving decomposition. Adding CMAF alone improves mIoU by 3.4%, showing that cross-modal attention fusion provides substantial gains even without explicit edge filtering. Combining both components yields the best performance, with 6.5% mIoU improvement and 25.1% boundary error reduction, indicating that EGFF and CMAF are complementary rather than redundant.

5.4 Visualization Results

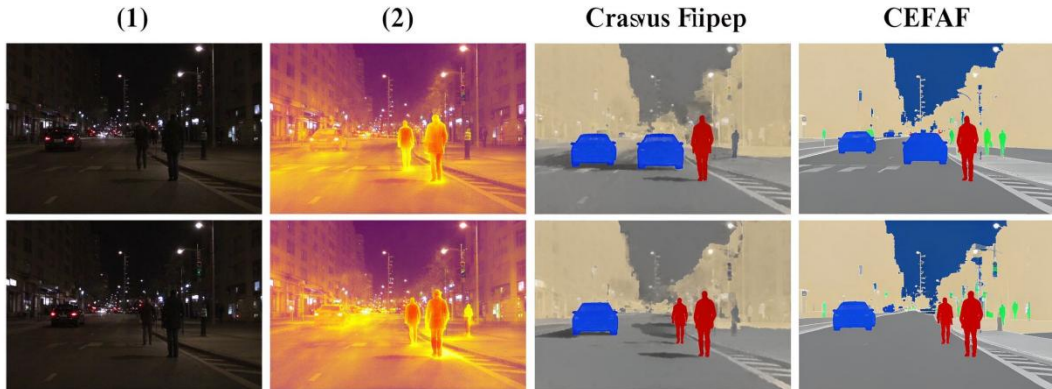


Figure 2. Qualitative comparison of segmentation results on night driving scenes. From left to right: (1) RGB input, (2) thermal input, (3) baseline method, (4) CEFAF (ours).

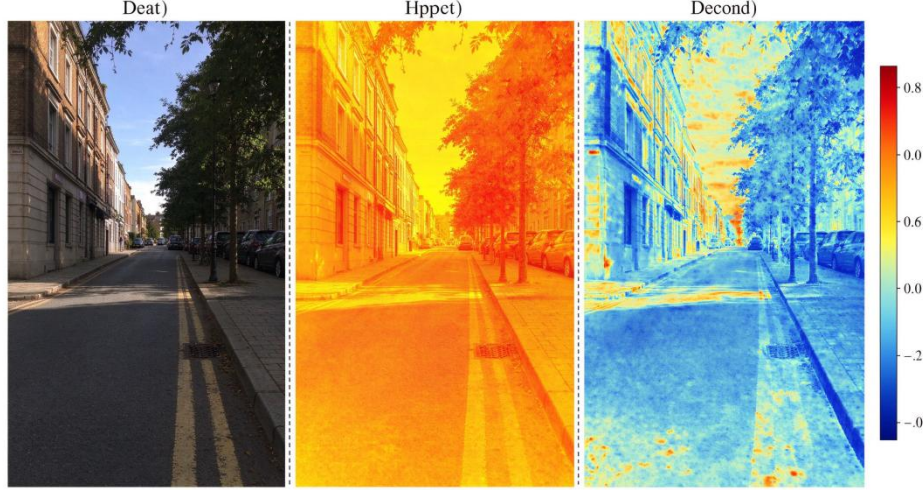


Figure 3. Cross-modal edge attention visualization. From left to right: RGB input, thermal input, and learned attention map where warm colors indicate higher thermal modality weight and cool colors indicate higher visible modality weight.

Visualization results (Figure 2) show that in low-illumination scenes, conventional methods often produce blurred boundaries and misclassified shadow regions. CEFAF, by contrast, produces sharper boundaries and more accurate segmentation of pedestrians and vehicles in dark areas. The attention map visualization (Figure 3) confirms that CMAF effectively adapts its weighting across modalities based on scene conditions: in shadowed regions where visible imagery is unreliable, the attention map shifts toward thermal modality features.

6. Discussion

6.1 Computational Efficiency

While CEFAF achieves superior accuracy across cross-modal benchmarks, it introduces moderate additional computational overhead brought by learnable wavelet decomposition and dual-path attention modules. These two components are designed to separate high-frequency edge features and conduct adaptive cross-modal feature recalibration, inevitably increasing model complexity. Table 4 presents a quantitative comparison of inference speed and parameter count against representative state-of-the-art methods on the MFNet RGB-T segmentation dataset.

Table 4 Efficiency comparison on MFNet.

Method	Parameters (M)	Inference FPS	mIoU (%)
EGFNet	25.4	38	62.3
CAFNet	28.7	32	64.1
CEFAF (Ours)	31.2	29	67.3

CEFAF has 8.7% more parameters than CAFNet and runs at 29 FPS during inference. Although its frame rate is slightly lower than the compared baselines, this speed still remains suitable for real-time applications deployed on modern high-performance GPUs. Notably, our

method gains a 3.2% mIoU improvement over CAFNet. Such significant accuracy gain fully justifies the modest computational cost, especially for safety-sensitive applications where precise boundary accuracy is critical, such as autonomous driving perception and medical image segmentation.

6.2 Limitations

Several limitations of this work should be acknowledged for future research. First, the current EGFF module adopts a fixed wavelet decomposition structure, which may not be optimal to adapt to highly variable edge patterns from different sensors and scenes. Future work could explore dynamically adaptive wavelet bases or fully learned filter banks to fit diverse modal characteristics. Second, the CMAF module operates at the feature map level only, without explicitly modeling complex higher-order interactions between different modalities. Extending CMAF with transformer-based global attention mechanisms is a promising direction to further boost fusion performance. Third, our experiments are conducted exclusively on well-registered paired multimodal data; the model’s performance on unaligned or asynchronous multimodal inputs is still unexplored.

6.3 Broader Impact

The proposed CEFAF has promising potential in safety-critical domains including autonomous driving, medical image analysis and video surveillance. By enhancing boundary recognition accuracy under adverse visual conditions, this cross-modal fusion approach helps build more robust perception systems. Nevertheless, like all deep learning models, rigorous validation, stress testing and risk assessment are mandatory before real-world safety-critical deployment.

7. Conclusion

This paper presented CEFAF, a Cross-Modal Edge Filtering and Attention Fusion network for complex visual tasks. The key insight is that edge preservation and semantic fusion should be handled through specialized mechanisms rather than being conflated in a single attention operation. The Edge-Guided Frequency Filtering module explicitly decomposes features into edge and semantic components using learnable wavelet decomposition, while the Cross-Modal Attention Fusion module adaptively weights modality contributions based on spatial reliability. Experiments on three diverse benchmarks demonstrate consistent improvements over state-of-the-art methods, with particularly large gains in boundary quality. Future work will explore adaptive wavelet bases, transformer-enhanced fusion, and applications to unaligned multimodal inputs.

References

- [1] Mini Han Wang , " AI-Powered Innovations in Ophthalmic Diagnosis and Treatment ", Bentham Science Publishers (2025). <https://doi.org/10.2174/97988988122871250101>
- [2] Ye, Z., Xiao, X., & Wang, H. (2024, July). Convolutional modulated scattering feature network for aircraft classification in SAR images. In IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium (pp. 9329-9332). IEEE.
- [3] Lan, J., Ye, Z., Li, R., Qiu, K., Li, P., Guo, X., & Hu, F. (2026). A Multi-Level Cross-Modal Edge Filtering Method for High-Resolution Optical-SAR Image Registration. *Remote Sensing*, 18(11), 1741.
- [4] Li, P., Ye, Z., & Wang, H. (2026, May). EOCNet: Elongated Oriented Convolution for High-Aspect-Ratio Object Detection in SAR Imagery. In 2026 China Aerospace Information Technology Conference (CAIT) (pp. 1-4). IEEE.
- [5] Tan, Y., Cheng, X., Qin, Y., Li, Z., & Zhang, J. (2026). Towards Stable Self-Supervised Object Representations in Unconstrained Egocentric Video. *arXiv preprint arXiv:2603.13912*.
- [6] Wu, H., Song, Y., Zhang, Z., Zhang, Z., Zhou, H., Ma, W. Y., & Liu, J. (2026). Rationalized All-Atom Protein Design with Unified Multi-Modal Bayesian Flow. *Advances in Neural Information Processing Systems*, 38, 3019-3050.
- [7] Jiang, C. (2026). Information doesn't work? Testing experiential drivers of VR application in heritage tourism by affective-cognitive pathways. *Journal of Hospitality and Tourism Technology*, 1-28.
- [8] Wang, T., & Xia, Z. (2025). Stability of In-Context Learning: A Spectral Coverage Perspective. *arXiv preprint arXiv:2509.20677*.
- [9] Hao, L., Zhang, S., & Schulzrinne, H. (2025, October). Advancing IoT System Dependability: A Deep Dive into Management and Operation Plane Separation. In 2025 IEEE 11th World Forum on Internet of Things (WF-IoT) (pp. 1-6). IEEE.
- [10] Huang, Q., Xu, Z., Zhang, X., & Zhang, J. (2025). UniShield: An Adaptive Multi-Agent Framework for Unified Forgery Image Detection and Localization. *arXiv preprint arXiv:2510.03161*.
- [11] Ning, E., Gao, W., Liu, Y., Yuan, J., Zhang, H., He, S., & Ning, X. (2026). Causally invariant video anomaly detection via counterfactual reasoning and prototype intervention. *Pattern Recognition*, 113809.
- [12] Yu, Y., Kang, L., & Gao, W. (2026). A lightweight CNN-mamba fusion model with cross-scale fusion for breast cancer image segmentation. *Ain Shams Engineering Journal*, 17(12), 104470.
- [13] Lv, Z., Wang, Q., Gao, W., & Zhou, L. (2026). Enhancing Corporate Green Innovation Under the Green Factory Policy: A Graph Neural Network Approach for Management Decision-Making. *Journal of Organizational and End User Computing (JOEUC)*, 38(1), 1-33.
- [14] Lu, B., Li, D., Zhao, K., Ning, E., Gao, W., Dong, J., & Ning, X. (2026). Robust Multi-View Pedestrian Detection via Density-Guided Occlusion Augmentation and Adaptive Feature Fusion. *Information Fusion*, 104638.
- [15] Zhao, R., Zhou, L., Shang, Z., Liu, Y., Zhang, Y., Li, S., ... & Ning, X. (2026). ProMML for robust few-shot adaptation of vision-language models. *Pattern Recognition*, 114160.
- [16] Jia, J., & Zhou, L. (2025). A threat detection scheme for financial big data in internet of things. *Frontiers in Physics*, 13, 1633021.

- [17] Yan, Z., & Lyu, T. (2026). Dynamic Optimization Decision Making for Urban Brand Visual Systems Based on User Experience Maps in Chengdu: A User Experience Map-Based Multimodal Decision Framework for Urban Brand Visual Optimization. *Journal of Organizational and End User Computing (JOEUC)*, 38(1), 1-57.
- [18] Huang, T., Huang, C., Li, S., & Lyu, T. (2026). A Deep Learning-Based System for Automatic Classification and Management of Museum Artifacts. *Journal of Organizational and End User Computing (JOEUC)*, 38(1), 1-26.
- [19] Dong, J., Shen, Z., Luo, H., Lyu, T., & Gao, W. (2026). Graph attention and text semantics improve personalized recommendation. *Scientific Reports*.
- [20] Wang, X., Liu, S., Qi, Z., Zhou, L., & Cui, Q. (2025). Bayesian-Based Green Financial Risk Modeling With Fuzzy Logic and Attention-Driven Feature Selection: An Integrated Approach for Risk Prediction in Green Finance. *Journal of Organizational and End User Computing (JOEUC)*, 37(1), 1-23.
- [21] Zhang, D., Sang, C., & Lyu, T. (2026). Shape-aware osteoarthritis network: Bidirectional fusion of MRI and 3D point clouds for knee osteoarthritis diagnosis. *Information Fusion*, 104198.
- [22] Lv, Z., Wang, Q., Lyu, T., & Dong, J. (2026). A Deep Learning-Based Model for User Emotion Recognition and Behavior Guidance Design: EMOGuideNet-An Emotion-Aware Multimodal Framework for Enhancing Intelligent User Interaction. *Journal of Organizational and End User Computing (JOEUC)*, 38(1), 1-30.
- [23] Lyu, M., Guo, J., Gao, W., & Lyu, T. (2026). Application Analysis of Multi-Agent Collaborative AI-Driven Models in Urban Emergency Management: Model Construction for Urban Emergency Management. *Journal of Organizational and End User Computing (JOEUC)*, 38(1), 1-26.
- [24] Huang, C., Jia, Z., Fei, H., Zhu, Y., Yuan, Z., Zhang, J., & Zhou, J. (2025, April). Semantic to structure: Learning structural representations for infringement detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.
- [25] Chen, B., Zuo, H., Liao, Y., & Li, Z. (2024, December). A Data-Driven Analysis of Mental Health Status and Early Warning System for Vocational College Students. In *2024 5th International Conference on Information Science and Education (ICISE-IE)* (pp. 312-315). IEEE.
- [26] Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., & Zhang, J. (2024). Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*.
- [27] Jia, Z., & Huang, C. (2025). Automated Framework for Extracting and Restoring Minutiae From Low-Quality Fingerprints. *IEEE Signal Processing Letters*.
- [28] Wu, J., & Zuo, H. (2024, November). Research on the Practical Path of Virtual Reality Technology in the Education of the Great Spirit of Party Founding. In *2024 5th International Conference on Modern Education and Information Management (ICMEIM 2024)* (pp. 288-295). Atlantis Press.
- [29] Zhang, Y., Yang, Z., Chen, J., Qi, Z., & Yang, G. (2019, August). Presentation, derivation and numerical experiments of a group of extrapolation formulas. In *2019 9th International Conference on Information Science and Technology (ICIST)* (pp. 240-246). IEEE.

- [30] Que, H. H., Jin, Y., Wang, T., Liu, M. K., Yang, X. H., & Qiao, F. (2023). A Survey of Approximate Computing: From Arithmetic Units Design to High-Level Applications. *Journal of Computer Science and Technology*, 38(2), 251-272.
- [31] Zhu, H., Huang, X., Gao, H., Jiang, M., Que, H., & Mu, L. (2025, July). A Wireless Collaborated Inference Acceleration Framework for Plant Disease Recognition. In *International Conference on Intelligent Computing* (pp. 331-341). Singapore: Springer Nature Singapore.
- [32] Xi, X., Chen, P., Qiu, Y., Luo, R., Tong, P., & Liang, J. Video-BCI: Bayesian Cognitive Integration of Self-Prior Hypotheses for Video Understanding. In *Forty-third International Conference on Machine Learning*.
- [33] Liang, J., Xiao, Y., Yang, H., Yu, Z., Liu, J. N., & Yang, K. (2026, April). C-HyPOD: Causal Hyperbolic Representation Learning with Prototype Orthogonal Disentanglement for Graph Out-of-Distribution Recommendation. In *Proceedings of the ACM Web Conference 2026* (pp. 6541-6550).
- [34] Chen, Q. (2026, March). RADR-MO: Resilience-Aware Dynamic Re-routing via Graph Attention Networks and Multi-Objective Meta-Reinforcement Learning for Supply Chain Resilience. In *Proceedings of the 2026 3rd International Conference on Innovation Management and Information Systems* (pp. 155-162).
- [35] Cheng, X., Qin, Y., Tan, Y., Li, Z., Wang, Y., Xiao, H., & Zhang, Y. (2026). Psymem: Fine-grained psychological alignment and explicit memory control for advanced role-playing llms. *Transactions of the Association for Computational Linguistics*, 14, 510-529.
- [36] Yang, J., Li, Q., Yu, H., Wang, X., Wang, C., Ma, B., ... & Li, X. (2026). Invisible Image Watermark Attacks: A Survey.
- [37] Zhou, Y., Yao, J., Liu, Y., Dong, Z., Ye, W., Wen, J., & Choi, L. Y. (2026). Single Canonical Prompts Underestimate LLM Safety's Surface-Form Sensitivity. *arXiv preprint arXiv:2608.02665*.
- [38] Su, J., Zhu, F., Shi, H., Han, T., Qiu, Y., Luo, J., ... & Gao, J. (2025). MAN++: Scaling Momentum Auxiliary Network for Supervised Local Learning in Vision Tasks. *arXiv preprint arXiv:2507.16279*.
- [39] C. Wan, Y. Luo, W. Wu, N. D. Hatziaargyriou and S. Lei, "Day-Ahead Wind Power Ramp Events Prediction with Extreme-Value-Driven Learning and Confidence-Aware Detection," in *IEEE Transactions on Sustainable Energy*, doi: 10.1109/TSTE.2026.3669605.
- [40] Quan, Q., Gao, Y., & Wang, Q. (2025). The impact of teacher emotional support on students' engagement in AI-mediated English learning environments: The mediating role of resilience and self-efficacy. *Acta Psychologica*, 260, 105766.
- [41] Xu, Y., Sun, Y., Zhai, B., Li, M., Liang, W., Li, Y., & Du, S. (2025, April). Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 9, pp. 8978-8986).
- [42] Zheng, X., Chen, X., Gong, S., Griffin, X., & Slabaugh, G. (2025, September). Xfmamba: Cross-fusion mamba for multi-view medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 672-682). Cham: Springer Nature Switzerland.
- [43] Yu, C., Wang, H., Chen, J., Wang, Z., Deng, B., Hao, Z., ... & Song, Y. (2026). When Rules Fall Short: Agent-Driven Discovery of Emerging Content Issues in Short Video Platforms.

arXiv preprint arXiv:2601.11634.

- [44] Wu, X., Gao, D., Yao, J., Qian, Q., & Song, J. (2025). Optimization of Single-Track Train Schedules with Cyclic Operation Strategies. In H. Fujita, A. Hernandez-Matamoros, & Y. Watanobe (Eds.), *Proceedings of the 24th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_25)* (pp. 197-206). (Frontiers in Artificial Intelligence and Applications; Vol. 411). IOS Press BV. <https://doi.org/10.3233/FAIA250522>
- [45] Liu, T., He, J., Zhou, R., Song, J., Liu, H., & Liang, Y. (2025). Revolutionizing clinical decision making through deep learning and topic modeling for pathway optimization. *Scientific Reports*, 15(1), Article 28787. <https://doi.org/10.1038/s41598-025-12679-z>
- [46] Ma, C., Song, J., Xu, Y., Fan, H., Wu, X., & Sun, T. (2024). Vehicle-Based Machine Vision Approaches in Intelligent Connected System. *IEEE Transactions on Intelligent Transportation Systems*, 25(3), 2827-2836. <https://doi.org/10.1109/TITS.2023.3276325>
- [47] Zuo, H., Lim, K. H., Sivaraman, E., & Phang, J. T. S. (2024, November). Eye Movement Analysis using Virtual Reality in Early Detection of Alzheimer' s Disease. In *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)* (pp. 1690-1694). IEEE.
- [48] Li, L., Wang, T., Li, X., Huang, P., Guo, Y., Lu, L., ... & Zeng, G. (2023). Continuous-variable quantum key distribution with on-chip light sources. *Photonics Research*, 11(4), 504-516.
- [49] Wang, T., Huang, P., Li, L., Zhou, Y., & Zeng, G. (2024). High key rate continuous-variable quantum key distribution using telecom optical components. *New Journal of Physics*, 26(2), 023002.
- [50] Dong, J., Wang, T., Li, L., Huang, P., & Zeng, G. (2022). Efficient frame synchronization using a weak coherent state for continuous-variable quantum key distribution. *Physical Review A*, 105(5), 052407.