

Article

DLiteNet: A Dual-Branch Lightweight Framework for Efficient and Precise Building Extraction from Visible and SAR Imagery

Zhe Zhao ^{1,2}, Boya Zhao ¹ , Ruitong Du ¹, Yuanfeng Wu ^{1,*} , Jiaen Chen ³ and Yuchen Zheng ³

¹ Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; zhaozhe22@mails.ucas.ac.cn (Z.Z.); zhaoby@aircas.ac.cn (B.Z.); durt@aircas.ac.cn (R.D.)

² College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

³ College of Information Science and Technology, Shihezi University, Shihezi 832000, China; 20222108012@stu.shzu.edu.cn (J.C.); zhengyuchen@shzu.edu.cn (Y.Z.)

* Correspondence: wuyf@aircas.ac.cn

Highlights

What are the main findings?

- A dual-branch lightweight multimodal framework (DLiteNet) is proposed. It decouples building extraction into a context branch for global semantics (via STDAC) and a CDAM-guided spatial branch for edges and details, with MCAM adaptively fusing visible–SAR features. Removing the complex decoding stage enables efficient segmentation.
- DLiteNet consistently outperforms state-of-the-art multimodal building-extraction methods on the DFC23 Track2 and MSAW datasets, achieving a strong efficiency–precision trade-off and demonstrating strong potential for real-time on-board deployment.

What is the implication of the main finding?

- By removing complex decoding and adopting a dual-branch, task-decoupled design, DLiteNet shows that accurate visible–SAR building extraction is achievable under tight compute/memory budgets, enabling large-area, high-frequency mapping and providing a reusable blueprint for other multimodal segmentation tasks (e.g., roads, damage, change detection).
- Its lightweight yet precise architecture makes real-time on-board deployment on UAVs and other edge platforms practical for city monitoring and rapid disaster response.



Academic Editor: Wen Liu

Received: 31 October 2025

Revised: 1 December 2025

Accepted: 3 December 2025

Published: 5 December 2025

Citation: Zhao, Z.; Zhao, B.; Du, R.; Wu, Y.; Chen, J.; Zheng, Y. DLiteNet: A Dual-Branch Lightweight Framework for Efficient and Precise Building Extraction from Visible and SAR Imagery. *Remote Sens.* **2025**, *17*, 3939. <https://doi.org/10.3390/rs17243939>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract

High-precision and efficient building extraction by fusing visible and synthetic aperture radar (SAR) imagery is critical for applications such as smart cities, disaster response, and UAV navigation. However, existing approaches often rely on complex multimodal feature extraction and deep fusion mechanisms, resulting in over-parameterized models and excessive computation, which makes it challenging to balance accuracy and efficiency. To address this issue, we propose a dual-branch lightweight architecture, DLiteNet, which functionally decouples the multimodal building extraction task into two sub-tasks: global context modeling and spatial detail capturing. Accordingly, we design a lightweight context branch and spatial branch to achieve an optimal trade-off between semantic accuracy and computational efficiency. The context branch jointly processes visible and SAR images, leveraging our proposed Multi-scale Context Attention Module (MCAM) to adaptively fuse multimodal contextual information, followed by a lightweight Short-Term Dense Atrous Concatenate (STDAC) module for extracting high-level semantics. The spatial branch focuses on capturing textures and edge structures from visible imagery and employs a Context-Detail Aggregation Module (CDAM) to fuse contextual priors

and refine building contours. Experiments on the MSAW and DFC23 Track2 datasets demonstrate that DLiteNet achieves strong performance with only 5.6 M parameters and extremely low computational costs (51.7/5.8 GFLOPs), significantly outperforming state-of-the-art models such as CMGFNet (85.2 M, 490.9/150.3 GFLOPs) and MCANet (71.2 M, 874.5/375.9 GFLOPs). On the MSAW dataset, DLiteNet achieves the highest accuracy (83.6% IoU, 91.1% F1-score), exceeding the best MCANet baseline by 1.0% IoU and 0.6% F1-score. Furthermore, deployment tests on the Jetson Orin NX edge device show that DLiteNet achieves a low inference latency of 14.97 ms per frame under FP32 precision, highlighting its real-time capability and deployment potential in edge computing scenarios.

Keywords: visible; synthetic aperture radar (SAR); building extraction; multimodal segmentation

1. Introduction

Rapid and accurate building extraction from high-resolution remote sensing images has significant applications in smart city development [1], disaster response [2], and UAV navigation. With the accelerating pace of urbanization, the demand for timely updates of building information has become increasingly urgent, which is essential for monitoring urban expansion, assessing disaster impacts, and optimizing resource allocation. As two major types of remote sensing data, visible images provide rich texture and color information of buildings [3], but are susceptible to obstructions from clouds, haze, and shadows. In contrast, synthetic aperture radar (SAR) imagery offers backscatter information and can penetrate clouds and haze [4], but often lacks sufficient texture and color cues, limiting the accuracy of building extraction when used alone. Therefore, integrating visible and SAR modalities holds great promise for achieving high-precision and all-weather building extraction.

However, the use of multimodal remote sensing data also presents significant challenges, including large data volumes and the complexity of data fusion operations, which hinder rapid and real-time building extraction. Most existing methods adopt separate feature extraction networks to process visible and SAR modalities independently, followed by intricate fusion mechanisms to integrate the extracted features for building segmentation [5]. While such strategies can improve accuracy, they substantially increase computational overhead and deployment difficulty, limiting their applicability on edge devices with constrained computational and memory resources. To improve efficiency, some researchers have explored lightweight network architectures, including MobileNet [6], ShuffleNet [7], and others, in remote sensing applications to reduce model size and computational complexity. However, in multimodal building extraction tasks, existing approaches often apply lightweight modules only in the encoder, while retaining complex structures in the decoder stage, such as computationally intensive Transformer-based decoders [8] or deep upsampling layers with skip connections [9]. As a result, these systems remain burdened by excessive parameters and limited inference speed, hindering genuine end-to-end efficiency improvements.

Furthermore, current methods often overlook the distinct imaging mechanisms of visible and SAR modalities and their complementary roles in building extraction. These two data sources are usually treated as independent, with their features simply concatenated or fused via complicated modules, lacking a targeted analysis of modality-specific contributions or structural decomposition. Visible images, rich in textures and fine-grained details, are particularly effective for capturing building contours and edges, while SAR

images exhibit strong backscattering responses and are sensitive to vertical structures, making them valuable for providing spatial context and positional priors, as shown in Figure 1. From a scattering perspective, man-made buildings often appear as bright regions in SAR images because the vertical walls and the adjacent ground form strong double-bounce reflectors that return a large portion of the incident radar energy back to the sensor. In contrast, smooth roofs, water surfaces, and vegetated areas typically produce weaker backscatter and therefore appear darker. However, most existing approaches fail to assign distinct functional roles to each modality within the network [10], leading to redundant information, inefficient feature utilization, and limited ability to simultaneously achieve high accuracy and high efficiency.

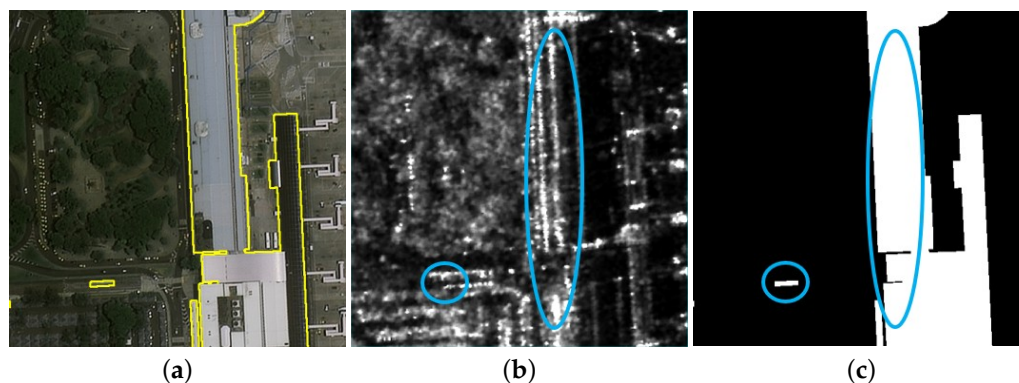


Figure 1. Illustration of modality-specific characteristics in building extraction. (a) Visible image with the building boundary highlighted in yellow, showing clear spatial details and well-defined contours. (b) SAR image with strong backscattering responses emphasized by the blue ellipse, indicating robust structural and contextual cues despite the lack of texture. (c) Ground truth mask. The comparison highlights the complementary nature of the two modalities, which motivates the dual-branch design of our proposed framework.

To overcome the limitations of existing methods, such as their lack of lightweight design and insufficient exploitation of modality complementarity, we propose DLiteNet (dual-branch lightweight network), a novel dual-branch architecture for efficient multi-modal building extraction. Distinctively, the architecture of DLiteNet is explicitly designed to bypass the computationally intensive and redundant decoding stage, thereby improving efficiency. It achieves this by functionally decoupling the building semantic segmentation task into two parallel subtasks: global context modeling and spatial detail capturing, each tailored to the characteristics of different modalities. The context branch processes both visible and SAR images to model global semantics. Specifically, it extracts texture and color representations from visible imagery and leverages the strong backscatter responses from SAR data to provide complementary structural priors for building localization and recognition. The spatial branch, operating exclusively on visible images, focuses on capturing edge information and fine-grained spatial details that are critical for delineating building boundaries and enhancing segmentation precision. Extensive experiments on two benchmark datasets validate the effectiveness of our method. DLiteNet achieves a favorable balance between segmentation accuracy and computational efficiency, outperforming several state-of-the-art methods in both metrics.

The main contributions are as follows:

1. We propose DLiteNet, a dual-branch lightweight network for multimodal building extraction using visible and SAR imagery. The overall task is functionally decoupled into two subtasks: global context modeling and spatial detail capturing. Accordingly, we design a context branch to learn global semantic representations by jointly leveraging visible and SAR information, and a spatial branch to extract fine-grained

structural and texture features from visible images. This design eliminates the need for a conventional decoder, enabling an efficient balance between accuracy and computational cost.

2. In the context branch, we propose the Multi-scale Context Attention Module (MCAM), which facilitates adaptive multimodal feature interaction through multi-scale context attention and a complementary selection gate. This module dynamically selects and integrates complementary contextual information from both modalities. Additionally, we propose the Short-Term Dense Atrous Concatenate (STDAC) module, which employs multi-rate atrous convolutions and hierarchical feature fusion to enhance multi-scale contextual understanding of buildings while significantly reducing the parameter burden.
3. In the spatial branch, a sequence of basic residual blocks is employed to extract high-resolution structural and textural features from visible images. To enhance spatial feature learning, we propose the Context-Detail Aggregation Module (CDAM), which incorporates contextual priors from the context branch to guide detail refinement. Furthermore, explicit edge supervision is incorporated to reinforce building boundary delineation, leading to more precise and structurally coherent segmentation results.

2. Related Work

2.1. Real-Time Semantic Segmentation in Natural Scenes

Various networks have been designed to simultaneously ensure high segmentation accuracy and efficient processing speed, which can be broadly categorized as follows:

Lightweight Encoder-Decoder Networks: Marin et al. designed SwiftNet [11], which utilizes a dual resolution input strategy: a low-resolution stream to capture high-level semantics, while the high-resolution counterpart supplies fine details for its lightweight decoder. Li et al. proposed DFANet [12], incorporating a depthwise separable convolution-based lightweight backbone into the Xception architecture to accelerate inference. Mostafa et al. introduced ShuffleSeg [13], employing ShuffleNet as the backbone with channel shuffling and group convolutions to reduce computational costs. Adam et al. developed E-Net [14], incorporating early feature downsampling and atrous convolutions to achieve accurate segmentation with reduced computational cost. Zhao et al. designed ICNet [15], which employs multi-scale feature cascade and feature pyramid pooling to improve segmentation speed.

Two-Branch Network Architectures: Yu et al. introduced BiSeNet [16], a two-branch network consisting of a context branch for global feature extraction and a detail branch for fine-grained feature learning, with a fusion module integrating both. Building on this, BiSeNetV2 [17] enhances context representation via an auxiliary semantic supervision branch, further improving inference speed while maintaining high accuracy. BiSeNetV3 [18], proposed by Fan et al., leverages Laplacian convolutions to generate detail-aware masks, guiding the network to learn fine-grained information while ensuring fast inference. Hong et al. developed DDRNet [19], which enhances bidirectional information exchange between the context and detail branches through bilateral connections. Xu et al. proposed PIDNet [20], extending the two-branch framework with an additional edge extraction branch, leveraging a designed Bag operation to incorporate edge information, achieving improved segmentation accuracy while maintaining efficient inference.

2.2. Semantic Segmentation in Building Extraction

Deep neural networks have led to major breakthroughs in the extraction of high-resolution remote sensing imagery due to their strong capacity to model spatial semantics

and structural patterns. Numerous approaches have been proposed based on visible or SAR images, each leveraging the unique strengths of the respective modality.

For visible imagery, researchers have developed a wide range of models based on convolutional neural networks (CNNs) to extract buildings using both spatial and spectral cues. Alshehhi et al. [21] employed classical CNN architectures for road and building extraction, improving the results by incorporating low-level features such as symmetry and structural refinement. Yang et al. [22] explored four CNN frameworks, including branched CNN, fully convolutional network, CRF-RNN, and SegNet, to extract building materials in a 56 km² area. Hui et al. [23] proposed a U-Net-based model with a specialized Xception module designed for remote sensing, achieving competitive results on two datasets.

For SAR imagery, researchers have taken advantage of the strong backscatter signatures of buildings to achieve accurate footprint delineation. Rapuzzi et al. [24] utilized CNNs to integrate multitemporal and multipolarization SAR data to build segmentation. Wu et al. [25] embedded residual blocks within a U-Net backbone to extract dense building clusters from SAR data. To address the limited semantic richness of SAR, Kang et al. [26] developed DisOptNet, which distills semantic knowledge from visible images into an SAR-only model, substantially improving segmentation accuracy.

Despite their individual merits, both visible and SAR modalities exhibit inherent limitations in building extraction tasks. Visible images often suffer from spectral confusion between buildings and other artificial structures (e.g., roads), and material diversity within buildings leads to intra-class variability, which undermines classification accuracy [27,28]. In contrast, SAR images lack color and texture cues, resulting in limited structural detail and reduced segmentation quality. To address these issues, recent efforts have focused on fusing visible and SAR modalities. For instance, Li et al. [29] proposed a progressive fusion learning framework to integrate phase characteristics from both modalities for accurate building localization. Zhang et al. [30] further introduced ASANet, an asymmetric semantic aligning network for RGB–SAR land-cover classification, which employs modality-specific feature extraction and cross-modal alignment to enhance multimodal interaction. Zhao et al. [31] applied a complex multimodal fusion pipeline to extract semantic information from visible and multispectral images, followed by refinement using edge-aware priors from visible data.

3. Materials and Methods

3.1. Datasets

3.1.1. MSAW Dataset

The MSAW dataset includes 3401 pairs of 900 × 900 visible and SAR images, each with a resolution of 0.5 m [32]. The scenes were captured in Rotterdam, the Netherlands. This dataset offers high-resolution and high-quality imagery, but features a large number of small-sized and densely distributed buildings. These characteristics render it particularly appropriate for assessing building extraction performance in dense urban environments.

3.1.2. DFC23 Track2 Dataset

The DFC23 track2 dataset includes 1773 pairs of 512 × 512 visible and SAR images for building extraction, with SAR images resampled to align with the visible resolution (0.5 m and 0.8 m) [33]. The DFC23 Track2 dataset exhibits a wide variety of building types with large intra-class variation. Additionally, due to differences in sensors and imaging conditions, some visible images in the dataset suffer from quality degradation, such as low contrast and thin cloud occlusion. These challenging conditions make the dataset suitable for assessing the robustness of building extraction models in complex scenes.

3.2. Evaluation Metrics

3.2.1. Accuracy Metrics

To evaluate building extraction performance, we employ four widely used metrics: Precision, Recall, F1-Score, and IoU.

- Precision indicates the fraction of correctly predicted building pixels among all pixels labeled as buildings.
- Recall captures the fraction of actual building pixels correctly detected by the model.
- F1-Score integrates Precision and Recall into a single measure, highlighting their balance via their harmonic mean.
- IoU evaluates spatial consistency by measuring the overlap ratio between predicted regions and ground truth annotations.

3.2.2. Computational Cost Metrics

In practical applications, model size, computational efficiency, and inference speed are crucial. We report three commonly used metrics: Parameters, GFLOPs, FPS and Inference Time.

- Parameters denote the count of trainable weights within the model, reflecting both storage requirements and architectural complexity. A smaller number of parameters typically indicates a more lightweight model.
- GFLOPs measure the computational workload per forward pass, offering a direct estimate of inference cost. Lower GFLOPs usually correspond to higher efficiency.
- FPS indicates the number of frames the model can process per second. In this paper, FPS is measured using an NVIDIA A5000 GPU.
- Inference Time refers to the per-frame latency of the model. For a given FPS value, the inference time τ (in milliseconds) is computed as $\tau = 1000/\text{FPS}$.

3.3. Implementation Details

3.3.1. Computing Facilities

High-Performance Training and Inference Platform

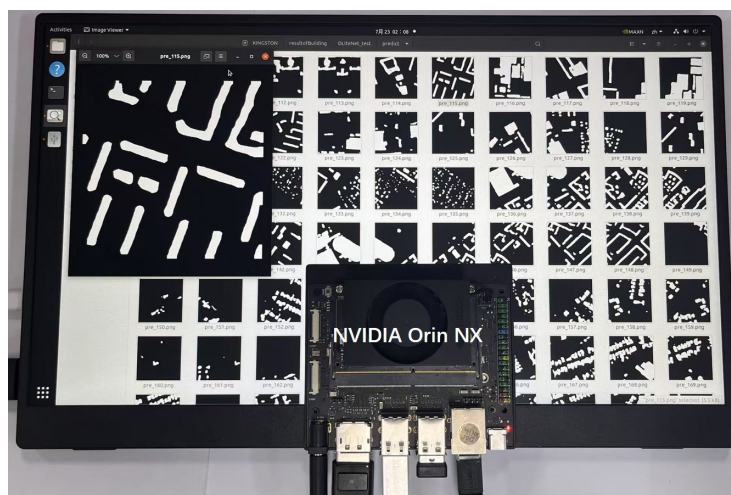
All model training and evaluation experiments were conducted on a high-performance workstation equipped with four NVIDIA RTX A5000 GPUs (24 GB each; NVIDIA Corp., Santa Clara, CA, USA) and an Intel Xeon Silver 4310 CPU (Intel Corp., Santa Clara, CA, USA).

Edge Inference Platform

To evaluate the practical deployment efficiency of the proposed model, especially under resource-constrained environments such as UAVs or satellite platforms, we performed edge inference tests on the NVIDIA Jetson Orin NX embedded computing platform (NVIDIA Corp., Santa Clara, CA, USA). The board and its hardware specifications are shown in Figure 2. Typical UAV-mounted or embedded platforms often run under strict resource limits, such as 10–25 W power budgets and 8–16 GB memory. Under such constraints, complex multimodal networks cannot sustain fast and efficient inference, which further motivates the lightweight design of DLiteNet.

3.3.2. Training and Inference Settings

In this study, building segmentation models were trained and evaluated on the MSAW dataset and the DFC23 Track2 dataset using an NVIDIA A5000 GPU (NVIDIA Corp., Santa Clara, CA, USA). Both datasets were split into training, validation, and testing sets following an 80:20 ratio. Specifically, the MSAW dataset consisted of 2770 images for training and 681 images for testing, while the DFC23 Track2 dataset contained 1418 training images and 355 testing images.



Module	Specification
CPU	8-core ARM Cortex-A78AE v8.2 64-bit
GPU	32 Tensor Cores + 1024 CUDA Cores
VRAM	16 GB 128-bit LPDDR5
Storage	M.2 SSD (expandable)
AI Performance	100 TOPS (INT8)

Figure 2. NVIDIA Jetson Orin NX board with operating screen and hardware specifications.

The experiments were implemented in the PyTorch 2.0.1 framework with CUDA 11.7. The network training utilized the Adam optimizer with an initial learning rate set to 1×10^{-3} . A StepLR scheduler was applied to adjust the learning rate, reducing it by a factor of 10 after 20 epochs. The total training spanned 100 epochs, with a batch size of 8 per epoch. This ensured that each model was fully trained until convergence.

Data augmentation was performed using random horizontal flipping to enhance model generalization. We did not employ additional strong augmentations such as arbitrary-angle rotation, large scaling, or brightness/contrast perturbations. Random horizontal flipping is a mild geometric transform that preserves local shapes and relative spatial structures while keeping the visible, SAR, and label maps strictly aligned. In contrast, aggressive geometric transforms (e.g., rotation and large scaling) may substantially alter the SAR imaging geometry and the apparent orientation of buildings, and naive photometric transforms on SAR can distort the physical meaning of backscatter values and the intensity relationships between visible and SAR.

3.4. Overall Network Architecture

As illustrated in Figure 3, most existing multimodal segmentation frameworks employ dual-stream deep encoders combined with heavy decoders, leading to substantial computational complexity and large model sizes. In this figure, the horizontal width of each rectangular block qualitatively reflects the number of convolutional layers in the corresponding module (i.e., its relative depth and parameter scale): wider blocks denote deeper and heavier components, whereas narrower blocks represent lightweight modules such as those used in DLiteNet.

To address these issues, we propose a novel dual-branch lightweight architecture, termed DLiteNet, which efficiently exploits the complementary characteristics of visible and SAR imagery through specialized processing branches, while significantly reducing model complexity.

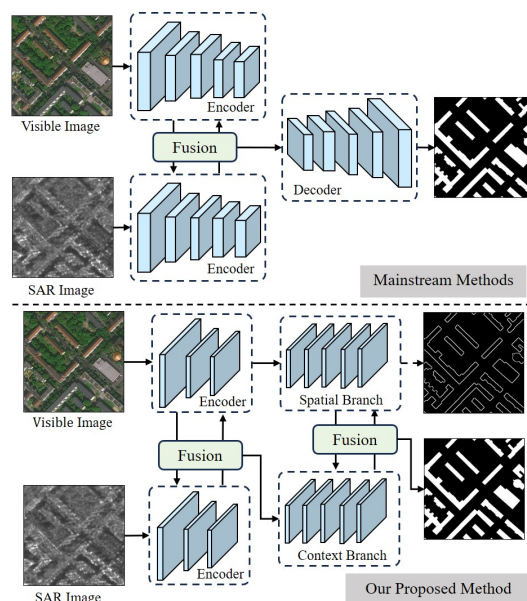


Figure 3. Schematic diagram comparing mainstream multimodal segmentation frameworks with DLiteNet. Mainstream methods use dual deep encoders with a large number of channels and cumbersome decoders, resulting in high computational complexity and bulky models. In contrast, DLiteNet achieves multimodal semantic fusion and detail extraction in lightweight context and spatial branches with fewer channels, omitting the decoder stage and achieving accurate segmentation results and lightweight model design.

As shown in Figure 4, DLiteNet consists of two parallel branches: the Context Branch and the Spatial Branch. By fully utilizing the complementary information from visible and SAR imagery, DLiteNet is capable of simultaneously capturing high-level semantic context and low-level spatial details essential for accurate building delineation.

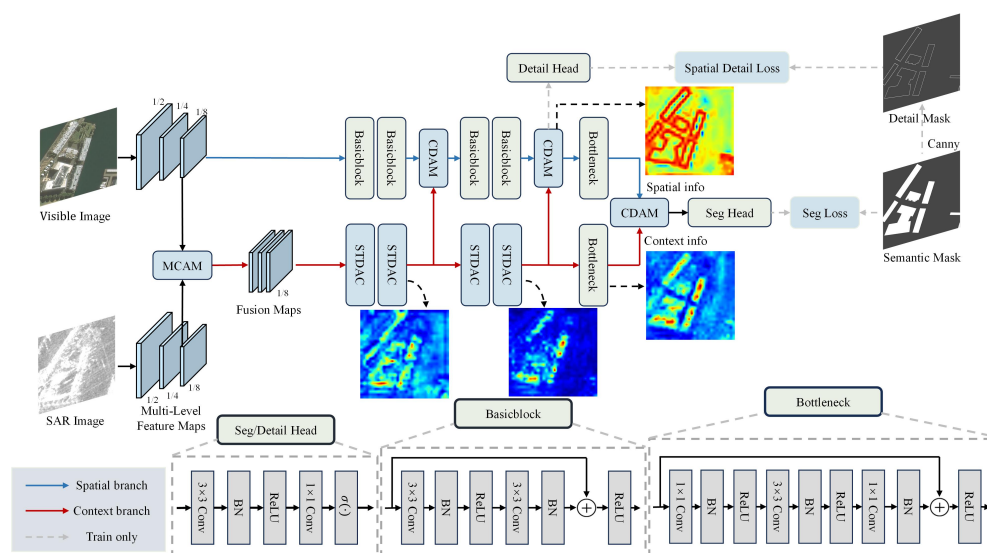


Figure 4. Overview of DLiteNet. DLiteNet consists of two lightweight branches—the context branch and the spatial branch—designed to balance building extraction accuracy and inference speed. The context branch takes both visible and SAR images as input to extract global contextual information using the lightweight STDAC and Bottleneck modules, capturing multi-scale context from both modalities. The spatial branch processes the visible image to extract fine-grained spatial details. During the detail aggregation process, contextual information is adaptively introduced through the proposed CDAM to guide the learning of structural building features. Black dashed arrows indicate intermediate feature maps.

Specifically, the Context Branch integrates modality-specific semantics from both inputs, while the Spatial Branch refines structural boundaries by focusing on fine-grained features derived from visible imagery. By adopting a dual-branch structure, DLiteNet effectively balances segmentation performance with computational efficiency.

Context Branch: The Context Branch is designed to capture high-level semantic context by aggregating information from both visible and SAR inputs. Visible imagery provides rich texture and structural cues, which are particularly effective for recognizing building outlines. In contrast, SAR imagery offers strong backscattering responses from vertical surfaces, making it especially sensitive to building facades and other man-made structures. To efficiently integrate this complementary information, we propose the Multi-modal Context Attention Module (MCAM) for balancing and fusing visible and SAR contexts. In addition, we design a lightweight multi-scale context extractor, the Short-Term Dense Atrous Concatenate (STDAC) module, which enables efficient multi-scale semantic aggregation with minimal computational overhead. For comparative evaluation and baseline strengthening, we also incorporate a standard Bottleneck module from ResNet [34], which facilitates deeper feature representation via residual learning. As shown by the intermediate feature maps (indicated by black dashed lines in Figure 4, this branch accurately captures the global layout and semantic boundaries of buildings.

Spatial Branch: The Spatial Branch is dedicated to capturing fine-grained spatial details from the visible modality, which typically contains richer structural information than SAR images. At the early stage of this branch, we adopt simple yet effective Basic Blocks [34]—a lightweight residual module widely used in efficient convolutional networks—to extract low-level texture and edge features. However, due to the limited representational capacity of such lightweight components, early-stage outputs often highlight generic edge structures (e.g., buildings, roads), lacking category-specific focus. To enhance the extraction of building-specific spatial details, we introduce cross-branch guidance from the Context Branch via the proposed Context-Detail Aggregation Module (CDAM). This guidance enables the spatial branch to selectively reinforce building-relevant contours while suppressing irrelevant edges. Additionally, we generate a Detail Mask from the ground-truth semantic mask using the Canny edge detector, which explicitly supervises the learning of building outlines. As illustrated in Figure 4, this branch effectively captures the precise planar contours of buildings, complementing the high-level semantics extracted by the context branch. The final predictions from both branches are fused by CDAM and refined through a lightweight segmentation head. The entire network is supervised using the ground truth Semantic Mask, ensuring accurate end-to-end training.

3.5. MCAM: Balancing the Context of Visible and SAR

We design an effective feature fusion mechanism called Multi-Scale Context Attention Module (MCAM) to selectively extract multi-scale contextual features from both visible and SAR images. This module enables adaptive learning of modality-specific contextual information, allowing the model to dynamically balance their contributions. MCAM consists of two key components: Multi-Scale Context Attention (MCA) and Complementary Selection Gate (CSGate).

Complementary Selection Gate (CSGate): The CSGate utilizes a gating strategy to simultaneously modulate information flow and capture complementary context from visible and SAR modalities. As shown in Figure 5, CSGate takes three inputs: (1) the contextual feature representation of visible images E_{VIS} ; (2) the SAR contextual features E_{SAR} ; and (3) the attention weights W_b learned from the MCA module.

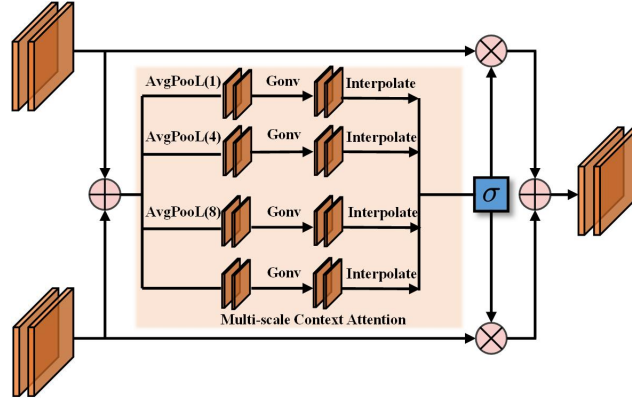


Figure 5. Illustration of the proposed MCAM. AvgPool denotes the global average pooling, σ represents the sigmoid function, and Interpolate indicates the upsampling operation implemented via interpolation.

The final fused contextual output E_r is formulated as:

$$E_r = \underbrace{E_{VIS} \odot \sigma(W_b)}_{\text{Selected Visible Context}} + \underbrace{E_{SAR} \odot \sigma(1 - W_b)}_{\text{Selected SAR Context}} \quad (1)$$

where $\sigma(\cdot)$ denotes the Sigmoid function, and $\odot$ represents element-wise multiplication. This operation allows the network to adaptively select complementary context features from both modalities.

Multi-Scale Context Attention (MCA): To capture broader contextual information and enrich feature diversity, we first construct a fused feature map by summing the visible and SAR features:

$$F_{VS} = F_{VIS} + F_{SAR} \quad (2)$$

Then, F_{VS} is passed through average pooling operations at multiple scales to extract context at different receptive fields. Specifically, we apply average pooling with kernel sizes of 1×1 , 4×4 , and 8×8 to obtain the multi-scale representations F_{c1} , F_{c2} , and F_{c3} , respectively. The original feature map F_{VS} is also retained as F_{c4} . Specifically, average pooling operations with kernel sizes of 1×1 , 4×4 , and 8×8 are applied to F_{VS} to obtain the multi-scale feature representations $F_{c1} \in \mathbb{R}^{C \times 1 \times 1}$ (global context), $F_{c2} \in \mathbb{R}^{C \times 4 \times 4}$ (mid-level context), and $F_{c3} \in \mathbb{R}^{C \times 8 \times 8}$ (local context), respectively. The original resolution feature map is retained as $F_{VS} \in \mathbb{R}^{C \times H \times W}$ to preserve fine-grained spatial details.

Here, $\text{Gonv}(\cdot)$ denotes a lightweight channel compression–expansion block implemented with two 1×1 convolutions:

$$\text{Gonv}(X) = \text{ReLU}(\text{BN}_2(\text{Conv}_{1 \times 1}^{(2)}(\text{ReLU}(\text{BN}_1(\text{Conv}_{1 \times 1}^{(1)}(X))))), \quad (3)$$

where $\text{Conv}_{1 \times 1}^{(1)}$ reduces the channel dimension from C to C/r , $\text{Conv}_{1 \times 1}^{(2)}$ restores it from C/r back to C , BN_1 and BN_2 are batch normalization layers, $\text{ReLU}(\cdot)$ is the activation function, and the reduction ratio r is set to 4 in our implementation. This design introduces non-linear channel interactions while keeping the computational overhead very small. These multi-scale contextual features are each processed through a lightweight convolutional projection (denoted as Gonv) [35], upsampled to the original resolution via nearest-neighbor interpolation (denoted as Up), and then aggregated to form the final attention weight map:

$$W_b = \sum_{i \in \{1, 4, 8\}} \text{Up}(\text{Gonv}(\text{AvgPool}_i(F_{VS}))) + \text{Gonv}(F_{VS}) \quad (4)$$

where, $\text{AvgPool}_i(F_{VS})$ refers to applying average pooling with kernel size $i \times i$ to F_{VS} , and $\text{Conv}(\cdot)$ represents the channel compression and expansion operations designed to capture non-linear cross-modal interactions at different scales.

3.6. STDAC: Short-Term Dense Atrous Concatenate

Inspired by the Short-Term Dense Concatenate (STDC) structure [18], we design a semantic segmentation-oriented contextual feature extraction module called Short-Term Dense Atrous Concatenate (STDAC), as illustrated in Figure 6. The STDAC module consists of a 1×1 convolution followed by three 3×3 atrous convolutions with different dilation rates, enabling the model to capture contextual information of buildings from progressively larger receptive fields.

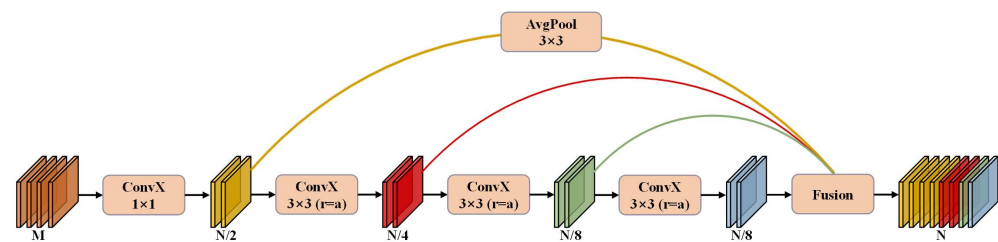


Figure 6. Illustration of the proposed STDAC. Where $r = a$ indicates the atrous rate = a .

The processing at each stage is defined as:

$$x_{i+1} = \text{ConvX}_i(x_i, k_i) \quad (5)$$

where x_i denotes the i -th feature map in the module, ConvX_i represents a sequence of a convolutional layer, batch normalization, and ReLU activation, and k_i indicates the kernel size and dilation rate of the convolution operation.

The input feature map, which encodes building-related contextual information from both visible and SAR modalities, is first projected to a suitable channel dimension using a 1×1 convolution. Subsequently, it undergoes three parallel atrous convolutional operations with varying dilation rates, enabling an expanded receptive field and capturing features at multiple scales. To reduce the number of parameters, the feature maps are progressively compressed into sizes of $N/2$, $N/4$, and $N/8$ channels. Finally, the extracted multi-scale contextual features are fused through a lightweight aggregation function F to obtain the final representation:

$$x_6 = F(x_2, x_3, x_4, x_5) \quad (6)$$

where x_6 denotes the output feature map of the STDAC module, and $F(\cdot)$ represents a fusion operation that combines multi-scale features from different levels to enhance semantic representation.

We adopt the STDAC module in the context branch because the extraction of object edges requires only limited local information, which can be handled by simple gradient operators (e.g., Sobel). However, the accurate recognition of buildings necessitates more abstract and high-level contextual understanding, which benefits from larger receptive fields. STDAC thus effectively balances multi-scale context aggregation and model efficiency.

3.7. CDAM: Fast Aggregation of Contexts and Details

We design a simple yet effective module called the Context-Detail Aggregation Module (CDAM) to efficiently fuse contextual and spatial detail information. This module provides a lightweight mechanism to combine feature maps rich in spatial details with those rich in building-related contextual semantics. Specifically, we use confidence maps generated

from both context and spatial branches to guide the network in dynamically adjusting the feature fusion strength across different spatial regions.

As illustrated in Figure 7, given the spatial feature map S and context feature map C from their respective branches, the CDAM is designed to fill high-frequency regions with spatial details and low-frequency regions with semantic context.

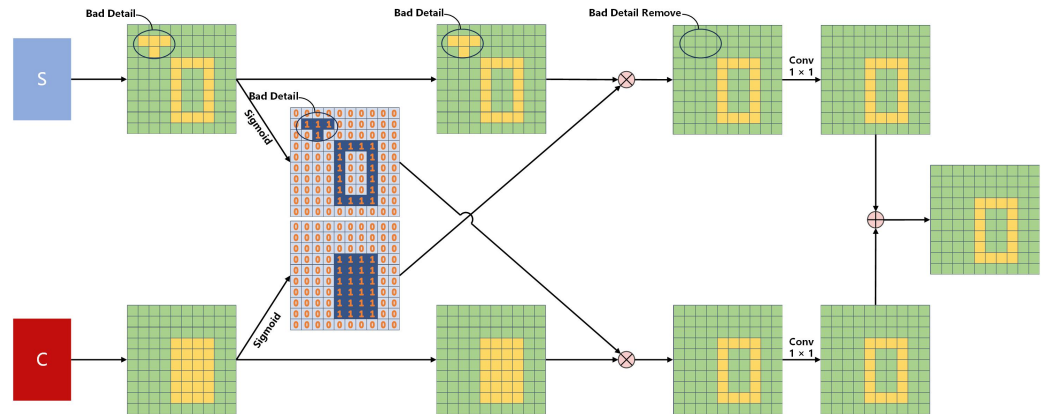


Figure 7. Illustration of the CDAM. “S” denotes the Spatial Branch, “C” denotes the Context Branch, and “Sigmoid” indicates the application of the Sigmoid activation function. The figure presents representative examples under extreme conditions.

The context branch typically produces semantically accurate predictions but often loses the structural details of buildings, such as precise boundaries and planar shapes. In contrast, the spatial branch preserves fine-grained spatial information. Therefore, we encourage the model to rely more on the spatial features near object boundaries, while allowing the context features to dominate in the interior or semantically consistent areas.

Let $\vec{v}_s$ and $\vec{v}_c$ denote the vector representations (per pixel) of spatial and context features, respectively. The confidence weights σ_s and σ_c are computed through a Sigmoid activation applied to each vector. Since the Sigmoid is applied independently at every spatial location, the confidence maps σ_s and σ_c are spatially varying and their values are constrained to $[0, 1]$, which enables CDAM to softly modulate the relative contribution of the two branches at each pixel. The fused output feature F_{sc} is computed as follows:

$$\sigma_s = \text{Sigmoid}(\vec{v}_s) \quad (7)$$

$$\sigma_c = \text{Sigmoid}(\vec{v}_c) \quad (8)$$

$$F_{sc} = f(\vec{v}_s \otimes \sigma_c) \oplus f(\vec{v}_c \otimes \sigma_s) \quad (9)$$

where $f(\cdot)$ denotes a composite operation consisting of convolution, batch normalization, and ReLU activation; $\otimes$ represents element-wise multiplication; and $\oplus$ denotes element-wise addition.

This formulation allows the model to adaptively balance spatial and contextual cues: when σ_s is high and σ_c is low, the model favors spatial features, and vice versa. In practice, σ_s tends to be larger near building boundaries where high-frequency structures dominate, while σ_c becomes larger inside homogeneous building regions and background areas where semantic context is more reliable, so CDAM behaves like a spatially adaptive mixture-of-experts between the spatial and context branches. As a result, CDAM effectively integrates spatial precision and semantic richness for better building extraction performance.

3.8. Loss Functions

Spatial Detail Loss: Since the number of pixels representing the actual planar contours of buildings is significantly smaller than that of non-contour pixels, the supervision of planar structure learning becomes a class-imbalance problem. To address this, we adopt a combination of Binary Cross-Entropy (BCE) loss and Dice loss [36] to jointly optimize the contour prediction. Dice loss measures the overlap between the prediction and ground truth and is insensitive to the number of foreground/background pixels, which helps mitigate the imbalance issue.

Let H and W denote the height and width of the predicted contour map. The Spatial Detail Loss $\mathcal{L}_{\text{Detail}}$ is defined as:

$$\mathcal{L}_{\text{Detail}}(y_d, \hat{y}_d) = \mathcal{L}_{\text{Dice}}(y_d, \hat{y}_d) + \mathcal{L}_{\text{BCE}}(y_d, \hat{y}_d) \quad (10)$$

where $y_d \in \mathbb{R}^{H \times W}$ denotes the predicted contour of buildings, and $\hat{y}_d \in \mathbb{R}^{H \times W}$ represents the corresponding ground truth.

The Dice loss term is calculated as:

$$\mathcal{L}_{\text{Dice}}(y_d, \hat{y}_d) = 1 - \frac{2 \sum_{i=1}^{H \times W} y_d^i \hat{y}_d^i + \varepsilon}{\sum_{i=1}^{H \times W} (y_d^i)^2 + \sum_{i=1}^{H \times W} (\hat{y}_d^i)^2 + \varepsilon} \quad (11)$$

where i denotes the pixel index and ε is a Laplace smoothing constant to avoid division by zero, typically set to $\varepsilon = 1$.

As illustrated in Figure 4, we generate a Detail Map using the Canny operator to guide the shallow layers in encoding spatial information. This Detail Map is used to supervise the output of the Detail Head for accurate boundary refinement. In our experiments, the Detail Head has been shown to significantly enhance feature representation. It is also worth noting that the Detail Head is discarded during inference, meaning that this auxiliary information can effectively improve segmentation accuracy without introducing additional inference cost.

In our preliminary experiments, we first trained the detail head using BCE loss only and observed that the spatial detail loss almost stopped decreasing during optimization. Closer inspection of the Canny-based detail maps showed that the supervision is extremely sparse: building pixels are already much fewer than background pixels, and edge pixels account for less than 1% of all pixels in an image. Under such severe class imbalance, BCE-only supervision tends to drive the network towards a degenerate solution where almost all pixels are predicted as background and the rare edge pixels receive very weak gradients. To alleviate this issue, we combine BCE and Dice loss for spatial detail supervision. The Dice term explicitly optimizes the overlap between predicted and ground-truth contours and provides stronger gradients for these very small positive regions, which stabilizes training and leads to noticeably better boundary preservation in our experiments.

Seg Loss: For building segmentation, predictions are generated by the Seg Head and supervised by the ground truth semantic mask of buildings. We employ binary cross-entropy (BCE) loss to optimize the learning process. The segmentation loss L_{Seg} is defined as:

$$L_{\text{Seg}}(y_s, \hat{y}_s) = L_{\text{BCE}}(y_s, \hat{y}_s) \quad (12)$$

where $y_s \in \mathbb{R}^{H \times W}$ denotes the predicted semantic segmentation map for buildings and $\hat{y}_s \in \mathbb{R}^{H \times W}$ denotes the mask of ground truth.

Overall Loss: Following the design of auxiliary detail supervision in real-time semantic segmentation networks proposed by Fan et al. [18], we simply sum the two terms to form the overall training objective:

$$L_{\text{Total}} = L_{\text{Seg}} + L_{\text{Detail}}. \quad (13)$$

4. Results

4.1. Results on DFC23 Track2 Dataset

To comprehensively evaluate the performance of the proposed DLiteNet, we conduct experiments on the DFC23 Track2 test set, which is characterized by diverse building types, significant intra-class variability, and challenging imaging conditions. In particular, the visible images suffer from sensor limitations and degraded image quality, which poses considerable challenges for accurate building extraction. These complex scenarios provide a rigorous benchmark for assessing the robustness and generalization capability of segmentation models. For comparison, we evaluate DLiteNet against both single-modal and multi-modal methods, including several lightweight networks. Given the high computational cost and architectural complexity of Transformer-based models, deploying such methods on edge computing platforms remains impractical. Consequently, we focus our comparative analysis on convolutional neural network (CNN)-based methods to highlight DLiteNet’s advantages in terms of efficiency and deployability. Meanwhile, several efficient Transformer variants such as SegFormer-B0 have been proposed to reduce the complexity of vision Transformers. However, SegFormer-B0 still requires about 8.4 GFLOPs and 3.7 M parameters for 512×512 inputs and is primarily designed for single-modality Visible segmentation rather than Visible–SAR fusion. In this work, we therefore focus on a lightweight CNN-based dual-branch architecture that is explicitly tailored for Visible–SAR building extraction and edge deployment. Table 1 summarizes the quantitative results of DLiteNet compared with several baseline models, including single-modal and multi-modal approaches.

Table 1. Quantitative results on the DFC23 track2 datasets.

Method	Modality	Precision↑	Recall↑	IoU↑	F1-Score↑	GFLOPs↓	Params (M)↓	Fps↑	Inf. Time (ms)↓
U-Net [9]	Visible	82.9	73.0	63.5	77.7	124.7	31.0	77.12	12.97
GCNet [37]	Visible	86.2	88.1	77.1	87.1	183.3	85.2	41.34	24.19
DANet [38]	Visible	87.3	86.6	76.9	86.9	500.6	46.2	26.54	37.68
U-Net++ [39]	Visible	87.8	87.0	77.6	87.4	800.5	45.1	16.89	59.21
DeepLabv3+ [40]	Visible	89.3	90.6	81.7	89.9	48.1	39.0	163.69	6.11
HRNet [41]	Visible	90.1	89.8	81.7	89.9	161.8	70.0	42.82	23.35
U-KAN [42]	Visible	89.6	88.0	79.9	88.8	27.6	9.4	39.14	25.55
BiSeNet [16]	Visible	86.2	84.8	74.7	85.5	40.7	23.1	382.68	2.61
BiSeNetV2 [17]	Visible	87.3	88.6	78.5	87.9	38.0	21.6	131.84	7.58
DDRNet-Slim [19]	Visible	85.4	83.1	72.8	84.2	4.53	5.7	258.23	3.87
U-Net [9]	SAR	44.3	45.7	29.1	45.0	124.4	31.0	77.16	12.96
DeepLabv3+ [40]	SAR	63.8	51.6	39.9	57.1	47.3	39.0	163.78	6.11
HRNet [41]	SAR	52.2	43.0	30.9	47.2	161.7	70.0	42.90	23.31
RedNet [43]	Visible + SAR	87.2	87.7	77.7	87.4	84.5	81.9	71.48	13.99
SA-Gate [44]	Visible + SAR	88.8	87.4	78.7	88.1	276.8	53.4	37.70	26.53
CMGFNet [45]	Visible + SAR	91.2	79.1	73.5	84.7	155.3	85.2	76.76	13.03
MCANet [5]	Visible + SAR	89.6	90.5	81.9	90.1	375.9	71.2	39.98	25.01
DLiteNet (Ours)	Visible + SAR	90.2	90.5	82.3	90.3	5.8	5.6	80.28	12.46

Note: For metrics marked with “↑”, higher is better; for those with “↓”, lower is better. **Red bold** = best, **Blue bold** = second-best.

When comparing with single-modal models, the results clearly demonstrate the benefits of multi-modal fusion. Models using only visible images as input achieve better segmentation performance than those relying solely on synthetic aperture radar (SAR) images, owing to the richer spectral and texture information present in the visible modality.

The best-performing single-modal visible-based model achieves an IoU of 81.7% and an F1-score of 89.9%. By contrast, SAR-only models obtain significantly lower performance, with the best IoU and F1-score recorded at 39.9% and 57.1%, respectively. Both single-modal models underperform relative to DLiteNet, highlighting the value of integrating complementary information from visible and SAR data. In comparison with existing multi-modal methods, DLiteNet achieves competitive or superior segmentation performance with significantly lower computational complexity. While prior multi-modal networks benefit from using dual-modality inputs, they generally rely on deep two-stream architectures, leading to large parameter counts and high computational overhead. For example, RedNet requires 84.5 GFLOPs, and MCANet involves 71.2 M parameters. Such models are less suitable for edge deployment. DLiteNet, by contrast, requires only 5.8 GFLOPs and 5.6 M parameters, achieving an IoU of 82.3% and an F1-score of 90.3%. The model also attains an inference speed of 80.28 FPS on an NVIDIA RTX A5000 GPU. In summary, multimodal baselines such as SA-Gate (53.4 M parameters, 276.8 GFLOPs), CMGFNet (85.2 M parameters, 155.3 GFLOPs), and MCANet (71.2 M parameters, 375.9 GFLOPs) remain substantially heavier than DLiteNet (5.6 M parameters, 5.8 GFLOPs), even though they target the same building extraction task. These results confirm that DLiteNet achieves a favorable balance between segmentation accuracy and computational efficiency, supporting its applicability in resource-constrained environments. To further highlight the architectural differences and computational costs of representative methods, Table 2 summarizes several typical single-modal and multimodal CNN-based networks on the DFC23 Track2 and MSAW dataset in terms of their modality usage, core designs, parameter counts, GFLOPs, and IoU.

To further validate the quantitative findings, representative qualitative results are illustrated in Figure 8. Four typical urban scenes with diverse building structures and imaging conditions are selected to compare DLiteNet with baseline models. Each row presents the visible image, SAR image, ground truth annotation, and prediction maps generated by competing methods.

In Scene (I), where the visible image is affected by low quality and spectral similarity between roads and buildings, U-Net and other CNN-based models exhibit substantial omission errors, missing large building areas. Although methods such as DeepLabv3+, HRNet, MCANet, and SA-Gate partially reduce these errors through contextual modeling, they still misclassify road regions as buildings. In comparison, DLiteNet effectively suppresses both false positives and false negatives by leveraging complementary context and spatial cues. Scene (II) involves a large building with heterogeneous roof textures and materials. While other models struggle with fragmented segmentation, DLiteNet produces a more consistent prediction across the entire structure, benefiting from its dual-branch architecture and edge-guided refinement. Scene (III) highlights DLiteNet's boundary delineation capability. The model accurately restores object contours, outperforming baseline methods that yield blurred or incomplete boundaries. In Scene (IV), small buildings with weak texture present challenges for both single-modal and existing multi-modal models. In contrast, DLiteNet successfully detects these small structures by integrating backscatter information from SAR imagery with fine spatial details from the visible image. Scene (V) presents a cloud- and haze-contaminated visible image, where large regions of buildings are partially obscured. In this case, visible-only models tend to miss or misclassify buildings inside the cloud region. In contrast, DLiteNet effectively leverages complementary SAR backscatter to recover the underlying building footprints within the cloud. Overall, DLiteNet demonstrates enhanced robustness, improved boundary localization, and superior small-object detection capability under diverse urban conditions. Compared with state-of-the-art multi-modal and lightweight networks, DLiteNet consistently achieves better segmentation accuracy

with reduced computational burden, confirming its practical value for real-world building extraction applications.

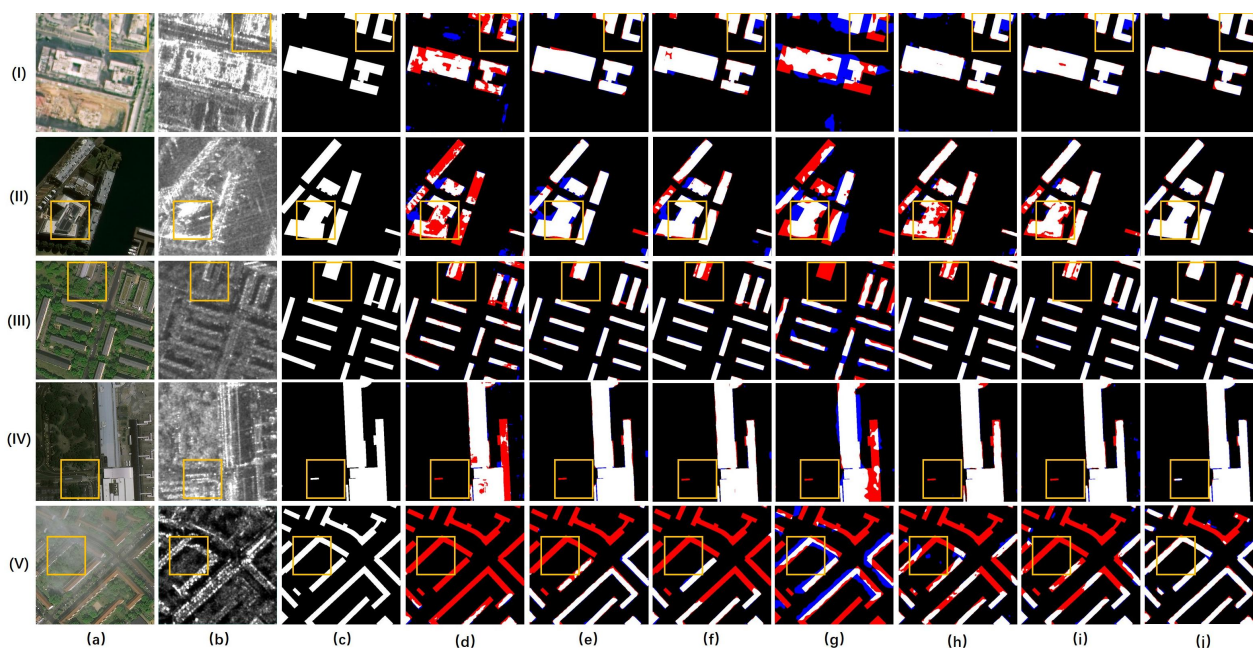


Figure 8. Building extraction results on the DFC23 track2 dataset. The yellow boxes highlight the zoom-in regions of interest (ROIs) that are further analyzed in the following discussion. The labels I, II, III, IV, and V correspond to five representative scenes selected from the dataset. (a) Visible image. (b) SAR image. (c) Ground Truth. (d) U-Net (Visible). (e) DeepLabv3+ (Visible). (f) HRNet (Visible). (g) DeepLabv3+ (SAR). (h) MCANet. (i) SA-Gate. (j) DLiteNet (Ours). The color coding is as follows: black indicates true negatives, white indicates true positives, blue represents false positives, and red denotes false negatives.

Table 2. Summary of representative methods for building extraction on the DFC23 Track2 and MSAW datasets.

Dataset	Method	Modality	Core Design	Params (M)	GFLOPs	IoU (%)
DFC23 Track2	BiSeNet	Visible	two-branch lightweight CNN	23.1	40.7	74.7
	BiSeNetV2	Visible	two-branch lightweight CNN	21.6	38.0	78.5
	DDRNet-Slim	Visible	two-branch lightweight CNN	5.7	4.53	72.8
	RedNet	Visible + SAR	dual-encoder multimodal CNN with decoder	81.9	84.5	77.7
	SA-Gate	Visible + SAR	dual-encoder multimodal CNN with decoder	53.4	276.8	78.7
	CMGFNet	Visible + SAR	dual-encoder multimodal CNN with decoder	85.2	155.3	73.5
	MCANet	Visible + SAR	dual-encoder multimodal CNN with decoder	71.2	375.9	81.9
	DLiteNet (Ours)	Visible+SAR	dual-branch lightweight multimodal fusion	5.6	5.8	82.3
MSAW	BiSeNet	Visible	two-branch lightweight CNN	23.1	128.4	76.8
	BiSeNetV2	Visible	two-branch lightweight CNN	21.7	118.8	77.2
	DDRNet-Slim	Visible	two-branch lightweight CNN	5.7	14.2	70.3
	SA-Gate	Visible + SAR	dual-encoder multimodal CNN with decoder	53.4	849.7	79.0
	CMGFNet	Visible + SAR	dual-encoder multimodal CNN with decoder	85.2	490.9	73.6
	MCANet	Visible + SAR	dual-encoder multimodal CNN with decoder	71.2	874.5	82.6
	DLiteNet (Ours)	Visible + SAR	dual-branch lightweight multimodal fusion	5.6	51.7	83.6

4.2. Results on MSAW Dataset

In this section, we evaluate the proposed DLiteNet on the MSAW test set, which contains high-resolution visible and SAR images with excellent imaging quality. However, the presence of numerous compact and closely packed buildings significantly complicates accurate extraction in dense urban areas. To assess the effectiveness of our approach in such scenarios, DLiteNet is compared against representative single-modal, multi-modal, and lightweight CNN-based models. Considering practical deployment requirements,

Transformer-based models are excluded from this comparison. The experimental results are summarized in Table 3.

Table 3 presents the quantitative results of DLiteNet and baseline models. Among single-modal methods using only visible images, HRNet achieves the best performance with an IoU of 78.4% and an F1-score of 87.9%. When using SAR images alone, DeepLabv3+ achieves the highest IoU of 64.3% and an F1-score of 78.3%. This performance gap highlights the advantages of visible data in providing detailed spectral and texture information, while SAR-only models suffer from the absence of such cues. Nevertheless, both single-modal approaches perform significantly worse than DLiteNet, confirming the superiority of multi-modal fusion in exploiting complementary information from visible and SAR imagery. In terms of multi-modal methods, MCANet achieves the strongest baseline performance, attaining an IoU of 82.6% and an F1-score of 90.5%. Other approaches such as SA-Gate and CMGFNet also yield competitive results. However, these models generally involve heavy dual-stream architectures, leading to high computational complexity. Specifically, MCANet requires 874.5 GFLOPs and 71.2 M parameters, while CMGFNet demands 490.9 GFLOPs with 85.2 M parameters. Such large resource consumption severely limits their applicability for edge computing. By contrast, DLiteNet achieves the best overall performance, with an IoU of 83.6% and an F1-score of 91.1%, while maintaining only 51.7 GFLOPs and 5.6 M parameters. Additionally, it delivers the fastest inference speed of 76.5 FPS. These results demonstrate that DLiteNet delivers a competitive balance between performance and efficiency, supporting its deployment in resource-constrained remote sensing scenarios.

Table 3. Quantitative Comparison of Building Extraction Methods on the MSAW Dataset.

Method	Modality	Precision↑	Recall↑	IoU↑	F1-Score↑	GFLOPs↓	Params (M)↓	Fps↑	Inf. Time (ms)↓
U-Net [9]	Visible	84.6	80.0	69.8	82.2	494.0	31.0	22.32	44.80
GCNet [37]	Visible	86.9	82.3	73.2	84.6	578.8	85.2	17.56	56.95
DANet [38]	Visible	85.9	81.0	71.5	83.4	1552.2	46.2	7.18	139.28
DeepLabv3+ [40]	Visible	87.8	86.4	77.1	87.1	151.8	39.0	51.91	19.26
HRNet [41]	Visible	88.6	87.2	78.4	87.9	504.2	70.0	15.35	65.15
U-KAN [42]	Visible	86.5	85.0	75.0	85.7	85.0	9.4	13.79	72.52
BiSeNet [16]	Visible	87.4	86.3	76.8	86.9	128.4	23.1	145.06	6.89
BiSeNetV2 [17]	Visible	86.5	87.8	77.2	87.1	118.8	21.7	70.94	14.10
DDRNet-Slim [19]	Visible	83.9	81.3	70.3	82.6	14.2	5.7	221.72	4.51
U-Net [9]	SAR	46.6	53.7	33.2	49.9	493.1	31.0	22.42	44.60
DeepLabv3+ [40]	SAR	79.3	77.3	64.3	78.3	149.2	39.0	51.98	19.24
HRNet [41]	SAR	66.3	69.4	51.3	67.9	503.2	70.0	15.39	64.98
SA-Gate [44]	Visible + SAR	89.0	87.6	79.0	88.3	849.7	53.4	14.82	67.48
CMGFNet [45]	Visible + SAR	88.4	81.5	73.6	84.8	490.9	85.2	23.24	43.03
MCANet [5]	Visible + SAR	91.0	89.9	82.6	90.5	874.5	71.2	15.47	64.64
DLiteNet (Ours)	Visible + SAR	91.7	90.4	83.6	91.1	51.7	5.6	76.51	13.07

Note: For metrics marked with “↑”, higher is better; for those with “↓”, lower is better. **Red bold** = best, **Blue bold** = second-best.

To further verify the quantitative results, Figure 9 shows representative visual comparisons on the test set, including visible images, SAR images, ground truth masks, and predicted building maps. Yellow bounding boxes highlight critical regions containing small or densely packed buildings and fine structural details.

As illustrated in Figure 9, single-modal methods often fail to detect small-scale buildings or preserve complete contours, particularly within densely clustered areas. For instance, in Scenes (II) and (III), HRNet and DeepLabv3+ frequently miss small buildings. Multi-modal baselines such as MCANet and SA-Gate achieve improved results but still produce fragmented predictions in compact areas due to insufficient edge modeling. In contrast, DLiteNet effectively integrates spatial detail from visible images with structural stability from SAR backscatter, leading to more accurate detection of small buildings and densely distributed structures. Additionally, Scenes (I) and (IV) demonstrate DLiteNet’s

superior edge delineation, producing continuous and sharp building boundaries where competing models exhibit blurred or broken contours. Scene (V) focuses on two small but spatially separated buildings within a cluttered background. Several baseline methods mistakenly merge them into a single connected region, whereas DLiteNet is able to stably separate and correctly detect both buildings, as indicated by the bounding box in Figure 9. Scene (VI) contains a large building with a complex and irregular footprint. In this scenario, competing methods tend to produce fragmented or jagged predictions along the boundary, whereas DLiteNet provides a more coherent segmentation mask that better preserves the global shape and fine boundary details of the building.

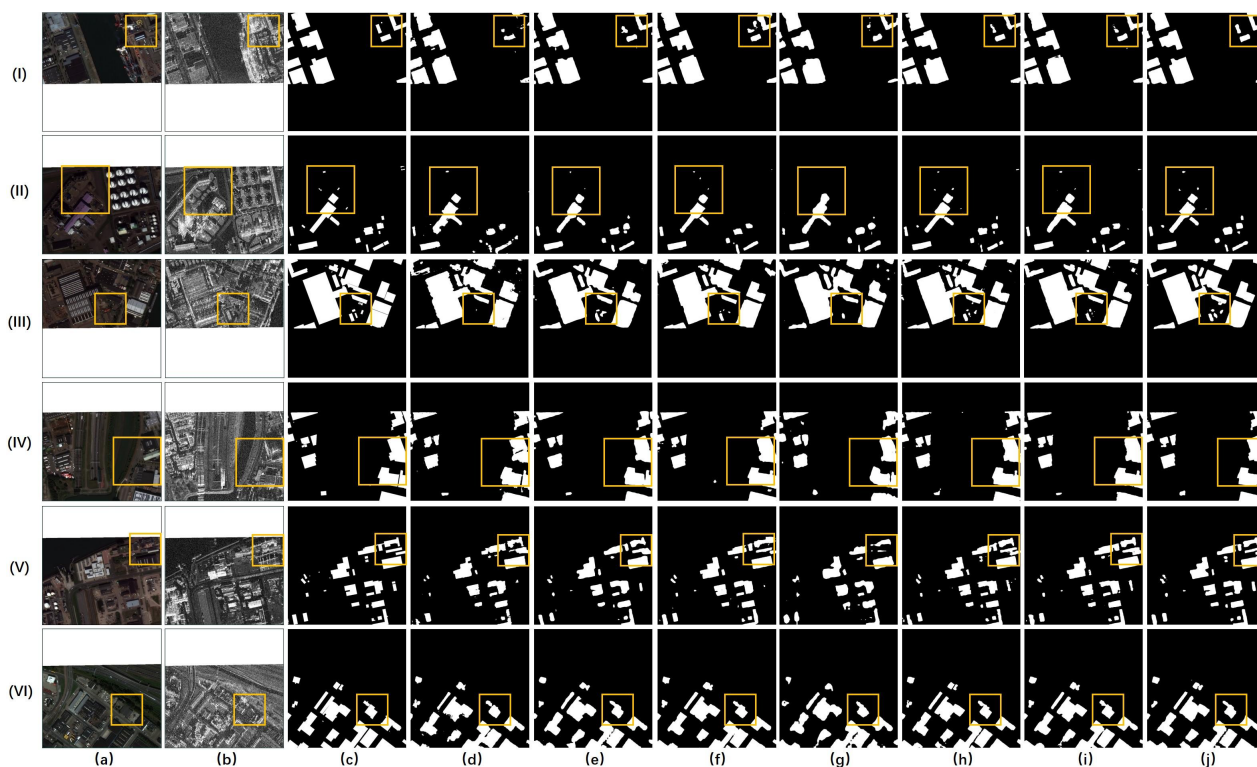


Figure 9. Building extraction results on the MSAW dataset. The yellow boxes highlight the zoom-in regions of interest (ROIs) that are further analyzed in the subsequent discussion. The labels I, II, III, IV, V, and VI correspond to six representative scenes selected from the dataset. (a) Visible image. (b) SAR image. (c) Ground Truth. (d) U-Net (Visible). (e) DeepLabv3+ (Visible). (f) HRNet (Visible). (g) DeepLabv3+ (SAR). (h) MCANet. (i) SA-Gate. (j) DLiteNet (Ours).

4.3. Deployment Results on Edge Platform (Jetson Orin NX)

To evaluate the practical deployability of the proposed DLiteNet model, we deployed it on the NVIDIA Jetson Orin NX 16GB embedded platform. Operating under the board's MAXN performance mode, the model was optimized using TensorRT with both FP16 and INT8 quantization strategies. These optimizations aim to reduce model size and inference latency, enabling real-time performance on resource-constrained edge devices. Table 4 summarizes the engine sizes and inference latencies of the DLiteNet model under different quantization settings, using a 512×512 visible and SAR image pair as input (batch size = 1).

Table 4. DLiteNet performance on Jetson Orin NX.

Quantization	Model Size (MB)	Inf. Time (ms)
FP32	27.4	14.97
FP16	16.6	12.33
INT8	11.0	11.18

As shown in Table 4, the FP32 model without quantization has an engine size of 27.4 MB and requires approximately 14.97 ms to infer a pair of 512×512 visible and SAR images on the Jetson Orin NX. After applying TensorRT FP16 quantization, the model size reduces to 16.6 MB and the inference time decreases to 12.33 ms. Further quantizing the model to INT8 reduces the model size to 11.0 MB and shortens the inference time to 11.18 ms. This performance corresponds to processing approximately 90 pairs of 512×512 visible and SAR images per second on the edge device, demonstrating the suitability of DLiteNet for real-time applications on resource-constrained platforms.

5. Discussion

In this section, we discuss the robustness and effectiveness of the three key modules in the proposed DLiteNet, including MCAM, CDAM, and STDAC. Each module is analyzed in terms of its contribution to the overall performance and its role in enhancing multimodal feature fusion, context–spatial interaction, and efficient representation learning.

5.1. Effectiveness of MCAM

As reported in Table 5, integrating the proposed MCAM improves the baseline IoU from 78.8% to 80.1% and the F_1 -score from 88.1% to 89.0%. By enabling cross-modal interactions between visible and SAR features, MCAM facilitates the fusion of complementary structural and spectral information. To further demonstrate the effectiveness of MCAM, we visualize the intermediate features from the visible and SAR branches as well as the fused features after MCAM in Figure 10. The visible branch alone fails to consistently highlight building regions due to intra-class appearance variations, while the SAR branch suffers from structural noise, resulting in incomplete building contours. In contrast, the fused feature map from MCAM clearly focuses on the complete building regions and suppresses irrelevant noise, evidencing its semantic enhancement capability.

On the DFC23 Track2 Dataset, we further measure the compactness of building-region features by the average cosine similarity to the class center. A baseline with simple Visible–SAR concatenation yields a mean similarity of 0.826, whereas the MCAM-fused features increase this value to 0.874, indicating more compact and semantically consistent representations for building structures.

Table 5. Ablation analysis of MCAM and CDAM modules on the DFC23 Track2 dataset. “✓” indicates the module is used.

Dataset	MCAM	CDAM	Precision (%)↑	Recall (%)↑	IoU (%)↑	F1-Score (%)↑
DFC23 Track2	✓		89.2	87.1	78.8	88.1
			89.0	88.9	80.1	88.9
		✓	88.9	89.1	80.2	89.0
	✓	✓	90.2	90.5	82.3	90.3

Note: “✓” indicates the corresponding module is enabled. For metrics marked with “↑”, higher values indicate better performance. **Red bold** denotes the best performance across all configurations.

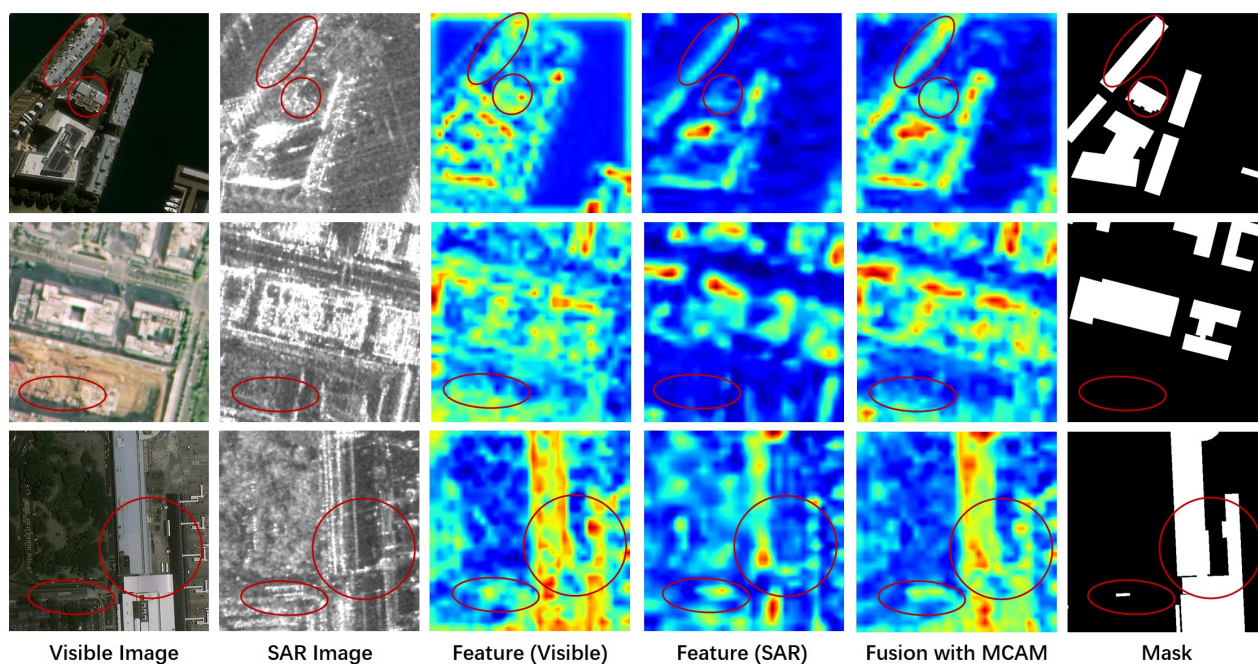


Figure 10. Feature heatmaps of the visible and SAR branches before and after MCAM fusion. The color intensity reflects the activation strength of the extracted features, where warmer colors correspond to higher responses. The visible and SAR images provide complementary yet incomplete features. As shown in the heatmaps, the visible branch fails to fully focus on building regions (elliptical red box), while the SAR branch lacks complete semantic structure (circular red box). After fusion through the MCAM, the building semantics from both modalities are correctly aligned, achieving complementarity and resulting in features that are more consistent with the final segmentation mask.

5.2. Effectiveness of CDAM

When the CDAM module is added to the baseline, the model benefits from joint edge and semantic reasoning, achieving 80.2% IoU and 89.0% F1-score. We visualize the feature maps from the context branch and the spatial branch in Figure 11. The context branch captures high-level semantic information using a lightweight structure and produces heatmaps with broad activation over building areas. However, these activations are often scattered and lack structural clarity. In contrast, the spatial branch effectively preserves fine-grained boundary and shape details, generating heatmaps with sharp and continuous contours even in complex scenes. The fusion of both branches combines global semantic reasoning with precise spatial localization, yielding a final activation map that is both semantically coherent and geometrically accurate. These results demonstrate the individual contribution of the CDAM module. Moreover, the combined use of MCAM and CDAM modules leads to further improvements. Our complete DLiteNet model, equipped with both modules, achieves state-of-the-art performance with an IoU of 82.3% and an F1-score of 90.3%. More specifically, the context-branch heatmaps exhibit strong responses inside large building footprints and homogeneous regions, indicating that this branch focuses on stabilizing semantic predictions and suppressing background clutter. The spatial-branch heatmaps, on the other hand, show peaked responses along building edges, corners and small protrusions, revealing that this branch is dedicated to capturing high-frequency boundary information. After CDAM fusion, the resulting activation maps simultaneously highlight building interiors and outlines while damping responses in non-building areas, which confirms that CDAM effectively combines complementary context and detail cues for more precise building delineation.

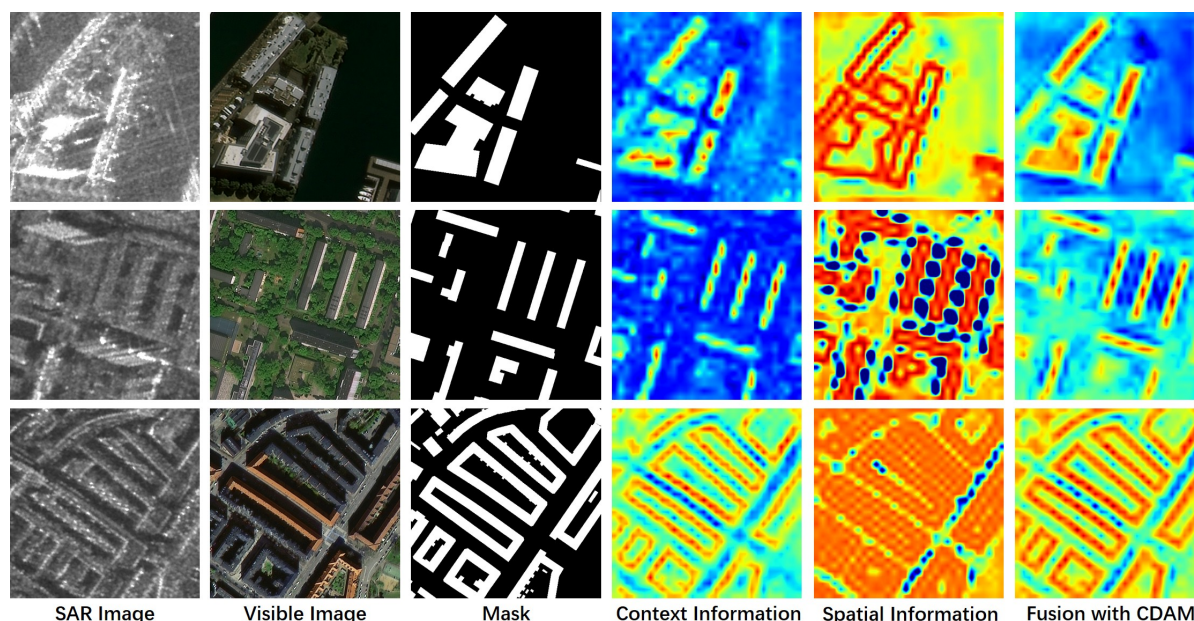


Figure 11. Feature heatmaps of the context and spatial branches before and after CDAM fusion. The color intensity reflects the activation strength of the extracted features. The context branch captures global semantic cues but suffers from boundary ambiguity and noise. In contrast, the spatial branch accurately highlights the sharp contours of buildings. The fused heatmap integrates the strengths of both branches, producing a clear and well-organized focus on building regions that closely matches the ground-truth mask.

5.3. Effectiveness of STDAC vs. Bottleneck and Basic Block

To evaluate the effectiveness of the proposed STDAC module for context feature extraction, we conduct a comparative experiment by replacing the STDAC module in the context branch with standard Bottleneck and Basic Block structures. The results are reported in Table 6. The model using STDAC achieves 82.3% IoU and 90.29% F1-score, outperforming the alternatives significantly. Replacing STDAC with Bottleneck results in a performance drop to 80.6% IoU and 89.25% F1-score, while the Basic Block further reduces performance to 79.5% IoU and 88.58% F1-score. These findings confirm that STDAC provides superior capability in extracting context features related to building structures. Its enhanced feature aggregation and multi-scale spatial perception help generate more complete and precise building representations.

Table 6. Ablation analysis for context feature extraction modules and different dilation configurations of STDAC on the DFC23 Track2 dataset.

Model	Context Feature Extraction	IoU (%) ↑	F1-Score (%) ↑
DLiteNet	STDAC (d = 1, 1, 1)	80.8	89.4
DLiteNet	STDAC (d = 2, 2, 2)	81.2	89.7
DLiteNet	STDAC (d = 3, 3, 3)	82.3	90.3
DLiteNet	Bottleneck	80.6	89.3
DLiteNet	Basic Block	79.5	88.6

Note: For metrics marked with “↑”, higher values indicate better performance. Bold numbers indicate the best results.

Furthermore, we analyze different dilation configurations within STDAC by setting the dilation rates of its convolutions to (1, 1, 1), (2, 2, 2) and (3, 3, 3) (the configuration used in DLiteNet). As summarized in Table 6, the (3, 3, 3) setting achieves the best IoU and F1-score, while smaller dilation rates (1, 1, 1) and (2, 2, 2) lead to slightly inferior performance.

This indicates that using a moderate dilated receptive field in STDAC is beneficial for capturing long-range building context without introducing severe gridding artifacts.

Overall, the three proposed modules collaboratively enhance the robustness and efficiency of DLiteNet. MCAM enables effective multimodal alignment, CDAM refines boundary localization through context–spatial fusion, and STDAC strengthens contextual representation with minimal computational overhead, jointly supporting the model’s superior balance between accuracy and efficiency.

5.4. Limitations and Failure Cases

Although DLiteNet demonstrates strong robustness across typical urban scenes, it still has limitations under extreme imaging conditions. In particular, when the visible modality is heavily degraded by dense fog or thick haze and the SAR backscatter of buildings becomes weak or ambiguous, the complementary cues from the two modalities are not sufficient to fully recover building structures. Figure 12 shows a representative failure case under dense fog conditions. In this example, large portions of the building regions are severely obscured in the visible image, and the corresponding SAR responses are also weak and contaminated by surrounding clutter. As a result, DLiteNet fails to detect the building roof within the highlighted bounding box, even though it performs well on nearby, less-degraded areas in the same scene. This indicates that the current model is still sensitive to simultaneous degradations of both modalities, and its ability to exploit subtle structural priors is limited in such extreme scenarios. These observations highlight two main directions for future work. First, constructing or simulating multimodal datasets with richer coverage of adverse weather (e.g., heavy rain, dense fog, and low-contrast SAR) would allow explicit training and evaluation under these conditions. Second, incorporating weather-aware priors or robustness-oriented fusion strategies (e.g., uncertainty-aware weighting of modalities, resolution- and quality-adaptive fusion) may further improve performance when one or both modalities are severely corrupted.

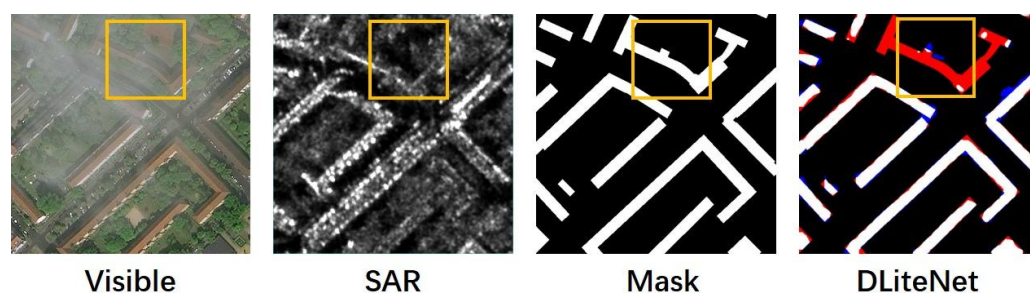


Figure 12. Representative failure case of DLiteNet under dense fog conditions. The yellow box highlights the region where the model fails due to severe occlusion in the visible image and weak or ambiguous backscatter in the SAR channel. This results in missing part of the building footprint within the highlighted area, illustrating the limitation of the model when both modalities are simultaneously degraded.

6. Conclusions

In this paper, we proposed DLiteNet, a novel dual-branch lightweight network designed for multi-modal building extraction, which efficiently integrates visible and SAR imagery to achieve high-precision and efficient segmentation. By decoupling feature representation into a context branch for semantic extraction and a spatial branch for detail preservation, DLiteNet effectively captures complementary information while maintaining a compact and deployable architecture. The proposed MCAM enhances multi-scale cross-modal fusion, the proposed STDAC module effectively extracts multi-scale contextual information while reducing computational cost, and the proposed CDAM module

adaptively integrates context and detail features for better building-specific representations. Extensive experiments conducted on the DFC23 Track2 and MSAW datasets demonstrate the effectiveness and generalization capability of the proposed approach. On the DFC23 Track2 dataset, DLiteNet achieves an IoU of 82.3% and an F1-score of 90.3% with only 5.6 M parameters and 80.28 FPS on an RTX GPU, outperforming state-of-the-art methods in both accuracy and efficiency. On the MSAW dataset, it achieves an IoU of 83.6% and an F1-score of 91.1% while requiring only 51.73 GFLOPs, highlighting its superior performance in densely built environments. In addition, deployment experiments on the Jetson Orin NX platform confirm the efficiency of DLiteNet under resource-constrained environments, achieving an inference latency of 11.18 ms after INT8 quantization. In summary, DLiteNet offers an effective and efficient solution for multi-modal building extraction, achieving a favorable trade-off between segmentation accuracy, model complexity, and inference speed, and showing strong potential for deployment in practical remote sensing applications.

Author Contributions: Z.Z. and B.Z. created and designed the framework. Z.Z. performed the algorithm on PyTorch. Z.Z., B.Z., Y.W., R.D., J.C. and Y.Z. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China under Grant 2021YFA0715203, and the Science and Disruptive Technology Program, AIRCAS under Grant 2024-AIRCAS-SDTP-12, and National Natural Science Foundation of China under Grant 62001455 and 41871245.

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

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

References

1. Benedek, C.; Descombes, X.; Zerubia, J. Building Development Monitoring in Multitemporal Remotely Sensed Image Pairs with Stochastic Birth–Death Dynamics. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 33–50. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Hou, Z.; Qu, Y.; Zhang, L.; Liu, J.; Wang, F.; Yu, Q.; Zeng, A.; Chen, Z.; Zhao, Y.; Tang, H.; et al. War City Profiles Drawn from Satellite Images. *Nat. Cities* **2024**, *1*, 359–369. [\[CrossRef\]](#)
3. Sirmacek, B.; Unsalan, C. Urban-Area and Building Detection Using SIFT Keypoints and Graph Theory. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 1156–1167. [\[CrossRef\]](#)
4. Zhang, B.; Wang, C.; Zhang, H.; Wu, F. A Review on Building Extraction and Reconstruction from SAR Image. *Remote Sens. Technol. Appl.* **2012**, *27*, 496–503.
5. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Wang, S.; Li, X.; Chen, Y.; Li, Z.; Zhang, L. MCANet: A Joint Semantic Segmentation Framework of Optical and SAR Images for Land Use Classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *106*, 102638. [\[CrossRef\]](#)
6. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [\[CrossRef\]](#)
7. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856. [\[CrossRef\]](#)
8. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-Attention Mask Transformer for Universal Image Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 1280–1289. [\[CrossRef\]](#)
9. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; Lecture Notes in Computer Science; Volume 9351, pp. 234–241. [\[CrossRef\]](#)
10. Li, X.; Lei, L.; Sun, Y.; Li, M.; Kuang, G. Multimodal Bilinear Fusion Network with Second-Order Attention-Based Channel Selection for Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1011–1026. [\[CrossRef\]](#)
11. Orsic, M.; Kreso, I.; Bevandic, P.; Segvic, S. In Defense of Pre-Trained ImageNet Architectures for Real-Time Semantic Segmentation of Road-Driving Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 12607–12616. [\[CrossRef\]](#)

12. Li, H.; Xiong, P.; Fan, H.; Sun, J. DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9522–9531.
13. Gamal, M.; Siam, M.; Abdel-Razek, M. ShuffleSeg: Real-Time Semantic Segmentation Network. *arXiv* **2018**, arXiv:1803.03816.
14. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147. [[CrossRef](#)]
15. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 405–420. [[CrossRef](#)]
16. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 334–349. [[CrossRef](#)]
17. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [[CrossRef](#)]
18. Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; Wei, X. Rethinking BiSeNet for Real-Time Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 9716–9725. [[CrossRef](#)]
19. Pan, H.; Hong, Y.; Sun, W.; Jia, Y. Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Traffic Scenes. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 3448–3460. [[CrossRef](#)]
20. Xu, J.; Xiong, Z.; Bhattacharyya, S.P. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 19529–19539. [[CrossRef](#)]
21. Alshehhi, R.; Marpu, P.R.; Woon, W.L.; Dalla Mura, M. Simultaneous Extraction of Roads and Buildings in Remote Sensing Imagery with Convolutional Neural Networks. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 139–149. [[CrossRef](#)]
22. Yang, H.L.; Yuan, J.; Lunga, D.; Laverdiere, M.; Rose, A.; Bhaduri, B. Building Extraction at Scale Using Convolutional Neural Network: Mapping of the United States. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2600–2614. [[CrossRef](#)]
23. Hui, J.; Du, M.; Ye, X.; Qin, Q.; Sui, J. Effective Building Extraction from High-Resolution Remote Sensing Images with Multitask Driven Deep Neural Network. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 786–790. [[CrossRef](#)]
24. Rapuzzi, A.; Nattero, C.; Pelich, R.; Chini, M.; Campanella, P. CNN-Based Building Footprint Detection from Sentinel-1 SAR Imagery. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 1707–1710. [[CrossRef](#)]
25. Wu, F.; Wang, C.; Zhang, H.; Li, J.; Li, L.; Chen, W.; Zhang, B. Built-Up Area Mapping in China from GF-3 SAR Imagery Based on the Framework of Deep Learning. *Remote Sens. Environ.* **2021**, *262*, 112515. [[CrossRef](#)]
26. Kang, J.; Wang, Z.; Zhu, R.; Xia, J.; Sun, X.; Fernandez-Beltran, R.; Plaza, A. DisOptNet: Distilling Semantic Knowledge from Optical Images for Weather-Independent Building Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4706315. [[CrossRef](#)]
27. Li, Q.; Mou, L.; Sun, Y.; Hua, Y.; Shi, Y.; Zhu, X.X. A review of building extraction from remote sensing imagery: Geometrical structures and semantic attributes. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4702315. [[CrossRef](#)]
28. Guo, H.; Du, B.; Zhang, L.; Su, X. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *183*, 240–252. [[CrossRef](#)]
29. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Chen, Y.; Li, Z.; Li, H.; Wang, H. Progressive fusion learning: A multimodal joint segmentation framework for building extraction from optical and SAR images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 178–191. [[CrossRef](#)]
30. Zhang, P.; Peng, B.; Lu, C.; Huang, Q.; Liu, D. ASANet: Asymmetric Semantic Aligning Network for RGB and SAR image land cover classification. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 574–587. [[CrossRef](#)]
31. Zhao, Z.; Zhao, B.; Wu, Y.; He, Z.; Gao, L. Building Extraction From High-Resolution Multispectral and SAR Images Using a Boundary-Link Multimodal Fusion Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 3864–3878. [[CrossRef](#)]
32. Shermeyer, J.; Hogan, D.; Brown, J.; Van Etten, A.; Weir, N.; Pacifici, F.; Hansch, R.; Bastidas, A.; Soenen, S.; Bacastow, T. SpaceNet 6: Multi-Sensor All Weather Mapping Dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, DC, USA, 14–19 June 2020; pp. 768–777. [[CrossRef](#)]
33. Huang, X.; Ren, L.; Liu, C.; Wang, Y.; Yu, H.; Schmitt, M.; Hänsch, R.; Sun, X.; Huang, H.; Mayer, H. Urban Building Classification (UBC)—A dataset for individual building detection and classification from satellite imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, New Orleans, LA, USA, 19–20 June 2022; pp. 565–571.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
36. Milletari, F.; Navab, N.; Ahmadi, S.A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision, Stanford, CA, USA, 25–28 October 2016; pp. 565–571.

37. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large Kernel Matters: Improve Semantic Segmentation by Global Convolutional Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1743–1751. [[CrossRef](#)]
38. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3146–3154. [[CrossRef](#)]
39. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, Granada, Spain, 16 September 2018; pp. 3–11. [[CrossRef](#)]
40. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1–11.
41. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [[CrossRef](#)] [[PubMed](#)]
42. Li, C.; Liu, X.; Li, W.; Wang, C.; Liu, H.; Yuan, Y. U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation. *arXiv* **2024**, arXiv:2406.02918. [[CrossRef](#)]
43. Jiang, J.; Zheng, L.; Luo, F.; Zhang, Z. RedNet: Residual Encoder-Decoder Network for Indoor RGB-D Semantic Segmentation. *arXiv* **2018**, arXiv:1806.01054.
44. Chen, X.; Lin, K.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-Directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 561–577. [[CrossRef](#)]
45. Hosseinpour, H.; Samadzadegan, F.; Javan, F.D. CMGFNet: A Deep Cross-Modal Gated Fusion Network for Building Extraction from Very High-Resolution Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 96–115. [[CrossRef](#)]

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