MINISTERUL EDUCAȚIEI ȘI CERCETĂRII

UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA

Computer Science

PhD Thesis
– ABSTRACT –

Visual Perception and Scene Understanding Methods for Remote Sensing Applications

PhD Student:
Eng. Bianca-Cerasela-Zelia BLAGA

PhD Supervisor:
Prof. Dr. Eng. Sergiu NEDEVSCHI

Examination committee:
Chair: Prof. Dr. Eng. **Radu Gabriel Dănescu** – Technical University of Cluj-Napoca;
PhD Supervisor: Prof. Dr. Eng. **Sergiu Nedevschi** – Technical University of Cluj-Napoca;
Members:
-Prof. Dr. Eng. **Vladimir-Ioan Crețu** – Polytechnic University of Timișoara;
-Prof. Dr. Eng. **Vasile-Ion Manta** – "Gheorghe Asachi" Technical University of Iași;
-Assoc. Prof. Dr. Eng. **Raluca Didona Brehar** – Technical University of Cluj-Napoca.

– Cluj-Napoca –
2026

str. Memorandumului nr. 28, 400114 Cluj-Napoca, România
tel. +40-264-401200, fax +40-264-592055, secretariat tel. +40-264-202209, fax +40-264-202280
www.utcluj.ro

Thesis Contents

Introduction

Context, Motivation, and Research Objectives

Remote sensing imagery acquired by satellites and unmanned aerial vehicles (UAVs) enables large-scale characterization of land cover, infrastructure, vegetation, and disaster impacts. As spatial resolution and revisit rates increase, data volumes grow rapidly and acquisition conditions become more diverse. Manual interpretation does not scale, which motivates automated visual perception and scene understanding methods that deliver consistent, timely, and spatially accurate outputs.

Deep learning forms the dominant methodological basis for these capabilities. Convolutional neural networks (CNNs) remain effective for dense prediction, particularly semantic segmentation, due to strong local representation learning and preservation of fine spatial detail. In contrast, attention mechanisms and transformer-based architectures strengthen global context aggregation, long-range dependency modeling, and multi-scale feature interaction. For aerial imagery, global–local designs are especially relevant because they reconcile boundary fidelity with scene-level reasoning.

Temporal analysis is also central to many operational tasks. Change detection aims to identify relevant differences between observations acquired at different times, supporting monitoring of land-cover transitions, urban expansion, and infrastructure evolution. Fine-grained change detection further requires precise delineation of subtle, object-level changes, which motivates region and instance-aware representations for bi-temporal alignment and comparison.

Disaster assessment exposes these requirements under strict time and reliability constraints. Visual Question Answering (VQA) provides a query-driven formulation for damage analysis, allowing models to return targeted attributes and quantities directly from imagery. Counting-oriented questions are particularly decision-relevant, yet they remain difficult under clutter, occlusion, and large scale variation.

Across segmentation, VQA, and change detection, a persistent limitation is the reliance on large and consistently annotated datasets. Dense supervision is expensive and uneven across sensors, environments, and acquisition regimes. This motivates scalable data generation through simulation, together with learning paradigms that reduce dependence on exhaustive labeling, including weakly supervised, semi-supervised, and self-supervised approaches.

Within this context, the thesis targets five challenges: achieving dataset scale with consistent semantic coverage; enabling sequence-aware learning that exploits temporal coherence under

limited supervision; improving urban semantic segmentation by balancing boundary accuracy with contextual reasoning; strengthening counting-oriented multimodal VQA through instance-sensitive, scale-aware fusion; and advancing fine-grained change detection via structure-aware temporal reasoning with instance-aligned representations and correlation-preserving prediction.

We define the following research objectives and sub-objectives for this thesis:

Objective 1: Semantic segmentation for UAV and aerial scene understanding.
Dense semantic scene interpretation in aerial imagery, with emphasis on forest environments and constraints related to viewpoint variation and labeling effort.

Objective 1.1: Study of UAV semantic segmentation requirements and the limitations of existing datasets for forest and vegetation understanding.

Objective 1.2: Design and release of a synthetic UAV dataset for forest semantic segmentation with controllable acquisition variability and dense pixel-level labels.

Objective 1.3: Development of a weakly supervised spatio-temporal segmentation framework for UAV videos that reduces annotation cost via label propagation and temporal refinement.

Objective 1.4: Development of a transformer-based semantic segmentation architecture that improves global context modeling while preserving local boundaries in high-resolution aerial imagery.

Objective 2: Remote-sensing VQA with emphasis on post-disaster assessment and counting.
Multimodal question answering for remote sensing, with emphasis on post-flood assessment and on improving counting accuracy under scale variation and high answer variability.

Objective 2.1: Study of remote-sensing VQA characteristics, with focus on counting as a recurrent source of error in post-disaster assessment.

Objective 2.2: Development of a VQA architecture for post-flood assessment integrating language and visual encoding with multimodal fusion and structured decision modeling.

Objective 2.3: Development of a counting-oriented VQA framework that improves multi-scale visual representation, multimodal alignment, and count prediction under long-tailed answer distributions.

Objective 3: Fine-grained change detection with structure-aware tokenization and temporal reasoning.
Accurate change localization in structured remote-sensing scenes (e.g., croplands and buildings), with emphasis on structure-aligned representations and reliable temporal fusion.

Objective 3.1: Study of limitations in existing CD pipelines, with emphasis on the mismatch between uniform patch tokenization and structure-centric domains.

Objective 3.2: Development of a region-aware CD framework for structured agricultural scenes using semantic region tokenization and region-level temporal reasoning.

Objective 3.3: Development of a building-focused CD framework that enforces instance-consistent representation and improves boundary-preserving change prediction.

Thesis Structure

Chapter 2. State of the Art provides the state-of-the-art review corresponding to the three technical pillars of the thesis. It summarizes representative datasets, evaluation protocols, and model families for semantic segmentation in aerial/UAV imagery, VQA in remote sensing (with emphasis on the challenges of counting and mixed answer spaces), and deep learning-based change detection, highlighting recurring limitations that motivate the methodological choices developed in the subsequent chapters.

Chapter 3. Semantic Segmentation for UAV-based Remote Sensing addresses semantic segmentation for UAV-based remote sensing by combining data-centric and model-centric contributions. It first introduces a synthetic UAV dataset for forest inspection with controllable acquisition variability and dense pixel-level labeling, designed to study generalization under viewpoint and environmental changes. It then investigates weakly supervised spatio-temporal learning for UAV video segmentation via label propagation and temporal refinement, targeting scenarios where dense annotation is impractical. Finally, it presents a transformer-based segmentation architecture (SwinFAN) that augments the decoder with global-local context fusion and feature refinement to improve boundary quality and fine detail recovery. The chapter validates these components through benchmark experiments, controlled ablations, and qualitative analyses that relate observed behavior to data properties and architectural design choices.

Chapter 4. Visual Question Answering in Remote Sensing studies visual question answering in remote sensing for post-disaster damage assessment, with an explicit focus on improving counting accuracy under heterogeneous question types and mixed answer spaces. It first introduces DiViNeCaps, a pipeline that combines a pretrained language encoder with Vision Mamba visual embeddings and an attention-driven fusion module, and uses a Capsule Network head to support both categorical answers and numerical count estimation within a unified formulation. It then advances to DeVANet, which strengthens counting-oriented reasoning by incorporating local-global visual modeling (LGA-ViM), bidirectional multimodal fusion (self-attention and bidirectional cross-attention), and task-specific prediction heads for classification and count estimation, trained with objectives that account for class imbalance and overdispersed count distributions. The chapter evaluates both systems on FloodNet and RescueNet, compares against representative VQA baselines, and reports quantitative results, error analyses across low and high-count regimes, and ablation studies that isolate the impact of encoders, fusion, resolution, and loss design.

Chapter 5. Change Detection in Remote Sensing focuses on change detection in remote sensing and develops region and instance-aware reasoning mechanisms for fine-grained change localization. It first presents SFCNet for structured agricultural scenes, where semantic region masks guide feature aggregation, temporal reasoning over region-level tokens, and spatial decoding to improve boundary consistency and reduce spurious detections. It then introduces RADNet for building change detection in high-resolution imagery, combining instance-consistent tokenization, a global-local attention encoder to refine representations, and a cross-temporal cross-scale decoding strategy coupled with correlation-based prediction to preserve spatial detail while maintaining temporal identity. The proposed methods are validated on cropland monitoring and building change detection benchmarks through standard metrics, qualitative comparisons, and analyses that emphasize semantic alignment, boundary precision, and robustness to appearance variability.

Chapter 6. Conclusions summarizes the thesis contributions with respect to the stated objectives and consolidates the main empirical findings across the three research directions. It discusses how the proposed datasets, learning architectures, and evaluation protocols jointly support the thesis goal of reliable perception for aerial and remote sensing applications, and it concludes with limitations, future research directions, and a dissemination summary of the resulting publications.

Chapter 7. Appendix provides supplementary material that supports reproducibility and deeper inspection of the results, including additional experimental details, extended ablation studies, and complementary quantitative tables for the models analyzed in the main chapters.

Semantic Segmentation for UAV-based Remote Sensing

Synthetic Forest Dataset for Semantic Segmentation

We introduce the Forest Inspection dataset, a synthetic unmanned aerial vehicle (UAV) dataset for semantic segmentation in forest environments. The dataset is generated in Unreal Engine 4 using physics-based flight simulation via AirSim and automatic mask extraction through a class-specific stencil encoding and a fixed labeling palette. It provides paired RGB images and dense pixel-level semantic maps at high resolution, together with UAV pose metadata, and it follows a controlled sweep-trajectory acquisition protocol that samples three altitudes (30, 50, 80m), three pitch angles (0°, 60°, 90°), and two weather conditions (sunny, overcast) under a consistent illumination setup. The taxonomy contains 11 classes, explicitly separates deciduous from coniferous trees, and includes an inspection-relevant fallen tree class. **Figure 1** shows representative RGB-label pairs, where each row presents a camera image followed by its corresponding semantic map.

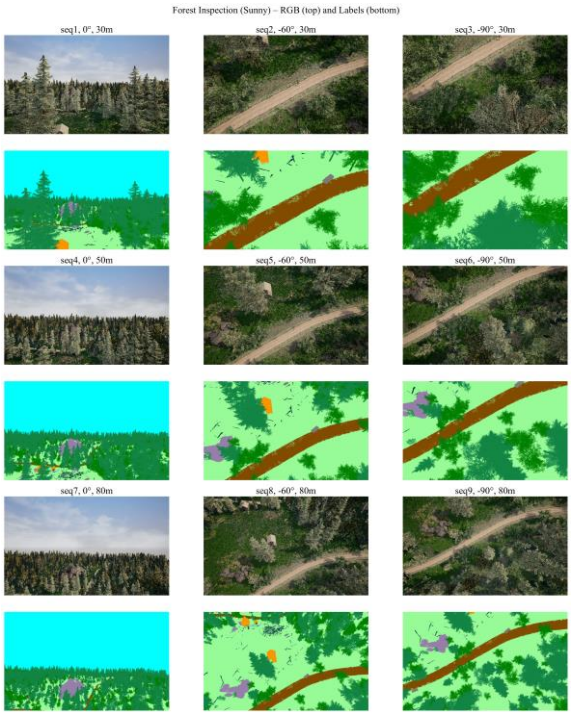

Figure 1. Sample RGB images (top row) and their corresponding semantic segmentation labels (bottom row) from the Forest Inspection synthetic dataset, shown for each flight sequence under sunny conditions.

Table 1 positions Forest Inspection with respect to representative UAV benchmarks by contrasting dataset type, scale, acquisition geometry, resolution, semantic design, and operational relevance for forest inspection. UAVid [1], Ruralscapes [2], and WildUAV [3] are real-world collections acquired at very high resolutions and largely under sunny conditions, while Mid-Air [4] and SPREAD [5] are synthetic datasets, focused on semantic and instance segmentation. Forest Inspection is a sequence-based synthetic dataset designed for controlled inspection trajectories, enabling pixel-accurate ground truth and reproducible acquisition settings. The dataset contains 26,319 images recorded at 3 altitudes and 3 pitch angles with a uniform high resolution, providing higher detail than common synthetic baselines while remaining practical for large-scale experimentation. Unlike previous UAV segmentation datasets that represent trees as a single class, Forest Inspection introduces explicit deciduous and coniferous tree separation and a fallen tree class to support hazard and damaged scene understanding. The dataset further targets realistic yet controlled environmental variability by focusing on sunny and overcast conditions, consistent with typical operational flight conditions, and its sequence organization yields lower frame overlap than other benchmarks such as Mid-Air and Ruralscapes, indicating reduced redundancy while preserving continuity.

Table 1. Comparison of UAV-based real and synthetic datasets for forest and aerial scene analysis.

Dataset	Mid-Air [72]	UAVid [70]	Ruralscapes [71]	WildUAV [182]	SPREAD [8]	**Forest Inspection**
Year	2019	2020	2020	2021	2025	**2025**
Type	Synthetic	Real	Real	Real	Synthetic	**Synthetic**
#Imgs	420,000	420	812	2,608	19,000	**26,319**
Resolution	512×512	4096×2160 3840×2160	4096×2160	3840×2160	960×540	1280×768
Altitude(m)	< 5	50	Varying	30	100	**30, 50, 80**
Pitch(°)	0	45	Random	80	90	**0, 60, 90**
#Classes	12	8	12	9	255	**11**
Tree Types	One	One	One	One	Instance	**Two (Dec., Conif.)**
Fallen Tree	No	No	No	No	No	**Yes**
Weather	Multiple conditions	Sunny	Sunny	Sunny	Multiple conditions	**Sunny, Overcast**
Overlap (%)	99.03	75.60	98.73	75.36	N/A	**90.02**

Weakly Supervised UAV Video Segmentation

We address semantic segmentation in UAV video sequences under limited pixel-level supervision by formalizing a weakly supervised setting in which dense labels are available only on every k-th frame while predictions are refined over time. The processing pipeline is illustrated in **Figure 2**. The approach couples a per-frame semantic segmentation network with optical flow estimation to propagate information between consecutive frames and applies spatio-temporal warping and recurrent refinement to improve temporal consistency. The proposed spatio-temporal gated recurrent unit integrates warping-based propagation with a modified recurrent module that uses depthwise separable convolutions to implement gating and state updates, producing refined segmentation outputs at each time step. Optical flow is used both as a source of supervision in the weakly supervised configuration and as a component that can be trained jointly with the segmentation network in an end-to-end setting.

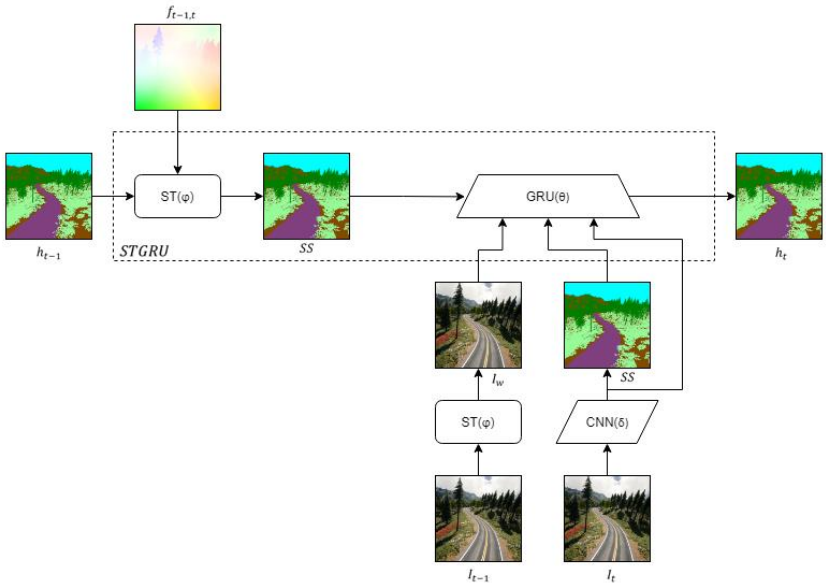

Figure 2. Framework flow of the proposed methodology, where the input RGB frames together with the semantic segmentation and optical flow are fed to the STGRU component to propagate and refine the labeling.

We evaluate the framework using the Mid-Air dataset, and report end-to-end training results with two optical-flow networks in **Table 2**. Joint optimization of the semantic segmentation and motion-estimation components increases segmentation accuracy, supporting the interpretation that improved flow estimation facilitates more consistent spatio-temporal scene reasoning. The gains are most visible for classes that benefit from temporal cues, including ground vegetation, rocky ground, boulders, and road sign. Overall, the end-to-end configuration achieves an improvement of approximately 4% over the static segmentation baseline, with a corresponding increase in computational cost.

Qualitative predictions on the test set are presented in **Figure 3**. The examples cover both forest and road scenes and include diverse textures and illumination conditions. The most challenging categories are boulders and traffic signs, which are sparsely represented in the training data and therefore exhibit less stable segmentation, particularly at object boundaries and in cluttered regions.

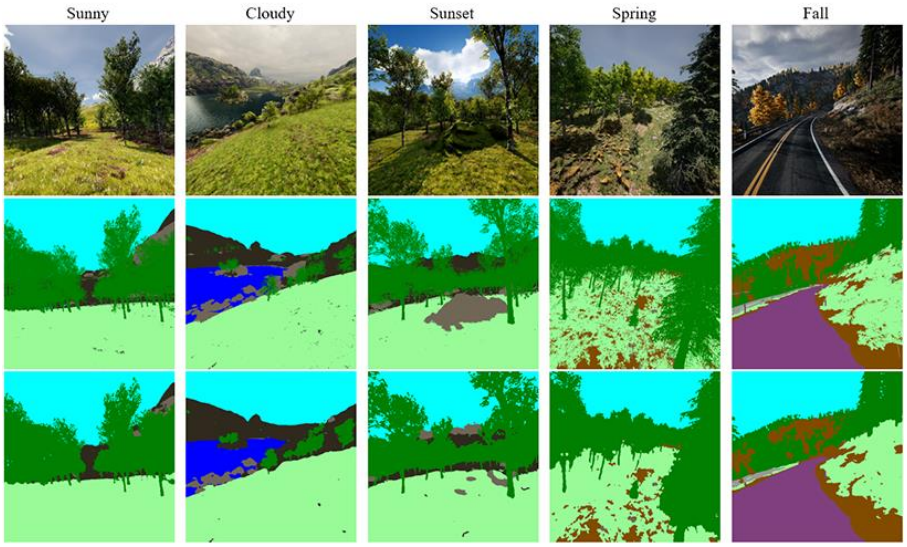

Figure 3. Results of the weakly supervised framework, presented for the 5 different scenarios from Mid-Air dataset.

Table 2. IoU results on the Mid-Air test set, where GRU(f,k) represents the full system, trained with every k-th label, f is using either the ground truth (GT) optical flow from PWC-Net or the one generated from the retrained VCN.

Class	ERFNet(8)	GRU(GT, 8)	GRU(VCN, 8)
Sky	92.34	92.50	**92.67**
Trees	85.23	89.45	**89.75**
Dirt ground	70.73	**73.49**	73.19
Ground vegetation	79.61	81.58	**83.47**
Rocky ground	67.17	68.85	**72.35**
Boulders	65.74	70.61	**71.82**
Water plane	86.27	88.47	**88.82**
Road	86.22	90.45	**91.21**
Train track	73.92	**76.89**	76.45
Road sign	36.54	39.02	**40.83**
Man-made objects	55.41	**59.58**	59.45
Mean IoU	72.65	75.54	**76.36**

Transformer-based Semantic Segmentation

We propose a transformer-based U-Net for semantic segmentation of remote sensing imagery and introduce SwinFAN (Swin-based Focal Axial Attention Network), illustrated in **Figure 4**, a four-level U-Net architecture that combines a hierarchical Swin Transformer encoder with a decoder built on Guided Focal–Axial (GFA) attention and an Attention-based Feature Refinement Head (AFRH) for output refinement. The encoder represents an image as a sequence of non-overlapping patches and employs window-based self-attention with shifted windows across multiple stages to obtain multi-scale features; the Swin block design is detailed in **Figure 5 (a)**. The decoder reconstructs predictions in a coarse-to-fine manner and uses GFA attention, shown in **Figure 5 (b)**, to couple axial attention with a focal module that produces a multi-scale spatial mask guiding the attention response, thereby combining long-range contextual modeling with spatially selective emphasis. Encoder and decoder features are merged through a weighted fusion mechanism, and AFRH (**Figure 6**) further refines the fused representation by aligning residual features, applying spatial attention reweighting, and using depthwise separable convolutions with residual connections to preserve fine structures. Training uses a hybrid loss function that combines cross-entropy and Dice loss to jointly optimize per-pixel classification and region overlap under class imbalance.

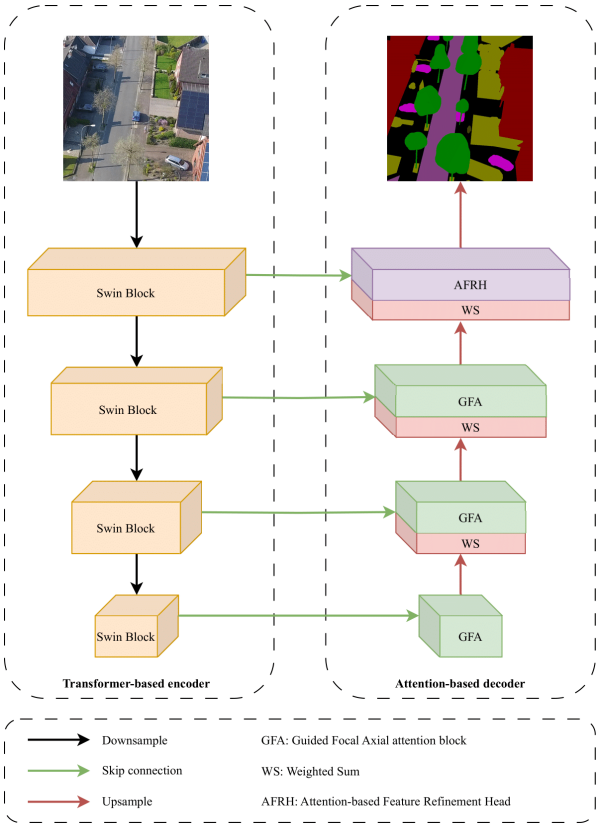

Figure 4. SwinFAN architecture. The encoder captures features at multiple scales using Swin transformer blocks, while the decoder employs Guided Focal Axial attention for integrating global and local attention features. As a final step, the Attention-based Feature Refinement Head refines the output to produce a precise segmentation map.

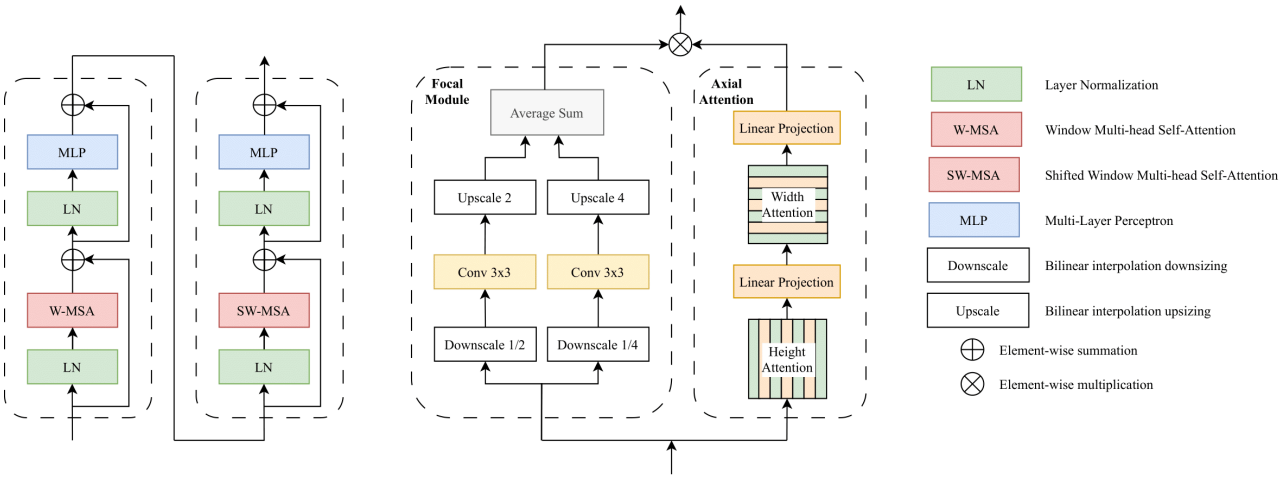

Figure 5. Key building blocks of the SwinFAN architecture. (a) Swin Block: Consists of a series of layer normalization, multi-head self-attention mechanisms (window-based and shifted window), and multi-layer perceptrons. (b) Guided Focal-Axial attention Block: Implements an element-wise multiplication of focal module which refines features at two scales to emphasize local detail and of axial attention which sequentially targets the feature map's vertical and horizontal axes to capture global dependencies.

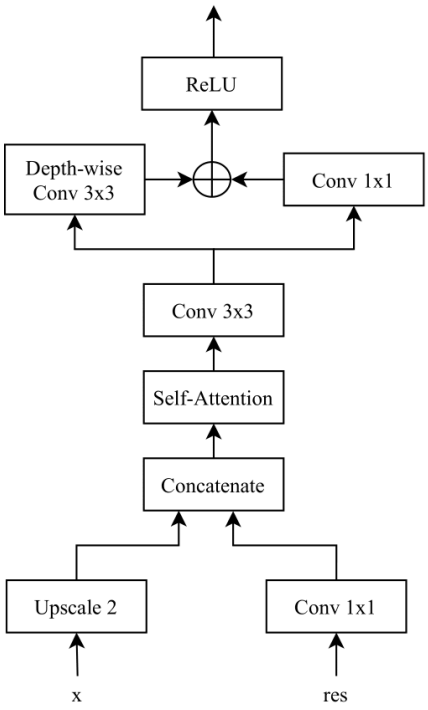

Figure 6. Architecture of the Attention-based Feature Refinement Head.

We evaluate SwinFAN on four public benchmarks (UAVid, Potsdam [6], Vaihingen [7], and LoveDA [8]) under a unified training protocol. The method is compared against a broad set of CNN, attention, and transformer-based baselines, including global-local variants. On UAVid, **Table 3** showcases that SwinFAN achieves the best mIoU, with notable gains on building, road, and human classes, and the qualitative comparison presented in **Figure 7** highlights improved boundaries and small-object detection. On Potsdam and Vaihingen, the model ranks first on the main evaluation metrics (F1, mIoU, and OA), indicating consistent generalization across high-resolution urban scenes, and it outperforms both ViT and Swin-based competitors. On LoveDA, SwinFAN attains the highest mIoU while reaching 196.1 FPS, performing strongly on building, road, water, and agriculture, with the main weakness observed on the Barren class due to limited texture and class imbalance. Ablation studies on UAVid confirm the contribution of the decoder and refinement design: Guided Focal–Axial attention yields the best decoder strategy, and the proposed Attention-based Feature Refinement Head adds a further 1.1% mIoU, with the largest improvements on small objects. A Swin-Base backbone comparison on Potsdam additionally shows that SwinFAN achieves superior accuracy with only 45 epochs, reducing training cost relative to longer-schedule Swin baselines.

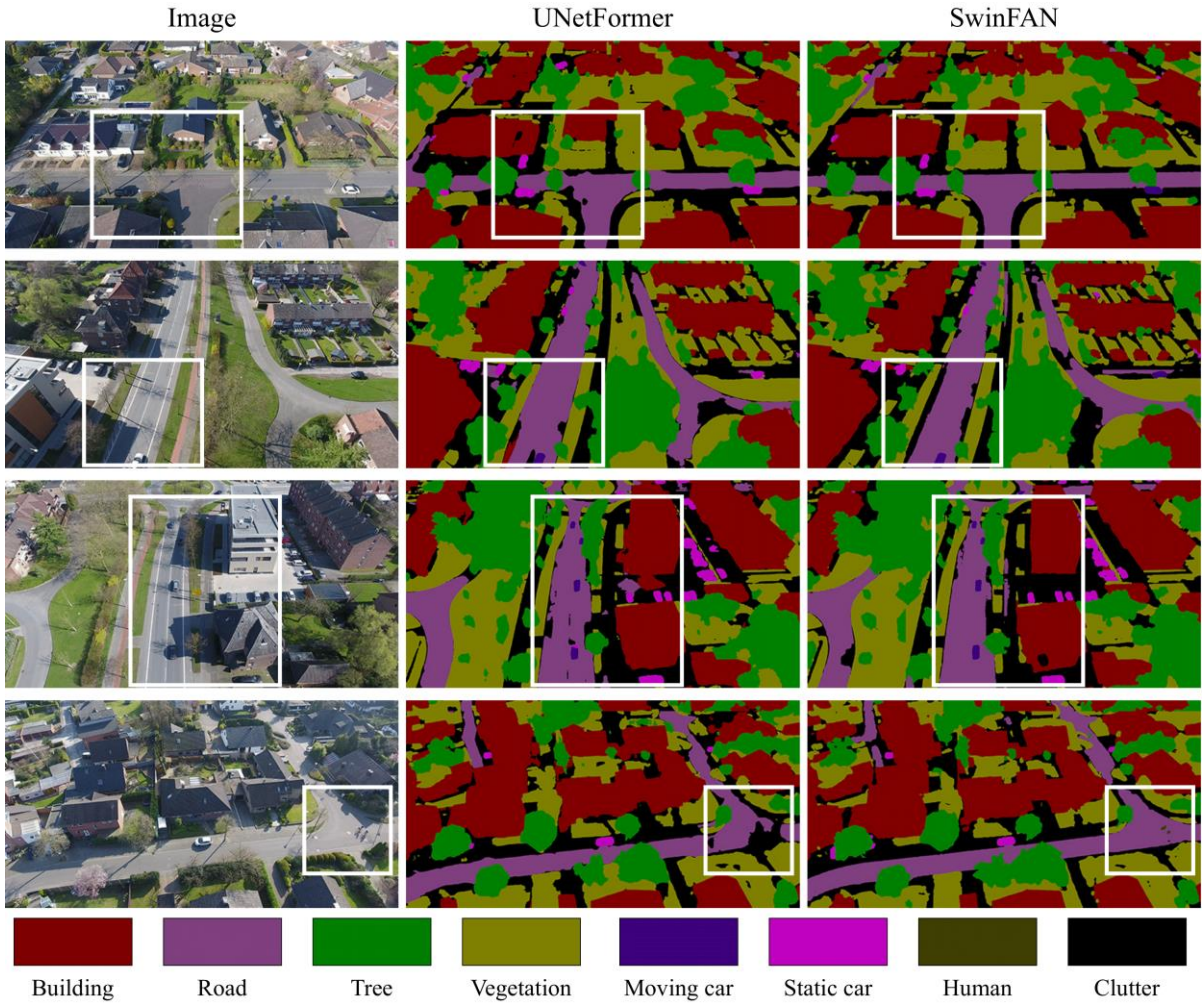

Figure 7. Comparative qualitative analysis of semantic segmentation performance of UNetFormer and SwinFAN, on the UAVid test set in Germany, from sequences 29, 30, and 38. With the white box we highlighted the improved road, car, and human detections.

Table 3. Performance evaluation on UAVid test dataset, highlighting the best results in bold and second best with an underline.

Network	Backbone	Building	Road	Tree	Vegetation	Moving Car	Static Car	Human	Clutter	mIoU
MSD[70]	-	79.8	74.0	74.5	55.9	62.9	32.1	19.7	57.0	57.0
Segmenter[222]	ViT-Tiny	84.4	79.8	76.1	57.6	59.2	34.5	14.2	64.2	58.7
SE-UNet[214]	ResNet18	77.0	69.9	75.0	58.9	72.3	54.9	18.8	56.3	60.4
DANet[104]	ResNet18	85.9	77.9	78.3	61.5	59.6	47.4	9.1	64.9	60.6
C-PNet[211]	-	60.5	72.3	79.3	55.4	66.5	49.8	29.0	**73.5**	60.7
SwiftNet[92]	ResNet18	85.3	61.5	78.3	76.4	51.1	62.1	15.7	64.1	61.7
UNet++[215]	ResNet18	86.4	72.6	74.4	59.3	65.3	55.8	24.0	65.7	63.0
BoTNet[107]	ResNet18	84.9	78.6	77.4	60.5	65.8	51.9	22.4	64.5	63.2
CANet[119]	ResNet18	86.6	62.1	79.3	**78.1**	47.8	68.3	19.9	66.0	63.6
A^2-FPN[103]	ResNet18	87.2	80.2	63.7	63.7	70.1	53.3	23.4	67.4	63.9
BANet[106]	ResT-Lite	85.4	80.7	78.9	62.1	69.3	52.8	21.0	66.7	64.6
CoaT[218]	CoaT-Mini	88.5	80.0	79.3	62.0	70.0	59.1	18.9	69.0	65.8
SegFormer[114]	MiT-B1	86.3	80.1	79.6	62.3	72.5	52.5	28.5	66.6	66.0
Mamba-UNet[213]	VMamba	88.1	74.2	74.2	62.9	69.9	60.0	30.8	69.7	66.2
UNetFormer[112]	ResNet18	86.9	80.9	79.9	64.8	72.0	61.3	30.4	67.2	67.9
EDDformer[120]	ResNet18	86.8	80.4	79.7	63.2	63.1	**73.0**	30.6	67.0	68.0
RDNet[105]	ResNet18	87.6	81.2	80.1	63.7	73.0	58.9	32.8	68.5	68.2
MFNet[216]	ResNet50	88.6	82.5	81.3	66.4	73.2	60.6	27.4	69.7	68.7
SwinFAN	**Swin-Base**	**89.5**	**82.6**	**81.8**	66.4	**76.8**	64.0	**33.0**	70.8	**70.6**

Visual Question Answering in Remote Sensing

Post-flood Remote Sensing VQA

We introduce the DiViNeCaps (DistilBERT Vision Network with CapsuleNet classifier) network for visual question answering (VQA) in post-flood damage assessment, focusing on counting-oriented queries and mapping each question-image pair to a mixed answer space that includes both categorical responses and numerical counts. The proposed pipeline (seen in **Figure 8**) uses DistilBERT to encode the input question and Vision Mamba to compute image embeddings that preserve visual evidence relevant to fine-grained counting, then integrates the two modalities with an attention-based fusion procedure that applies self-attention over textual features, gated attention over visual features, and bidirectional cross-attention to align question semantics with image content. Prediction is performed by a Capsule Network head that operates on the fused representation to support structured category prediction and count regression within a unified formulation, and training is driven by a joint objective that combines cross-entropy for categorical answers with RMSE for numerical outputs to reflect the FloodNet [9] answer vocabulary.

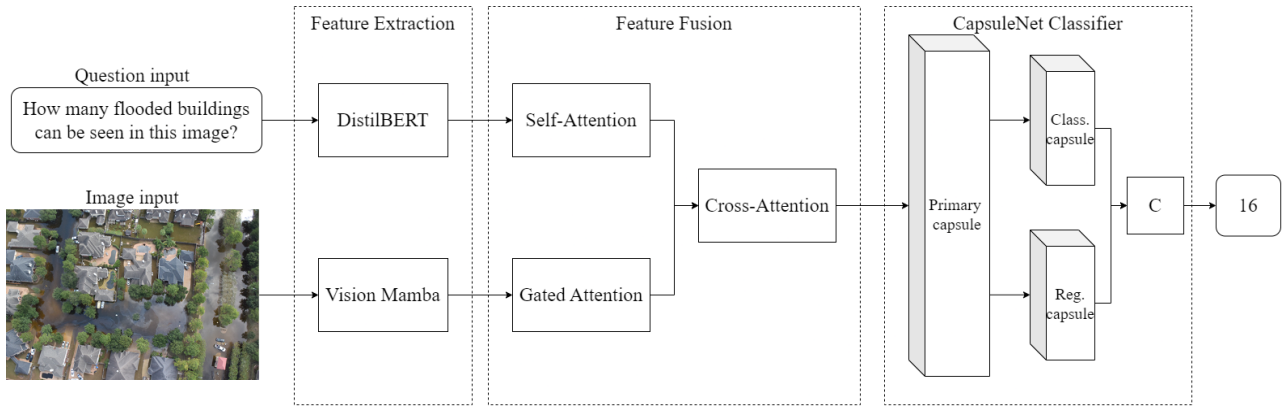

Figure 8. Overview of the proposed DiViNeCaps (DistilBERT Vision Network with CapsuleNet classifier) network architecture for VQA on the FloodNet dataset. The network consists of three main components: Feature Extraction, Feature Fusion, and CapsuleNet Classifier. Feature Extraction uses DistilBERT to process the input question and Vision Mamba to process the image. Feature Fusion employs Self-Attention and Gated Attention mechanisms, which are combined using Cross-Attention. The CapsuleNet Classifier then classifies the concatenated features into the final output, determining the number of flooded buildings in the image.

DiViNeCaps is evaluated on the FloodNet post-flood remote sensing VQA benchmark. The method is compared against a diverse set of VQA baselines ranging from early LSTM+VGG16 architectures and attention models (SAN [10], MFB with Co-Attention [11]) to recent pretrained language and vision approaches employing RoBERTa or DistilBERT-style encoders and ConvNeXt, ResNet-152, or CLIP visual backbones with fusion by concatenation, elementwise multiplication, or MFB-based attention. Relative to these baselines, DiViNeCaps combines DistilBERT question encoding with Vision Mamba image features, an attention-based fusion pipeline (self-attention, gated attention, and cross-attention), and a CapsuleNet prediction head trained with a joint objective that couples cross-entropy for categorical answers with RMSE for count regression to match FloodNet's mixed answer space. As reported in **Table 4**, the proposed model achieves the best overall accuracy (98.84%) and ranks first across all reported question categories, including counting (40.02% Simple Count and 40.78% Complex Count) and classification-style questions (98.92% Yes/No, 98.73% Image Condition, and 98.55% Road Condition). Qualitative results in **Figure 9** support the

quantitative trends and indicate that the primary failure cases occur on questions involving very large building counts, where predictions become less reliable; compared to strong recent baselines such as CNX-mul and Grad-Cam, DiViNeCaps improves counting through its dedicated regression component, while the closest overall competitor, CLIP-mul-taskwise, does not report counting performance.

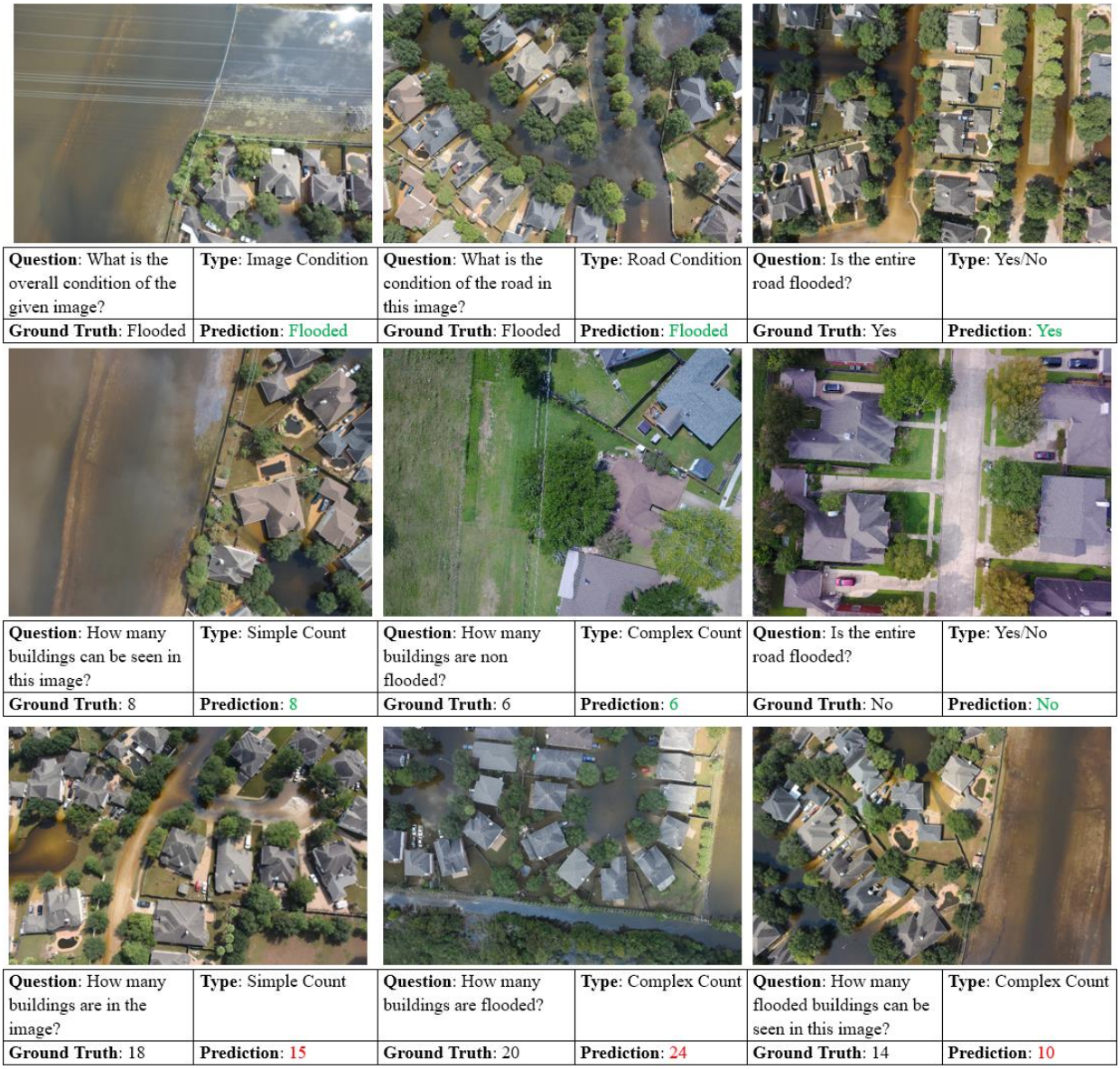

Figure 9. Results of our proposed network on the FloodNet dataset, with the first two rows demonstrating successful predictions for various question types (Image Condition, Road Condition, Yes/No, Simple Count, Complex Count). The bottom row highlights scenarios where the prediction model struggled, particularly with Simple Count and Complex Count questions.

Table 4. Performance comparison of various VQA networks on the FloodNet dataset across multiple question types, sorted by overall accuracy score. The bolded and underlined values represent the best and second-best results across each category.

Network	Overall	Simple Count	Complex Count	Yes/No	Image Condition	Road Condition
MFB with Co-Attention [130]	21.00	21.00	21.00	98.00	96.00	-
SAN [129]	32.00	32.00	32.00	62.00	96.00	-
Concatenation [125]	52.00	6.00	3.00	31.00	88.00	-
RSVQA [141]	64.98	17.14	17.89	50.17	86.12	81.08
Point-wise Multiplication [128]	77.00	30.00	28.00	97.00	97.00	-
DATWEP [139]	78.00	27.00	28.00	**99.00**	98.00	-
CNX-mul [132]	81.26	<u>37.07</u>	<u>37.40</u>	98.31	<u>98.62</u>	97.18
Grad-Cam [135]	82.00	36.00	32.00	-	94.00	<u>98.00</u>
CLIP-mul-taskwise [136]	<u>98.33</u>	-	-	98.85	98.43	97.71
DiViNeCaps	**98.84**	**40.02**	**40.78**	<u>98.92</u>	**98.73**	**98.55**

Robust Counting for Post-Disaster VQA

We propose the DeVANet architecture (**Figure 10**) for post-disaster visual question answering (VQA) in remote sensing imagery, with a primary focus on improving object counting accuracy through efficient multimodal fusion. DeVANet uses DeBERTa to encode the input question and a Vision Mamba visual backbone augmented with a Local-Global Attention (LGA) module to preserve both fine-scale evidence and broader contextual cues required for counting in cluttered scenes, presented in **Figure 11**. The image representation is formed by guiding Vision Mamba features with the local-global feature map to emphasize consistently informative regions, and the text and image features are integrated through a bidirectional fusion module that combines modality-specific self-attention with bidirectional cross-attention and co-attention to promote joint reasoning. For prediction, DeVANet employs two task-specific MLP heads, one for categorical answers and one for count estimation, selected according to the question type, and it trains the model using a joint objective that pairs weighted cross-entropy for imbalanced classification with a Negative Binomial loss to model overdispersed count data.

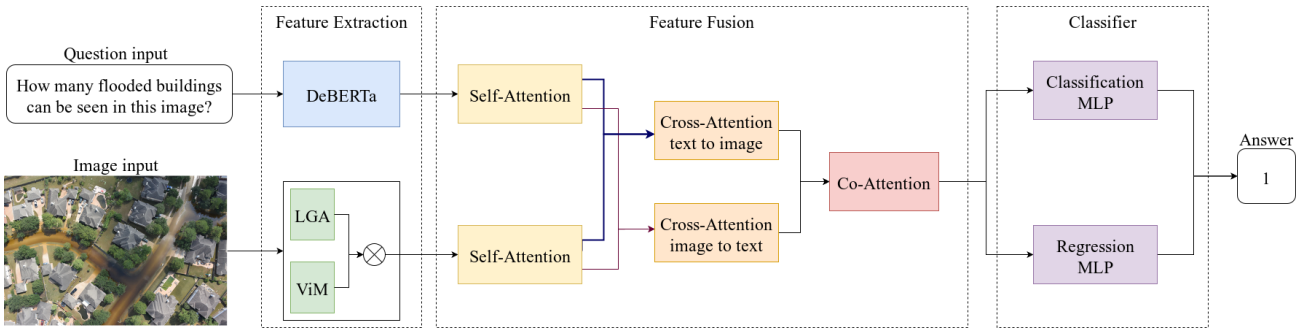

Figure 10. Overview of DeVANet architecture. The system takes two inputs, a question and an image, which are processed by DeBERTA and LGA-ViM to extract text and image features. These are processed through a bidirectional feature fusion module, and lastly passed through a dual MLP classifier to output the final answer.

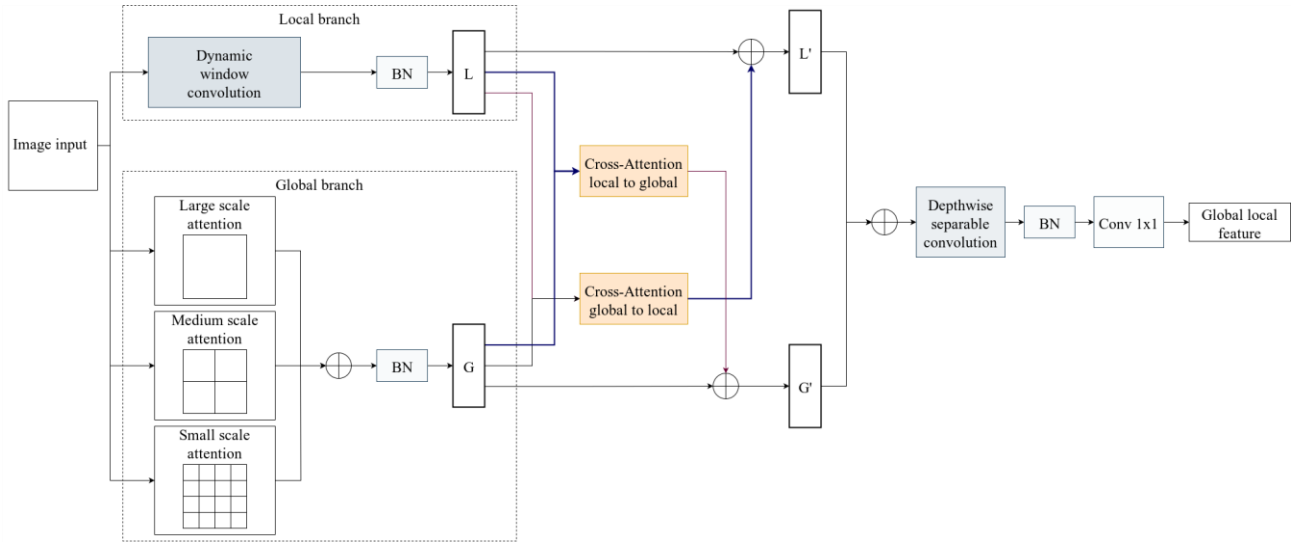

Figure 11. Architecture of LGA, highlighting the local and global branch components.

We evaluate DeVANet on FloodNet and RescueNet [12] post-disaster datasets. The method is compared against representative multimodal baselines from natural-image and remote-sensing VQA, including MCAN [13], VisualBERT [14], ViLBERT [15], UNITER [16], RSVQA [17], SHRNet [18], SAM-VQA [19], and RSAdapter [20], covering non-transformer co-attention models, single and dual-stream transformers, and remote-sensing adaptations. On FloodNet, DeVANet achieves the highest overall accuracy (98.93%) and substantially improves counting performance (59.72% Simple Count, 54.48% Complex Count) compared to the strongest baselines RSAdapter and UNITER, as seen in **Table 5**. Counting-specific error analysis further confirms improved robustness across both low-count and high-count regimes, where DeVANet attains the lowest RMSE/MAE and reduced logarithmic errors (LRMSE/LMAE), indicating better relative accuracy as object quantities increase (**Table 6**). On RescueNet, which introduces denser scenes and broader task diversity, DeVANet again ranks first overall (88.47%), improving over RSAdapter by 2.24% and over UNITER by 3.74%, with the largest gains on counting (60.11% Simple Count, 55.32% Complex Count) and consistent improvements on damage-related and reasoning tasks (**Table 7**). The corresponding counting error metrics on RescueNet show the same stability trend for higher counts, with DeVANet yielding the lowest errors across RMSE/MAE and their logarithmic variants (**Table 8**). An efficiency comparison on RescueNet shows that DeVANet achieves the best accuracy with competitive computational cost, comparable to transformer baselines. Extensive ablations on RescueNet identify the main contributors: 512x512 image pixel inputs provide the best accuracy–cost trade-off; DeBERTa is the strongest language encoder; the proposed LGA–ViM visual backbone yields the largest improvements on challenging tasks; three LGA scales balance accuracy and efficiency; Bidirectional Attention Fusion is the most effective fusion strategy; and a dual-MLP predictor trained with Negative Binomial loss delivers the best overall performance, particularly for complex counting and other imbalanced categories.

Table 5. Accuracy performance evaluation of DeVANet on FloodNet dataset.

Network	Overall	Simple Count	Complex Count	Yes/No	Image Condition	Road Condition
MCAN [240]	70.88	23.45	20.43	96.45	96.11	96.04
VisualBERT [241]	82.82	32.53	29.02	96.90	96.94	97.04
RSVQA [41]	88.50	30.10	27.85	96.80	97.40	97.55
ViLBERT [242]	90.76	35.61	33.01	97.34	97.77	98.42
SHRNet [144]	92.20	36.25	32.50	97.60	98.10	98.20
UNITER [243]	94.69	39.69	35.09	97.79	98.60	98.75
RSAdapter [42]	96.10	45.85	41.60	98.40	98.75	98.80
DeVANet	**98.93**	**59.72**	**54.48**	**99.29**	**98.92**	**99.03**

Table 6. Counting error metrics on FloodNet for small (<10) and larger (>=10) object counts.

	RMSE		MAE		LRMSE		LMAE	
Network	< 10	≥ 10	< 10	≥ 10	< 10	≥ 10	< 10	≥ 10
MCAN [240]	2.83	9.66	1.98	7.41	0.72	1.35	0.48	1.02
RSVQA [41]	2.60	8.94	1.75	6.84	0.68	1.28	0.44	0.96
VisualBERT [241]	2.18	6.74	1.43	5.17	0.55	0.98	0.36	0.74
SHRNet	2.08	6.32	1.33	4.80	0.52	0.94	0.34	0.71
ViLBERT [242]	1.95	6.02	1.38	4.57	0.50	0.89	0.33	0.68
UNITER [243]	1.91	5.83	1.25	4.47	0.49	0.88	0.32	0.67
RSAdapter [42]	1.52	4.27	0.93	3.15	0.38	0.62	0.24	0.45
DeVANet	**1.19**	**3.23**	**0.70**	**2.35**	**0.29**	**0.41**	**0.18**	**0.31**

Table 7. Accuracy performance evaluation of DeVANet on RescueNet dataset.

Network	Overall	Simple Count	Complex Count	Building Condition	Road Condition	Level of Damage	Risk Assessment	Density Estimation	Positional	Change Detection
MCAN [240]	75.89	30.04	24.76	90.32	92.87	85.11	72.35	70.27	69.49	70.85
RSVQA [41]	77.42	32.45	27.21	91.02	93.48	84.11	73.12	71.85	70.42	72.36
VisualBERT [241]	78.27	33.56	28.93	90.72	94.21	84.85	74.48	74.69	72.67	74.13
ViLBERT [242]	80.13	36.56	34.12	91.34	95.12	85.56	76.23	74.78	75.12	74.56
SHRNet [144]	81.95	37.82	33.65	92.85	96.21	86.90	77.10	75.42	75.26	74.95
UNITER [243]	84.73	40.12	36.57	94.11	97.32	87.78	78.49	76.94	76.57	75.33
RSAdapter [42]	86.23	46.90	42.77	95.23	97.85	88.73	79.64	78.15	79.12	77.88
DeVANet	**88.47**	**60.11**	**55.32**	**96.74**	**98.29**	**89.72**	**81.79**	**80.36**	**83.64**	**79.12**

Table 8. Counting error metrics on RescueNet for small (<10) and larger (>=10) object counts.

	RMSE		MAE		LRMSE		LMAE	
Network	< 10	≥ 10	< 10	≥ 10	< 10	≥ 10	< 10	≥ 10
MCAN [240]	3.41	12.80	2.37	9.60	0.85	1.62	0.58	1.18
RSVQA [41]	3.26	11.92	2.15	8.91	0.81	1.55	0.54	1.11
VisualBERT [241]	2.77	9.19	1.89	6.80	0.68	1.18	0.45	0.87
SHRNet [144]	2.68	8.73	1.74	6.54	0.65	1.12	0.43	0.83
ViLBERT [242]	2.33	8.45	1.67	6.32	0.63	1.10	0.41	0.82
UNITER [243]	2.51	8.12	1.66	6.13	0.62	1.05	0.41	0.78
RSAdapter [42]	2.02	6.07	1.24	4.45	0.48	0.78	0.31	0.56
DeVANet	**1.53**	**4.66**	**0.98**	**3.22**	**0.36**	**0.52**	**0.23**	**0.38**

Change Detection in Remote Sensing

Region-Aware Change Detection

We introduce SFCNet (Semantic Field-aware Change Network), a change detection architecture designed for structured agricultural landscapes in which changes tend to follow coherent field boundaries rather than isolated pixels. As shown in **Figure 12**, SFCNet follows a region-aware pipeline that extracts multiscale convolutional features from bitemporal inputs, aggregates them into field-aligned region tokens using a Semantic Region Tokenizer (SRT)

driven by predefined crop-region masks, and performs region-level temporal reasoning with a Field-Aware Region Encoder (FARE) based on slot-style token grouping to capture shared temporal patterns across fields. The resulting context-aware embeddings are decoded back to dense spatial representations through a Contextual Crop Decoder (CCD), enabling pixel-level reconstruction that preserves field structure, and a subtraction-based prediction head produces the final change map, trained with a binary cross-entropy objective.

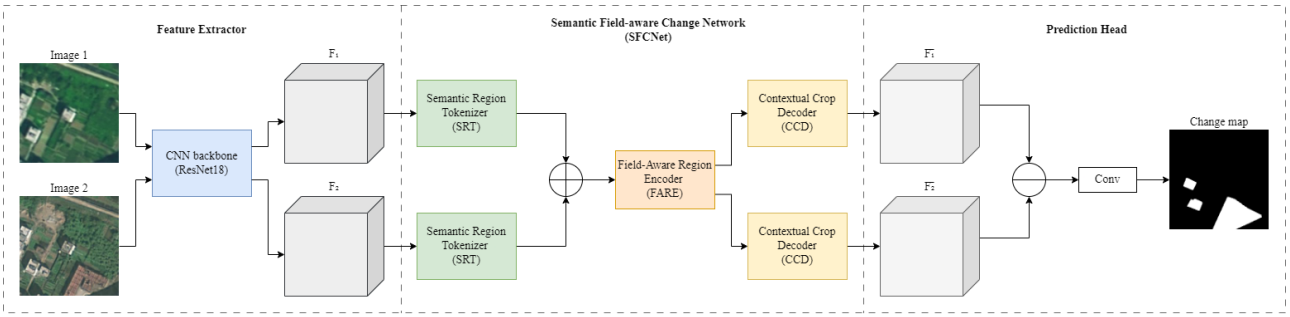

Figure 12. Overview of the proposed SFCNet architecture for change detection in structured agricultural regions. The model consists of a shared CNN backbone for bitemporal feature extraction, a Semantic Region Tokenizer (SRT), a Field-Aware Region Encoder (FARE), and two Contextual Crop Decoders (CCD), followed by a subtraction and convolutional prediction head that produces the final change map.

We evaluate SFCNet on the CLCD fine-grained cropland change detection benchmark [21], and compare it against seven representative CNN, hybrid, and transformer-based baselines (SiamUNet-Diff [22], A2Net [23], BIT [24], ChangeFormer [25], ICIF-Net [26], DMINet [27], and RDP-Net [28]). As reported in **Table 9**, SFCNet achieves the best results across all metrics, reaching 52.47% IoU and 67.23% F1 with balanced precision and recall (67.85% and 68.37%), and improving over the strongest baseline BIT by 3.31% IoU. Qualitative comparisons in **Figure 13** show that SFCNet produces cleaner boundaries and fewer false detections, reducing both over-segmentation observed in transformer baselines and missed thin structures typical of CNN-based models. Finally, the efficiency analysis indicates that SFCNet attains the highest IoU and accuracy while maintaining competitive computational cost, outperforming heavier transformer designs such as ChangeFormer and remaining comparable in latency to lighter baselines.

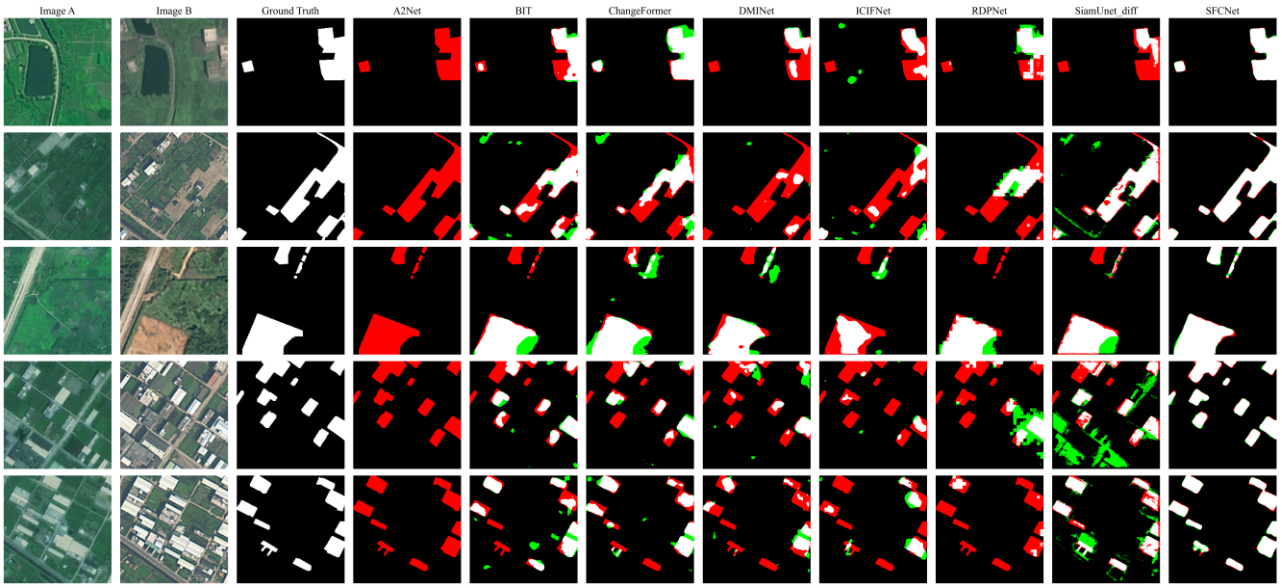

Figure 13. Qualitative comparison of change detection results on the CLCD dataset. White: true positives; Red: false negatives; Green: false positives; Black: true negatives.

Table 9. Quantitative comparison of change detection models on the CLCD dataset.

Model	IoU	Accuracy	F1	Precision	Recall
A2Net [48]	23.42	92.04	37.95	45.18	32.71
SiamUNet-Diff [45]	32.57	93.58	49.13	59.83	41.68
RDPNet [47]	37.28	92.88	54.31	51.95	56.90
ICIFNet [54]	39.17	94.16	56.29	63.54	50.52
DMINet [55]	40.28	93.91	57.42	59.86	55.18
ChangeFormer [52]	45.07	94.31	62.14	61.50	62.79
BIT [51]	49.16	94.88	65.91	65.26	66.58
SFCNet	**52.47**	**95.20**	**67.23**	**67.85**	**68.37**

Building Change Detection

We propose RADNet (Region-Aware Dual-path Network), a token-based global-local change detection architecture for high-resolution urban imagery that focuses on building changes and aims to reduce common errors caused by structural clutter, viewpoint shifts, and illumination variation. As presented in **Figure 14**, RADNet extracts multi-scale bitemporal features with a shared convolutional backbone and converts them into building-aligned tokens using an Instance-Consistent Tokenizer (ICT) that infers soft instance regions from attention cues and spatial continuity, enabling semantically consistent tokenization without relying on external instance masks, shown in **Figure 15**. The resulting tokens are refined by a Global-Local Attention Encoder (GLAE) that combines global token interactions with localized geometric evidence to preserve both long-range context and boundary-level detail, and a Cross-Temporal Cross-Scale Decoder (CTCSD) aggregates information across pyramid levels while keeping time-specific embeddings separate for instance-level comparison. A Correlation-Based Prediction Head (CBPH) reconstructs dense representations from instance embeddings and estimates change likelihood through similarity-based differencing, supporting spatially coherent change maps that attenuate appearance variations not related to structural change.

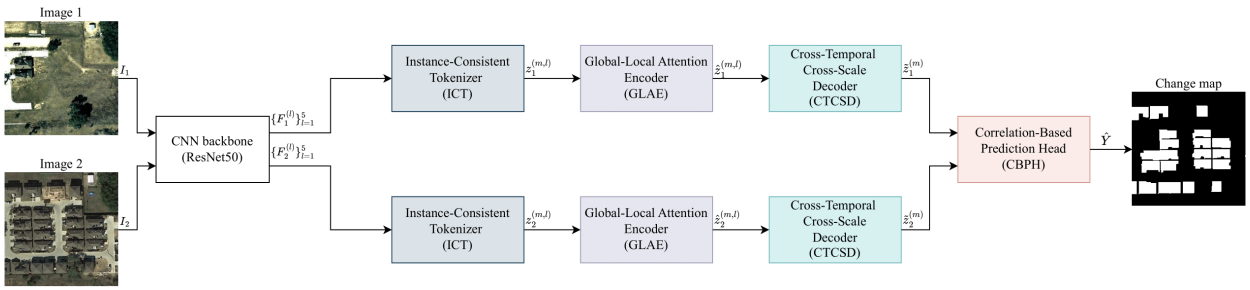

Figure 14. RADNet architecture. The model processes two input images through a shared ResNet50 backbone to extract multi-scale features, which are then tokenized, contextually enriched, and temporally aligned through a series of instance-aware modules. The final Correlation-Based Prediction Head (CBPH) produces a dense change probability map.

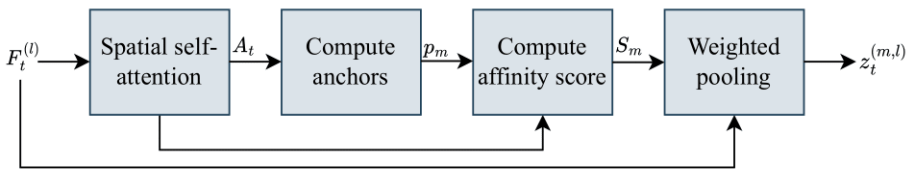

Figure 15. Instance-Consistent Tokenizer (ICT). Attention responses are used to compute peak anchor points, from which spatial affinity maps are derived to form soft masks. These masks are then used to extract instance-level tokens via weighted pooling.

We evaluate RADNet on three public building change detection benchmarks (EGY-BCD [29], LEVIR-CD [30], and SYSU-CD [31]), and compare it against representative CNN, hybrid, and transformer-based baselines (SiamUNet-Diff, A2Net, BIT, CASP-Mb0 [32], ChangeFormer, ICIF-Net, DMINet, and RDP-Net). On EGY-BCD, **Table 10** shows that RADNet achieves the best performance across all metrics (67.51% IoU, 81.27% F1), improving over BIT and CASP-Mb0 while maintaining the strongest precision–recall balance. On LEVIR-CD, **Table 11** reports that RADNet again ranks first (82.34% IoU, 90.26% F1, 99.05% accuracy), confirming consistent gains on large-scale building change detection, and ablation studies further indicate a favorable accuracy–efficiency trade-off relative to transformer-heavy designs (notably lower FLOPs and substantially reduced training time compared to ChangeFormer) while remaining competitive in inference time. On SYSU-CD, **Table 12** shows that RADNet attains the highest IoU (65.42%) and F1 (78.12%), with the best recall (82.38%), indicating improved sensitivity to diverse change patterns under heterogeneous scene content. Qualitative results presented in **Figure 16** corroborate the quantitative trends across datasets, showing cleaner building boundaries, fewer spurious detections, and less fragmented change maps. Finally, ablations on LEVIR-CD show complementary contributions from instance-consistent tokenization (largest drop when removed), global–local attention encoding, cross-scale temporal decoding, and correlation-based prediction, supporting the role of token alignment and correlation reasoning in robust building change localization.

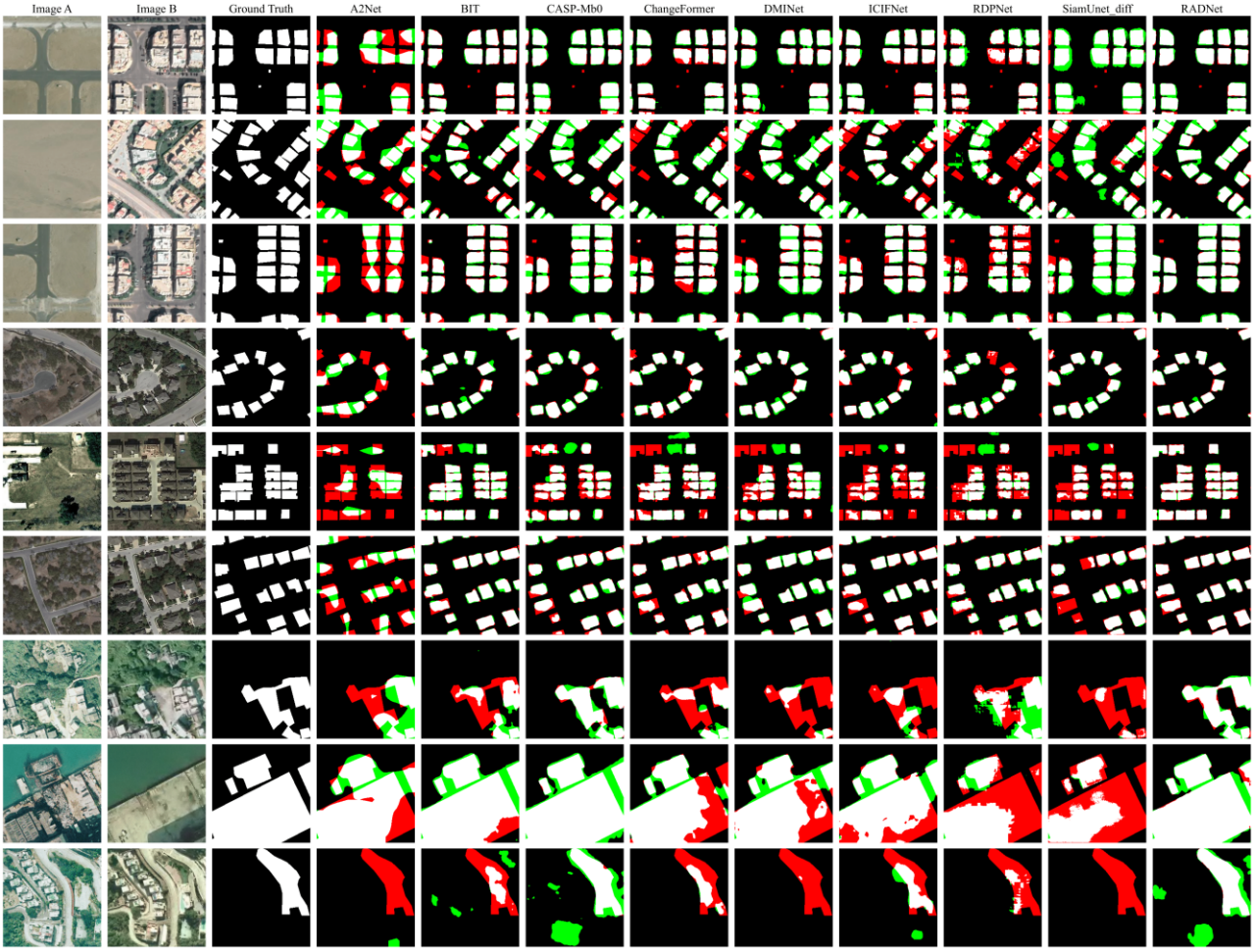

Figure 16. Visual comparison of change detection results. Rows 1-3 show examples from the EGY-BCD dataset, rows 4-6 are from the LEVIR-CD dataset, and rows 7-9 are from the SYSU-CD dataset. White: true positives, Red: false negatives, Green: false positives, Black: true negatives.

Table 10. Quantitative results on the EGY-BCD dataset.

Model	IoU	Accuracy	F1	Precision	Recall
SiamUNet-Diff [45]	47.29	95.24	64.22	61.91	66.70
RDPNet [47]	52.20	96.12	68.59	71.25	66.12
ICIFNet [54]	52.71	96.21	69.03	72.41	65.95
A2Net [48]	52.79	96.23	69.10	72.81	65.75
DMINet [55]	53.02	95.85	69.29	65.90	73.06
ChangeFormer [52]	56.03	96.51	71.82	74.32	69.48
CASP-Mb0 [246]	64.85	96.88	78.43	79.81	77.09
BIT [51]	65.38	97.35	79.06	80.16	78.00
RADNet	**67.51**	**98.15**	**81.27**	**82.36**	**80.50**

Table 11. Quantitative results on the LEVIR-CD dataset.

Model	IoU	Accuracy	F1	Precision	Recall
A2Net [48]	50.20	96.99	66.84	76.14	59.57
SiamUNet-Diff [45]	64.23	98.05	78.22	90.73	68.74
RDPNet [47]	67.47	98.08	80.57	83.15	78.16
ICIFNet [54]	68.95	98.27	81.62	88.71	75.58
ChangeFormer [52]	69.96	98.26	82.32	85.17	79.67
DMINet [55]	72.87	98.45	84.31	87.08	81.71
CASP-Mb0 [246]	78.12	98.87	87.30	88.21	86.41
BIT [51]	79.91	98.88	88.84	90.45	87.28
RADNet	**82.34**	**99.05**	**90.26**	**91.58**	**89.04**

Table 12. Quantitative results on the SYSU-CD dataset.

Model	IoU	Accuracy	F1	Precision	Recall
SiamUNet-Diff [45]	42.77	85.46	59.91	85.62	46.08
ICIFNet [54]	54.21	86.41	70.30	72.51	68.23
DMINet [55]	59.54	88.81	74.64	80.14	69.84
BIT [51]	59.96	87.30	74.97	70.03	80.65
RDPNet [47]	60.47	88.52	75.36	76.30	74.45
ChangeFormer [52]	61.21	89.12	75.93	79.37	72.78
A2Net [48]	61.27	88.12	75.98	72.63	79.66
CASP-Mb0 [246]	63.52	89.07	76.61	73.40	80.19
RADNet	**65.42**	**90.45**	**78.12**	**75.25**	**82.38**

Future Work

We identify future work directions along three complementary axes: data, modeling, and deployment. On the data side, a priority is to expand controllable synthetic generation toward richer ecological diversity and more realistic sensing effects (e.g., seasonal phenology, shadows, atmospheric haze, motion blur, and sensor noise), and to establish stronger links between simulated and real imagery through domain adaptation and cross-domain calibration. On the modeling side, the next steps include uncertainty-aware learning (e.g., Bayesian approximations and calibrated confidence estimation) to support risk-sensitive decision making, together with more explicit incorporation of geometric priors and boundary cues for segmentation and change detection, and improved multimodal counting via distributional objectives and region-conditioned reasoning. For temporal analysis, extending change detection from bi-temporal inputs to longer sequences is a natural direction, enabling persistence-aware change characterization and trend estimation rather than binary change maps. Finally, toward operational use, future work should emphasize efficiency and

reliability: lightweight backbones for edge inference on UAV platforms, robustness testing under adverse conditions, and human-in-the-loop evaluation protocols that quantify how model explanations and confidence scores impact downstream inspection and navigation decisions.

Personal Contributions

This thesis addresses the overarching objective of advancing visual perception and scene understanding methods that support control and navigation-relevant decision making in aerial and remote sensing contexts. The personal contributions align with this objective by developing data and evaluation resources that reflect operational constraints, proposing architectures tailored to the structure of aerial imagery and multimodal supervision, and providing systematic experimental evidence through comparisons, ablations, and failure analyses across representative tasks.

For the objective of improving UAV and aerial semantic understanding, the thesis contributes a structured perspective on how data availability, label quality, and controllable variability affect segmentation performance and generalization. In this direction, the work emphasizes simulation-driven data generation as a practical mechanism to obtain scalable, densely annotated imagery under controlled acquisition factors (altitude, viewpoint, and weather). The resulting resources and analyses are used to study model behavior under domain shift and to clarify which scene factors most strongly influence class-wise segmentation robustness in forest and vegetation-centric environments.

For the objective of enhancing multimodal reasoning in remote sensing, the thesis develops a Visual Question Answering (VQA) pipeline specialized for post-disaster assessment, with a particular focus on counting accuracy and answer reliability under heterogeneous scene content. The contributions include a coherent multimodal formulation that combines language representations with visual features and an explicit treatment of counting as a distribution-sensitive prediction problem. Through targeted design choices and controlled studies, the thesis identifies practical mechanisms that reduce common counting errors (undercounting in cluttered regions and overcounting due to repeated textures), while keeping the architecture compatible with remote-sensing image resolutions and dataset constraints.

For the objective of improving temporal reasoning and fine-grained change detection, the thesis introduces an encoder-free change detection framework that directly models cross-temporal feature interactions and emphasizes locality-aware reasoning for small and subtle changes. The work contributes modular components for temporal fusion and discrepancy modeling, together with an evaluation methodology that isolates the impact of each design choice via ablation studies. Beyond aggregate scores, the thesis provides qualitative and error-oriented analyses to characterize typical failure cases (illumination-induced false positives and boundary ambiguities) and to motivate design decisions that trade off sensitivity and precision in operational settings.

In summary, we bring the following original contributions:

Semantic segmentation for UAV and aerial scene understanding.

1. A synthetic UAV dataset for forest semantic segmentation, designed with controllable acquisition conditions and dense pixel-level annotations to support reproducible training and evaluation in forest environments.
2. A weakly supervised spatio-temporal segmentation framework for UAV videos, based on label propagation and temporal refinement via an ST-GRU module, including an efficiency-oriented GRU variant for practical deployment.
3. A transformer-based semantic segmentation architecture (SwinFAN) that strengthens global-local feature integration for high-resolution aerial imagery via decoder-side attention and feature refinement.
4. Methodological improvements in the SwinFAN decoder to improve global context modeling and boundary precision in UAV segmentation:
    a) Guided Focal-Axial Attention (GFA) for global-local context fusion, integrating focal attention to emphasize salient regions at multiple scales with axial attention to efficiently capture long-range dependencies along spatial axes, enabling accurate segmentation of both large structures and small objects in high-resolution UAV imagery.
    b) Attention-Based Feature Refinement Head (AFRH) for boundary and detail enhancement, refining decoder outputs through self-attention and convolutional operations to improve object boundary delineation and recognition of small or underrepresented classes.
5. A validation protocol combining benchmarking, ablation studies, and transfer evaluation, supporting a structured analysis of dataset utility and model behavior under domain shift.

Remote-sensing VQA with emphasis on post-disaster assessment and counting.

1. A task-driven analysis of remote-sensing VQA, with emphasis on the difficulty of counting under scale variation, clutter, and long-tailed answer distributions.
2. DiViNeCaps, a post-flood VQA architecture that integrates text and image representations through multimodal fusion and a structured classification module, evaluated on FloodNet.
3. DeVANet, a counting-oriented VQA framework for remote-sensing VQA, integrating multi-scale image embedding, multimodal feature fusion, and dual-head prediction for discrete and continuous count modeling.
4. Methodological improvements introduced in DeVANet to improve counting robustness under scale variation and imbalanced count distributions:
    a. Local-Global Attention Vision Mamba (LGA-ViM) for multi-scale image embedding, integrating local-global attention with Vision Mamba through element-wise multiplication to strengthen representations across multiple spatial scales in high-resolution aerial imagery.
    b. Bidirectional Fusion Attention (BFA) for multimodal feature integration, combining self-attention in each modality with bidirectional cross-attention and co-attention to improve alignment between questions and relevant image regions.
    c. A distribution-aware loss formulation for imbalanced and overdispersed counts, combining weighted cross-entropy loss for discrete classification with negative binomial loss for count regression to address class imbalance and

overdispersion, improving optimization stability and counting accuracy for large object counts.

5. A comprehensive experimental evaluation on FloodNet and RescueNet, supported by ablation studies that isolate the impact of the visual backbone, fusion mechanism, prediction heads, and loss design.

Fine-grained change detection with structure-aware tokenization and temporal reasoning.

1. A structured analysis of change detection representation choices, motivating structure-aligned tokenization and temporal reasoning in domains where objects and coherent regions define the change semantics.
2. SFCNet, a region-aware change detection framework for structured agricultural scenes, including a semantic region tokenizer, a region-level temporal encoder, and a region-to-pixel decoding strategy for dense change map prediction.
3. RADNet, an instance-oriented change detection framework for building change detection, including instance-consistent tokenization, global-local attention encoding, and boundary-aware decoding for precise change localization.
4. Methodological improvements introduced in RADNet to support instance-level semantic alignment and contextual reasoning in dense urban scenes:
    a. Instance-Consistent Tokenizer (ICT) for building-level semantic alignment, generating semantically aligned region tokens corresponding to individual buildings or coherent structures by leveraging internal attention cues to infer region boundaries directly from feature maps, enabling instance-aware reasoning without requiring ground-truth segmentation masks during inference.
    b. Global-Local Attention Encoder (GLAE) for contextual and geometric reasoning, refining instance-level tokens by jointly modeling global semantic relationships across regions using multihead self-attention and local geometric consistency using convolutional attention, supporting accurate building-level change detection under dense urban layouts.
5. A unified evaluation and ablation methodology across change detection benchmarks, quantifying the role of tokenization, temporal fusion, decoding strategy, and prediction head design.

Dissemination

Publications in ISI journals (4 as first author):

1. **Semantic Segmentation of Remote Sensing Images With Transformer-Based U-Net and Guided Focal-Axial Attention.** Bianca-Cerasela-Zelia Blaga and Sergiu Nedevschi. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 18303–18318, 2024. DOI: 10.1109/JSTARS.2024.3470316. **Q1, IF: 5.3**
2. **Data Augmentation for Environment Perception with Unmanned Aerial Vehicles.** Vivian Chiciudean, Horațiu Florea, Bianca-Cerasela-Zelia Blaga, Radu Beche, Florin Oniga, and Sergiu Nedevschi. *IEEE Transactions on Intelligent Vehicles*, 2024. DOI: 10.1109/TIV.2024.3374117. **Q1, IF: 14.3**
3. **Token-Based Global–Local Framework for Building Change Detection.** Bianca-Cerasela-Zelia Blaga. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 24162–24173, 2025, DOI: 10.1109/JSTARS.2025.3609073. **Q1, IF: 5.3**

4. **Forest Inspection Dataset: A Synthetic UAV Dataset for Semantic Segmentation of Forest Environments.** Bianca-Cerasela-Zelia Blaga and Sergiu Nedevschi. *Scientific Data (Springer Nature)*, 2026. DOI: 10.1038/s41597-026-06665-x. **Q1, IF: 6.9**
5. **Improving Counting Accuracy of Post-Disaster Visual Question Answering for Remote Sensing.** Bianca-Cerasela-Zelia Blaga and Sergiu Nedevschi. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, vol. 19, pp 10065-10082, 2026, DOI: 10.1109/JSTARS.2026.3670705. **Q1, IF: 5.3**

Presentations at international conferences (5 as first author):

1. **A Critical Evaluation of Aerial Datasets for Semantic Segmentation.** Bianca-Cerasela-Zelia Blaga and Sergiu Nedevschi. *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, 2020, pp. 353–360. DOI: 10.1109/ICCP51029.2020.9266169.
2. **Weakly Supervised Semantic Segmentation Learning on UAV Video Sequences.** Bianca-Cerasela-Zelia Blaga and Sergiu Nedevschi. *2021 29th European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, 2021, pp. 731–735. DOI: 10.23919/EUSIPCO54536.2021.9616055.
3. **Employing Simulators for Collision-Free Autonomous UAV Navigation.** Bianca-Cerasela-Zelia Blaga and Sergiu Nedevschi. *2022 IEEE 18th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, 2022, pp. 313–318. DOI: 10.1109/ICCP56966.2022.10053959.
4. **Static Mesh Enrichment with Dynamic Entities for Training Sets Generation.** Vivian Chiciudean, Radu Beche, Bianca-Cerasela-Zelia Blaga, Florin Oniga, and Sergiu Nedevschi. *2023 IEEE 19th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, 2023, pp. 111–119. DOI: 10.1109/ICCP60212.2023.
5. **Improving VQA Counting Accuracy for Post-Flood Damage Assessment.** Bianca-Cerasela-Zelia Blaga and Sergiu Nedevschi. *2024 IEEE 20th International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, 2024, pp. 1–8. DOI: 10.1109/ICCP63557.2024.10793007.
6. **Region-Aware Token Aggregation for Fine-Grained Change Detection.** Bianca-Cerasela-Zelia Blaga. *2025 IEEE 21st International Conference on Intelligent Computer Communication and Processing (ICCP)*, Cluj-Napoca, Romania, 2025.

Workshop presentations:

1. **Forest Inspection Dataset for Aerial Semantic Segmentation and Depth Estimation.** Bianca-Cerasela-Zelia Blaga and Sergiu Nedevschi. *arXiv preprint* arXiv:2403.06621, 2024. Presented at the workshop "Large aerial dataset creation using real and synthetic data for forest monitoring", ICCP 2021.

The research presented in this thesis was carried out within several projects of the Image Processing and Pattern Recognition Research Center, coordinated by Prof. Dr. Eng. Sergiu Nedevschi. The contributions presented here were produced as part of the following projects:

1. SEPCA – "Integrated Semantic Visual Perception and Control for Autonomous Systems", grant funded by the Romanian Ministry of Education and Research, code PN-III-P4-ID-PCCF-2016-0180.
2. DeepPerception – "Deep Learning Based 3-D Perception for Autonomous Driving", grant funded by the Romanian Ministry of Education and Research under Grant PN-III-P4-PCE-2021-1134.
3. Distributed Sensemaking – project funded by Lockheed Martin Australia.

References

[1] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," ISPRS J. Photogramm. Remote Sens., vol. 165, pp. 108–119, 2020.

[2] A. Marcu, V. Licaret, D. Costea, and M. Leordeanu, "Semantics through Time: Semi-supervised Segmentation of Aerial Videos with Iterative Label Propagation," in Proceedings of the Asian Conference on Computer Vision (ACCV), November 2020.

[3] H. Florea, V.-C. Miclea, and S. Nedevschi, "Wilduav: Monocular uav dataset for depth estimation tasks," in 2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP). IEEE, 2021, pp. 291–298.

[4] M. Fonder and M. Van Droogenbroeck, "Mid-Air: A multi-modal dataset for extremely low altitude drone flights," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[5] Z. Feng, Y. She, and S. Keshav, "SPREAD: A large-scale, high-fidelity synthetic dataset for multiple forest vision tasks," Ecol. Inform., vol. 87, p. 103085, 2025.

[6] ISPRS, "ISPRS Benchmark on Semantic Labeling in Urban Remote Sensing - 2D Semantic Labeling - Potsdam," https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx.

[7] ISPRS, "ISPRS Benchmark on Semantic Labeling in Urban Remote Sensing - 2D Semantic Labeling - Vaihingen," https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx.

[8] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation," in Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, J. Vanschoren and S. Yeung, Eds., vol. 1, 2021.

[9] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding," IEEE Access, vol. 9, pp. 89 644–89 654, 2021.

[10] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked Attention Networks for Image Question Answering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.

[11] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-Modal Factorized Bilinear Pooling With Co-Attention Learning for Visual Question Answering," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1821–1830.

[12] A. Sarkar and M. Rahnemoonfar, "RESCUENet-VQA: A Large-Scale Visual Question Answering Benchmark for Damage Assessment," in IGARSS 2023-2023 IEEE Int. Geosci. Remote Sens. Symp. IEEE, 2023, pp. 1150–1153.

[13] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 6281–6290.

[14] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," arXiv preprint arXiv:1908.03557, 2019.

[15] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining TaskAgnostic Visiolinguistic Representations for Vision-and-Language Tasks," Adv. Neural Inf. Process. Syst., vol. 32, 2019.

[16] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: UNiversal Image-TExt Representation Learning," in Eur. Conf. Comput. Vis. (ECCV). Springer, 2020, pp. 104–120.

[17] C. Chappuis, E. Walt, V. Mendez, S. Lobry, B. L. Saux, and D. Tuia, "The curse of language biases in remote sensing VQA: the role of spatial attributes, language diversity, and the need for clear evaluation," arXiv preprint arXiv:2311.16782, 2023.

[18] Z. Zhang, L. Jiao, L. Li, X. Liu, P. Chen, F. Liu, Y. Li, and Z. Guo, "A spatial hierarchical reasoning network for remote sensing visual question answering," IEEE Trans. Geosci. Remote Sens., vol. 61, pp. 1–15, 2023.

[19] A. Sarkar, T. Chowdhury, R. R. Murphy, A. Gangopadhyay, and M. Rahnemoonfar, "SAM-VQA: Supervised Attention-Based Visual Question Answering Model for Post-Disaster Damage Assessment on Remote Sensing Imagery," IEEE Trans. Geosci. Remote Sens., vol. 61, pp. 1–16, 2023.
[20] Y. Wang and P. Ghamisi, "RSAdapter: Adapting Multimodal Models for Remote Sensing Visual Question Answering," IEEE Trans. Geosci. Remote Sens., 2024.
[21] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-Transformer Network With Multiscale Context Aggregation for Fine-Grained Cropland Change Detection," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 4297–4306, 2022.
[22] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully Convolutional Siamese Networks for Change Detection," in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 4063–4067.
[23] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully Convolutional Siamese Networks for Change Detection," in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 4063–4067.
[24] H. Chen, Z. Qi, and Z. Shi, "Remote Sensing Image Change Detection With Transformers," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–14, 2022.
[25] W. G. C. Bandara and V. M. Patel, "A Transformer-Based Siamese Network for Change Detection," in IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, 2022, pp. 207–210.
[26] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-Scale Cross-Interaction and Inter-Scale Feature Fusion Network for Bitemporal Remote Sensing Images Change Detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–13, 2022.
[27] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change Detection on Remote Sensing Images Using Dual-Branch Multilevel Intertemporal Network," IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–15, 2023.
[28] H. Chen, F. Pu, R. Yang, R. Tang, and X. Xu, "RDP-Net: Region Detail Preserving Network for Change Detection," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–10, 2022.
[29] S. Holail, T. Saleh, X. Xiao, and D. Li, "AFDE-Net: Building Change Detection Using Attention-Based Feature Differential Enhancement for Satellite Imagery," IEEE Geosci. Remote Sens. Lett., vol. 20, pp. 1–5, 2023.
[30] H. Chen and Z. Shi, "A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection," Remote Sens., vol. 12, no. 10, 2020.
[31] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A Deeply Supervised Attention Metric-Based Network and an Open Aerial Image Dataset for Remote Sensing Change Detection," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–16, 2022.
[32] Q. Wang, M. Zhang, J. Ren, and Q. Li, "Exploring Context Alignment and Structure Perception for Building Change Detection," IEEE Trans. Geosci. Remote Sens., vol. 63, pp. 1–10, 2025.