MINISTERUL EDUCAȚIEI ȘI CERCETĂRII

UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA

Calculatoare și Tehnologia Informației

TEZĂ DE DOCTORAT

– REZUMAT –

Metode de percepție vizuală și înțelegere a scenei pentru aplicații de teledetecție

Student-doctorand:
Ing. Bianca-Cerasela-Zelia BLAGA

Conducător științific:
Prof. Dr. Ing. Sergiu NEDEVSCHI

Comisia de evaluare a tezei de doctorat:

Preşedinte: Prof. Dr. Ing. **Radu Gabriel Dănescu** – Universitatea Tehnică din Cluj-Napoca;
Conducător ştiinţific: Prof. Dr. Ing. **Sergiu Nedevschi** – Universitatea Tehnică din Cluj-Napoca;
Referenţi:
- Prof. Dr. Ing. **Vladimir-Ioan Crețu** –Universitatea Politehnica Timișoara;
- Prof. Dr. Ing. **Vasile-Ion Manta** – Universitatea Tehnică „Gheorghe Asachi" din Iaşi;
- Conf. Dr. Ing. **Raluca Didona Brehar** – Universitatea Tehnică din 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

Cuprins teză

1 Introducere 7
1.1 Context 7
1.2 Înţelegerea vizuală în teledetecţie 9
1.3 Provocări 15
1.4 Obiective de cercetare 18
1.5 Contribuţii. 20
1.6 Structura tezei. 23

2 Stadiul actual al cercetării 27
2.1 Segmentare semantică în imagini aeriene şi din UAV 27
2.2 Răspunsuri la întrebări vizuale 33
2.3 Detectarea schimbărilor în teledetecţie 37
2.4 Seturi de date. 41
2.5 Metrici 45

3 Segmentare semantică pentru teledetecţie din UAV 49
3.1 Introducere 49
3.2 Metode 52
3.2 Set de date sintetic pentru segmentarea pădurilor. 52
3.2.2 Segmentare video UAV slab supervizată 55
3.2.3 Segmentare semantică bazată pe transformeri. 58
3.3 Experimente 65
3.3.1 Set de date sintetic pentru segmentarea pădurilor 65
3.3.2 Segmentare video UAV slab supervizată 70
3.3.3 Segmentare semantică bazată pe transformeri 72
3.4 Concluzii 79

4 Răspunsuri la întrebări vizuale în teledetecţie 81
4.1 Introducere 81
4.2 Metode 83
4.2.1 VQA în teledetecţie post-inundaţie. 83
4.2.2 Numărare robustă pentru VQA post-dezastru. 88
4.3 Experimente 96
4.3.1 VQA în teledetecţie post-inundaţie. 96
4.3.2 Numărare robustă pentru VQA post-dezastru. 99
4.4 Concluzii. 107

5 Change Detection in Remote Sensing 109
5.1 Introducere. 109
5.2 Metode. 111
5.2.1 Detectarea schimbărilor bazată pe regiune. 111
5.2.2 Detectarea schimbărilor clădirilor. 115
5.3 Experimente. 122
5.3.1 Detectarea schimbărilor bazată pe regiune. 122
5.3.2 Detectarea schimbărilor clădirilor. 124
5.4 Concluzii. 132

6 Concluzii **133**
6.1 Contribuţii personale 133
6.2 Diseminare 138

7 Anexe **141**
A.1 Figuri suplimentare 141
A.2 Tabele suplimentare 148

Bibliografie **157**

Introducere

Context, Motivaţie, şi Obiective de Cercetare

Imaginile de teledetecţie obţinute cu sateliţi şi cu vehicule aeriene fără pilot (UAV) permit descrierea, la scară largă, a acoperirii terenului, a infrastructurii, a vegetaţiei şi a efectelor produse de dezastre. Odată cu creşterea rezoluţiei spaţiale şi a frecvenţei de revizitare, volumele de date se măresc rapid, iar condiţiile de achiziţie devin tot mai variate. În aceste condiţii, interpretarea manuală nu mai este fezabilă, ceea ce impune metode automate de percepţie vizuală şi de înţelegere a scenei, capabile să furnizeze rezultate consecvente, rapide şi corecte din punct de vedere spaţial.
Învăţarea profundă reprezintă în prezent baza principală pentru astfel de capabilităţi. Reţelele neuronale convoluţionale (CNN) rămân eficiente pentru sarcini cu predicţie densă, în special segmentarea semantică, datorită capacităţii de a învăţa reprezentări locale robuste şi de a păstra detalii fine. În schimb, mecanismele de atenţie şi arhitecturile de tip transformer îmbunătăţesc integrarea contextului global, modelarea dependenţelor pe distanţe mari şi interacţiunea între caracteristici la scări diferite. În cazul imaginilor aeriene, combinaţiile global–local sunt deosebit de utile deoarece îmbină precizia contururilor cu raţionamentul la nivel de scenă.
Pentru multe aplicaţii, analiza în timp este la fel de importantă ca analiza unei singure imagini. Detectarea schimbărilor urmăreşte identificarea diferenţelor relevante dintre observaţii realizate la momente diferite, fiind esenţială pentru monitorizarea tranziţiilor de utilizare a terenului, expansiunea urbană şi evoluţia infrastructurii. Detectarea fină a schimbărilor cere, în plus, localizarea exactă a modificărilor subtile, la nivel de obiect, ceea ce motivează reprezentări care ţin cont de regiuni şi instanţe pentru alinierea şi compararea imaginilor bi-temporale.
Evaluarea post-dezastru pune aceste cerinţe în prim-plan, sub constrângeri stricte de timp şi fiabilitate. Abordarea Visual Question Answering (VQA) formulează analiza daunelor prin întrebări, astfel încât modelul poate furniza direct din imagini atribute şi cantităţi specifice. Întrebările de tip numărare sunt adesea critice pentru decizie, însă rămân dificile din cauza aglomerării, a ocluziunilor şi a variaţiilor mari de scară.
În segmentare, VQA şi detectarea schimbărilor, o limitare constantă este dependenţa de seturi de date mari, adnotate uniform. Adnotarea densă este costisitoare şi inegală între senzori, zone şi condiţii de achiziţie. De aici rezultă nevoia de generare scalabilă a datelor prin simulare, împreună cu strategii de învăţare care reduc dependenţa de etichetare exhaustivă, precum abordările slab supervizate, semi-supervizate şi auto-supervizate.
În acest cadru, teza abordează cinci direcţii principale: construirea unor seturi de date de dimensiuni mari, cu acoperire semantică coerentă; dezvoltarea învăţării pe secvenţe care valorifică coerenţa temporală în condiţii de supervizare redusă; îmbunătăţirea segmentării

semantice urbane prin echilibrarea preciziei la contururi cu utilizarea contextului; creşterea performanţei VQA multimodale pentru întrebări de numărare prin fuziune adaptată scării şi sensibilă la instanţe; şi avansarea detectării fine a schimbărilor prin raţionament temporal care urmăreşte structura scenei, cu reprezentări aliniate la nivel de instanţă şi cu predicţii care păstrează relaţiile spaţiale relevante.

Pentru această teză definim următoarele obiective de cercetare:

Objectiv 1: Segmentare semantică pentru înţelegerea scenelor aeriene.

Interpretarea semantică densă a scenelor din imagini aeriene, cu accent pe mediile forestiere şi pe constrângerile impuse de variaţia punctului de vedere şi de efortul de adnotare.

Objectiv 1.1: Studierea cerinţelor segmentării semantice în date UAV şi a limitărilor seturilor de date existente pentru înţelegerea pădurilor şi a vegetaţiei.

Objectiv 1.2: Proiectarea şi publicarea unui set de date sintetic pentru segmentare semantică în imagini UAV din medii forestiere, cu variabilitate controlabilă a achiziţiei şi etichete dense la nivel de pixel.

Objectiv 1.3: Dezvoltarea unei reţele de segmentare spaţio-temporală slab supervizat pentru videoclipuri UAV, care reduce costul adnotării prin propagarea etichetelor şi rafinare temporală.

Objectiv 1.4: Dezvoltarea unei arhitecturi de segmentare semantică bazate pe transformeri, care îmbunătăţeşte modelarea contextului global, menţinând în acelaşi timp precizia contururilor locale în imagini aeriene de înaltă rezoluţie.

Objectiv 2: VQA în teledetecţie, cu accent pe evaluarea post-dezastru şi pe numărare.

Adresarea întrebărilor multimodale în teledetecţie, cu accent pe evaluarea post-inundaţie şi pe creşterea acurateţii numărării în condiţii de variaţie de scară şi diversitate ridicată a răspunsurilor.

Objectiv 2.1: Studierea particularităţilor VQA în teledetecţie, cu focalizare pe numărare ca sursă recurentă de eroare în evaluarea post-dezastru.

Objectiv 2.2: Dezvoltarea unei arhitecturi VQA pentru evaluarea post-inundaţie, care integrează codificarea lingvistică şi vizuală cu fuziune multimodală şi modelare decizională structurată.

Objectiv 2.3: Dezvoltarea unei reţele VQA orientat spre numărare, care îmbunătăţeşte reprezentarea vizuală multi-scală, alinierea multimodală şi predicţia numărului în prezenţa distribuţiilor dezechilibrate ale răspunsurilor.

Objectiv 3: Detectarea fină a schimbărilor, cu tokenizare sensibilă la structură şi raţionament temporal.

Localizarea precisă a schimbărilor în scene structurate de teledetecţie (de exemplu, terenuri agricole şi clădiri), cu accent pe reprezentări aliniate la structură şi pe fuziune temporală robustă.

Objectiv 3.1: Studierea limitărilor fluxurilor de lucru existente pentru detectarea schimbărilor, cu accent pe nepotrivirea dintre tokenizarea uniformă pe patch-uri şi domeniile centrate pe structură.

Objectiv 3.2: Dezvoltarea unei arhitecturi de detectare a schimbărilor sensibil la regiuni pentru scene agricole structurate, utilizând tokenizare semantică pe regiuni şi raţionament temporal la nivel de regiune.

Objectiv 3.3: Dezvoltarea unei metode de detectare a schimbărilor axat pe clădiri, care impune consistenţa reprezentărilor la nivel de instanţă şi îmbunătăţeşte predicţia schimbărilor cu păstrarea contururilor.

Structura Tezei

Capitolul 2. State of the Art prezintă sinteza lucrărilor de referinţă corespunzătoare celor trei piloni tehnici ai tezei. Capitolul rezumă seturi de date reprezentative, protocoale de evaluare şi familii de modele pentru segmentarea semantică în imagini aeriene/UAV, VQA în teledetecţie (cu accent pe dificultăţile asociate numărării şi pe spaţiile de răspuns mixte) şi detectarea schimbărilor bazată pe învăţare profundă, evidenţiind limitări recurente care motivează alegerile metodologice din capitolele următoare.
Capitolul 3. Semantic Segmentation for UAV-based Remote Sensing abordează segmentarea semantică în teledetecţia bazată pe UAV prin combinarea contribuţiilor centrate pe date cu cele centrate pe model. Mai întâi, capitolul introduce un set de date sintetic UAV pentru inspecţia pădurilor, cu variabilitate controlabilă a achiziţiei şi etichete dense la nivel de pixel, conceput pentru a studia capacitatea de generalizare la schimbări de punct de vedere şi de mediu. Apoi investighează învăţarea spaţio-temporală slab supervizată pentru segmentarea videoclipurilor UAV, prin propagarea etichetelor şi rafinare temporală, vizând scenarii în care adnotarea densă nu este practică. În final, prezintă o arhitectură de segmentare bazată pe transformeri (SwinFAN), care extinde decodorul prin fuziune global–local a contextului şi rafinare a caracteristicilor, pentru îmbunătăţirea calităţii contururilor şi recuperarea detaliilor fine. Validarea se realizează prin experimente de tip benchmark, ablaţii controlate şi analize calitative care corelează comportamentul observat cu proprietăţile datelor şi cu alegerile de proiectare ale arhitecturii.
Capitolul 4. Visual Question Answering in Remote Sensing studiază VQA în teledetecţie pentru evaluarea daunelor post-dezastru, cu un accent explicit pe îmbunătăţirea acurateţii numărării în condiţiile unor tipuri eterogene de întrebări şi ale unor spaţii de răspuns mixte. Iniţial, capitolul prezintă DiViNeCaps, un flux de procesare care combină un encoder lingvistic pre-antrenat cu embedding-uri vizuale Vision Mamba şi un modul de fuziune ghidat de atenţie, utilizând un cap de tip Capsule Network pentru a susţine, într-o formulare unificată, atât răspunsuri categoriale, cât şi estimarea numerică a numărului. Ulterior, introduce DeVANet, care consolidează raţionamentul orientat spre numărare prin modelare vizuală local–globală (LGA-ViM), fuziune multimodală bidirecţională (self-attention şi cross-attention bidirecţional) şi head-uri de predicţie specifice sarcinii pentru clasificare şi estimarea numărului, antrenate cu obiective care ţin cont de dezechilibrul de clasă şi de distribuţiile supra-dispersate ale numărării. Capitolul evaluează ambele sisteme pe FloodNet şi RescueNet, compară cu baseline-uri VQA reprezentative şi raportează rezultate cantitative, analize ale erorilor în regimuri cu număr mic şi mare, precum şi studii de ablaţie care izolează impactul encoderelor, al fuziunii, al rezoluţiei şi al funcţiilor de pierdere.
Capitolul 5. Change Detection in Remote Sensing se concentrează pe detectarea schimbărilor în teledetecţie şi dezvoltă mecanisme de raţionament sensibile la regiuni şi instanţe pentru localizarea fină a schimbărilor. Mai întâi, capitolul prezintă SFCNet pentru scene agricole structurate, unde măştile semantice pe regiuni ghidează agregarea caracteristicilor, raţionamentul temporal pe tokeni la nivel de regiune şi decodarea spaţială, pentru a îmbunătăţi consistenţa contururilor şi a reduce detecţiile false. Apoi, introduce RADNet pentru detectarea schimbărilor la nivel de clădiri în imagini de înaltă rezoluţie, combinând tokenizare consistentă la nivel de instanţă, un encoder cu atenţie global–locală pentru rafinarea reprezentărilor şi o strategie de decodare multi-temporală şi multi-scală, împreună cu o predicţie bazată pe corelaţii, pentru a păstra detaliul spaţial menţinând identitatea temporală. Metodele propuse sunt validate pe benchmark-uri pentru monitorizarea culturilor şi detectarea schimbărilor la clădiri, prin metrici standard, comparaţii calitative şi analize care evidenţiază alinierea semantică, precizia contururilor şi

robusteţea la variabilitatea de aspect.
Capitolul 6. Conclusions sintetizează contribuţiile tezei în raport cu obiectivele formulate şi consolidează principalele rezultate empirice din cele trei direcţii de cercetare. Capitolul discută modul în care seturile de date, arhitecturile de învăţare şi protocoalele de evaluare propuse susţin împreună obiectivul tezei de a obţine percepţie fiabilă pentru aplicaţii aeriene şi de teledetecţie şi încheie cu limitări, direcţii viitoare de cercetare şi un rezumat al diseminării prin publicaţiile rezultate.
Anexele oferă materiale suplimentare care sprijină reproductibilitatea şi analiza detaliată a rezultatelor, incluzând detalii experimentale adiţionale, studii de ablaţie extinse şi tabele cantitative complementare pentru modelele analizate în capitolele principale.

Segmentare semantică pentru teledetecţie bazată pe UAV

Set de date sintetic forestier pentru segmentare semantică

Introducem Forest Inspection, un set de date sintetic pentru segmentare semantică în medii forestiere, construit pentru imagini capturate cu vehicule aeriene fără pilot (UAV). Setul de date este generat în Unreal Engine 4, folosind simulare de zbor bazată pe fizică prin AirSim, iar măştile semantice sunt obţinute automat printr-o codare stencil specifică fiecărei clase şi o paletă fixă de etichetare. Setul oferă perechi de imagini RGB şi hărţi semantice dense la nivel de pixel, la rezoluţie înaltă, împreună cu metadate privind poziţia şi orientarea UAV-ului. Achiziţia urmează un protocol controlat, de tip traiectorie „sweep", care eşantionează trei altitudini (30, 50 şi 80 m), trei unghiuri de înclinare (pitch) (0°, 60° şi 90°) şi două condiţii meteo (senin şi înnorat), păstrând un regim de iluminare consecvent. Taxonomia include 11 clase, separă explicit arborii foioşi de cei coniferi şi introduce o clasă relevantă pentru inspecţie, arbore căzut. **Figura 1** prezintă exemple de perechi RGB–etichetă, fiecare rând conţinând o imagine a camerei urmată de harta semantică asociată.

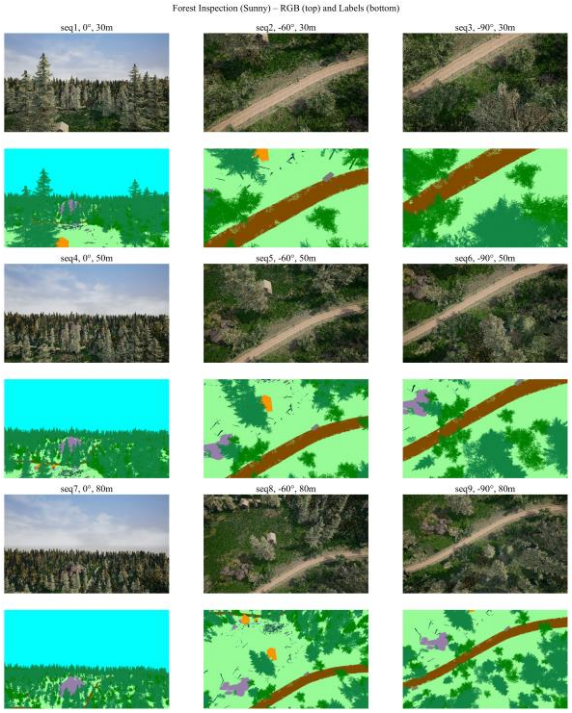

Figura 1. Exemple de imagini color (rândul de sus) şi etichetele corespunzătoare de segmentare semantică (rândul de jos) din setul de date sintetic Forest Inspection, prezentate pentru fiecare secvenţă de zbor în condiţii de vreme senină.

Tabel 1 poziţionează setul de date Forest Inspection în raport cu benchmark-uri UAV reprezentative, comparând tipul setului de date, dimensiunea, geometria de achiziţie, rezoluţia, proiectarea semantică şi relevanţa operaţională pentru inspecţia forestieră. UAVid [1], Ruralscapes [2] şi WildUAV [3] sunt colecţii din lumea reală, achiziţionate la rezoluţii foarte mari şi, în general, în condiţii de vreme senină, în timp ce Mid-Air [4] şi SPREAD [5] sunt seturi de date sintetice, orientate spre segmentare semantică şi segmentare la nivel de instanţă. Forest Inspection este un set de date sintetic, organizat pe secvenţe, conceput pentru traiectorii de inspecţie controlate, ceea ce permite obţinerea unei referinţe (ground truth) precise la nivel de pixel şi setări de achiziţie reproductibile. Setul conţine 26.319 imagini înregistrate la 3 altitudini şi 3 unghiuri de înclinare (pitch), la o rezoluţie înaltă uniformă, oferind un nivel de detaliu mai ridicat decât bazele sintetice uzuale, păstrând în acelaşi timp fezabilitatea pentru experimente la scară mare. Spre deosebire de seturile anterioare pentru segmentare UAV, care tratează arborii ca o singură clasă, Forest Inspection introduce separarea explicită între arbori foioşi şi coniferi, precum şi o clasă arbore căzut, pentru a sprijini analiza pericolelor şi înţelegerea scenelor afectate. În plus, setul urmăreşte o variabilitate de mediu realistă, dar controlată, prin includerea condiţiilor senin şi înnorat, în acord cu scenarii operaţionale tipice. Organizarea pe secvenţe conduce, totodată, la un grad de suprapunere între cadre mai redus decât în alte benchmark-uri, precum Mid-Air şi Ruralscapes, ceea ce indică o redundanţă mai mică, menţinând însă continuitatea temporală.

Tabel 1. Comparație între seturi de date reale și sintetice, achiziționate cu UAV, pentru analiza mediilor forestiere și a scenelor aeriene.

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**

Segmentare semantică în videoclipuri UAV cu supervizare redusă

Abordăm segmentarea semantică în secvenţe video UAV în condiţii de supervizare limitată la nivel de pixel, formulând un cadru slab supervizat în care etichetele dense sunt disponibile doar pentru fiecare al k-lea cadru, iar predicţiile sunt rafinate în timp. Fluxul de procesare este ilustrat în **Figura 2**. Metoda combină o reţea de segmentare semantică aplicată pe fiecare cadru cu estimarea fluxului optic, pentru a propaga informaţia între cadre consecutive, şi foloseşte deformare (warping) spaţio-temporală, împreună cu rafinare recurentă, pentru a creşte consistenţa temporală. Unitatea recurentă propusă, de tip GRU spaţio-temporal cu porţi (gated), integrează propagarea bazată pe warping cu un modul recurent modificat, în care porţile şi actualizările de stare sunt implementate prin convoluţii separabile pe adâncime (depthwise separable convolutions). Astfel, se obţin ieşiri de segmentare rafinate la fiecare pas temporal. Fluxul optic este utilizat atât ca sursă de supervizare în configuraţia slab supervizată, cât şi ca o componentă ce poate fi antrenată împreună cu reţeaua de segmentare într-un cadru end-to-end.

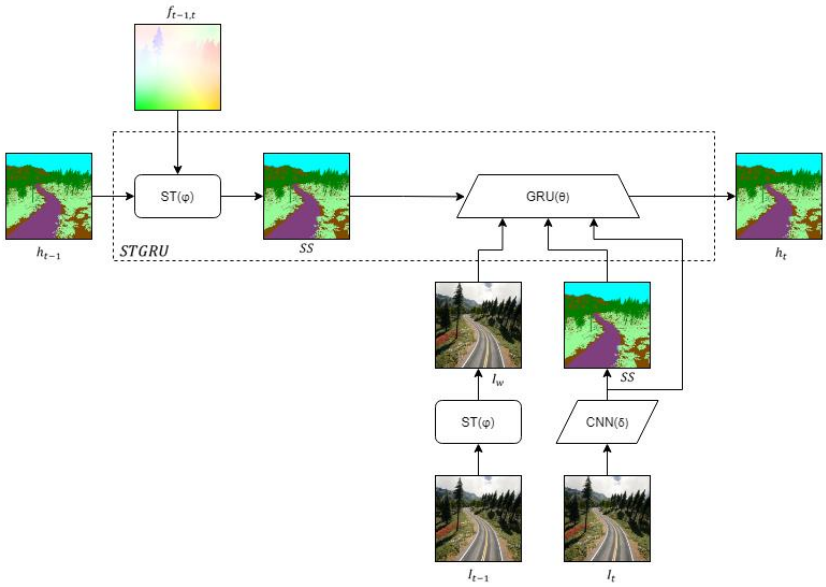

Figura 2. Fluxul de procesare al metodologiei propuse, în care imaginile color de intrare, împreună cu segmentarea semantică și fluxul optic, sunt introduse în componenta STGRU pentru propagarea și rafinarea etichetării.

Evaluăm cadrul propus folosind setul de date Mid-Air şi raportăm, în **Tabel 2**, rezultate de antrenare end-to-end pentru două reţele de estimare a fluxului optic. Optimizarea comună a componentelor de segmentare semantică şi estimare a mişcării creşte acurateţea segmentării, susţinând interpretarea că o estimare mai bună a fluxului optic facilitează un raţionament spaţio-temporal mai coerent. Câştigurile sunt mai vizibile pentru clasele care beneficiază de indicii temporale, precum vegetaţia de la sol, terenul stâncos, bolovanii şi indicatoarele rutiere. Per ansamblu, configuraţia end-to-end obţine o îmbunătăţire de aproximativ 4% faţă de baseline-ul de segmentare statică, cu un cost computaţional corespunzător mai ridicat.

Predicţiile calitative pe setul de test sunt prezentate în **Figura 3**. Exemplele includ atât scene forestiere, cât şi scene rutiere şi acoperă texturi variate şi condiţii diferite de iluminare. Cele mai dificile categorii sunt bolovanii şi indicatoarele rutiere, care apar rar în datele de antrenare şi, prin urmare, au o segmentare mai puţin stabilă, în special la contururile obiectelor şi în zone aglomerate.

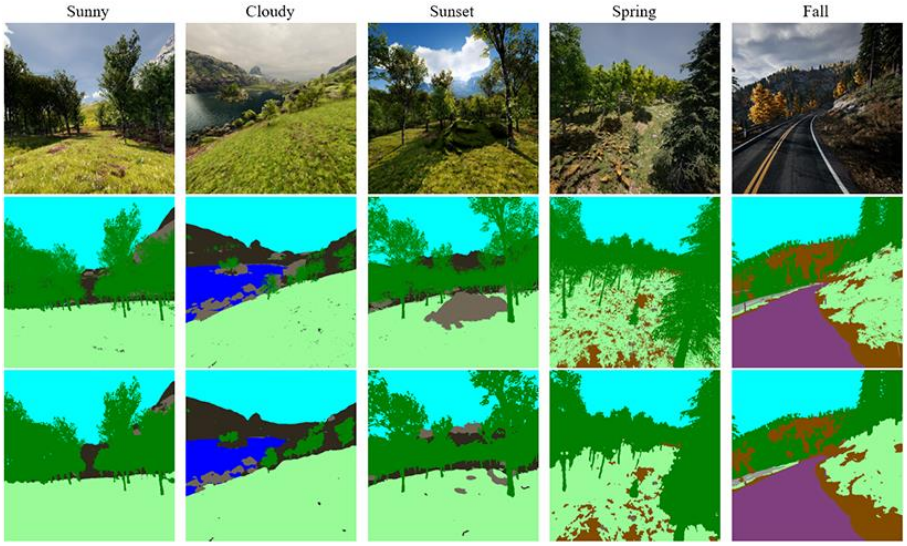

Figura 3. Rezultatele metodei slab supervizate, prezentate pentru 5 scenarii diferite din setul de date Mid-Air.

Tabel 2. Rezultatele IoU pe setul de test Mid-Air, unde GRU(f, k) reprezintă sistemul complet, antrenat folosind etichete pentru fiecare al k-lea cadru, iar f indică fluxul optic utilizat: fie fluxul de referință (GT) obținut cu PWC-Net, fie fluxul estimat de VCN reantrenat.

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**

Segmentare semantică bazată pe transformeri

Propunem o arhitectură de tip U-Net bazată pe transformeri pentru segmentarea semantică a imaginilor de teledetecţie şi introducem SwinFAN (Swin-based Focal Axial Attention Network), ilustrată în **Figura 4**. Aceasta este o arhitectură U-Net pe patru niveluri, care combină un encoder ierarhic Swin Transformer cu un decodor construit pe atenţie Guided Focal–Axial (GFA) şi un modul Attention-based Feature Refinement Head (AFRH) pentru rafinarea ieşirii.

Encoderul reprezintă imaginea ca o secvenţă de patch-uri ne-supravapuse şi utilizează auto-atenţie în ferestre (window-based self-attention), cu ferestre deplasate între etape, pentru a obţine caracteristici multi-scală; structura blocului Swin este detaliată în **Figura 5(a)**. Decodorul reconstruieşte predicţiile gradual, de la grosier la fin, şi foloseşte atenţia GFA, prezentată în **Figura 5(b)**, pentru a combina atenţia axială cu un modul focal care generează o mască spaţială multi-scală ce ghidează răspunsul atenţiei. În acest fel, se îmbină modelarea contextuală pe distanţe lungi cu o accentuare selectivă în spaţiu.

Caracteristicile encoderului şi ale decodorului sunt combinate printr-un mecanism de fuziune ponderată, iar AFRH (**Figura 6**) rafinează suplimentar reprezentarea fuzionată prin alinierea caracteristicilor reziduale, reponderare prin atenţie spaţială şi utilizarea convoluţiilor separabile pe adâncime (depthwise separable) cu conexiuni reziduale, pentru a păstra structurile fine. Antrenarea foloseşte o funcţie de pierdere hibridă, care combină cross-entropy şi Dice loss, optimizând simultan clasificarea la nivel de pixel şi suprapunerea pe regiuni în condiţii de dezechilibru între clase.

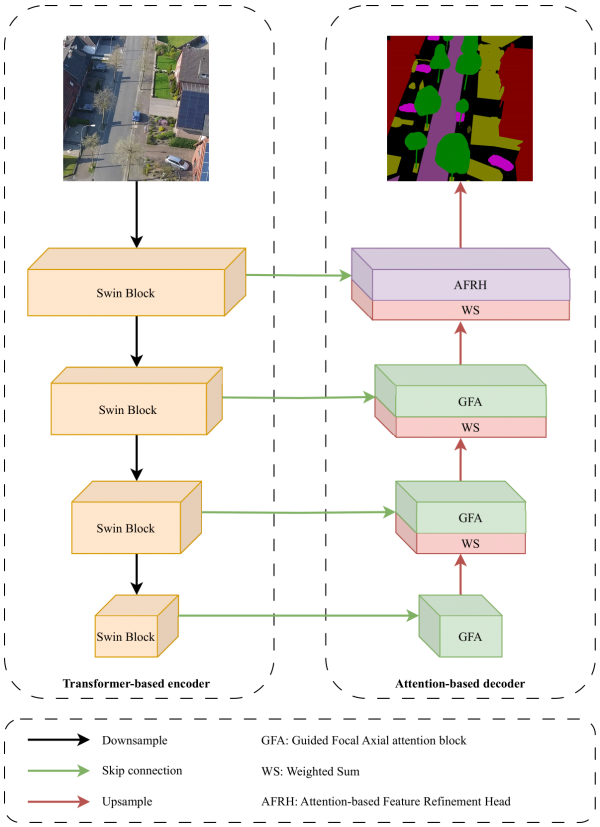

Figura 4. Arhitectura SwinFAN. Encoderul extrage caracteristici la mai multe scări folosind blocuri Swin bazate pe transformeri, iar decodorul utilizează atenția Guided Focal–Axial pentru integrarea informației globale și locale. În etapa finală, modulul Attention-based Feature Refinement Head rafinează ieșirea pentru a produce o hartă de segmentare precisă.

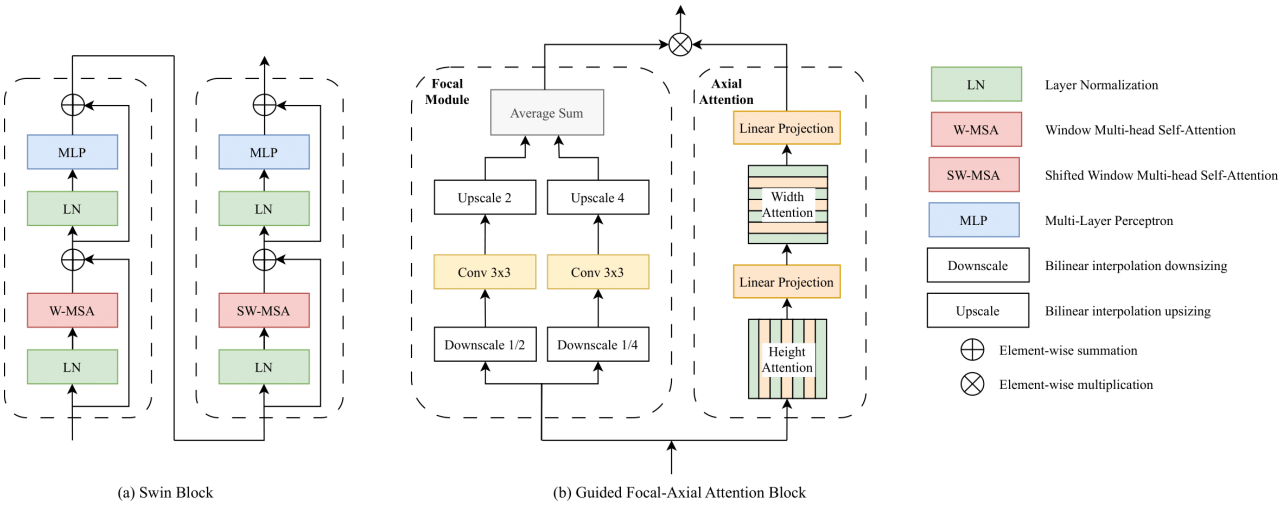

Figura 5. Blocurile principale ale arhitecturii SwinFAN. (a) Blocul Swin: include o succesiune de normalizare pe strat (layer normalization), mecanisme de self-attention multi-head (în ferestre și cu ferestre deplasate) și perceptroni multi-strat (MLP). (b) Blocul Guided Focal–Axial Attention: realizează o înmulțire element-cu-element între un modul focal, care rafinează caracteristicile la două scări pentru a evidenția detaliile locale, și atenția axială, care operează secvențial pe axele verticală și orizontală ale hărții de caracteristici pentru a surprinde dependențe globale.

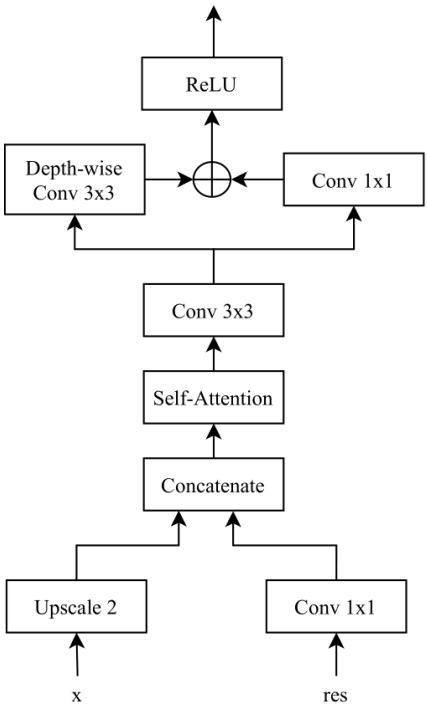

Figura 6. Arhitectura modulului Attention-based Feature Refinement Head.

Evaluăm SwinFAN pe patru benchmark-uri publice (UAVid, Potsdam [6], Vaihingen [7] şi LoveDA [8]) folosind un protocol de antrenare unificat. Metoda este comparată cu un set extins de baseline-uri bazate pe CNN, mecanisme de atenţie şi modele bazate pe transformeri, incluzând şi variante de tip global–local.

Pe UAVid, **Tabel 3** arată că SwinFAN obţine cel mai bun mIoU, cu îmbunătăţiri notabile pentru clasele building, road şi human, iar comparaţia calitativă din **Figura 7** evidenţiază contururi mai bine definite şi o detecţie mai bună a obiectelor mici. Pe Potsdam şi Vaihingen, modelul se clasează pe primul loc la metricile principale (F1, mIoU şi OA), ceea ce indică o generalizare consecventă în scene urbane de înaltă rezoluţie, depăşind atât competitori bazaţi pe ViT, cât şi pe Swin. Pe LoveDA, SwinFAN atinge cel mai mare mIoU, menţinând în acelaşi timp 196,1 FPS, şi obţine performanţe ridicate pentru clasele building, road, water şi agriculture; principala limitare apare la clasa Barren, din cauza texturii reduse şi a dezechilibrului între clase.

Studiile de ablaţie pe UAVid confirmă contribuţia decodorului şi a mecanismelor de rafinare: atenţia Guided Focal–Axial oferă cea mai bună strategie pentru decodor, iar modulul propus Attention-based Feature Refinement Head adaugă încă 1,1% mIoU, cu cele mai mari câştiguri la obiectele mici. O comparaţie suplimentară cu backbone-ul Swin-Base pe Potsdam arată că SwinFAN obţine o acurateţe superioară după doar 45 de epoci, reducând costul de antrenare în raport cu baseline-urile Swin antrenate cu programe mai lungi.

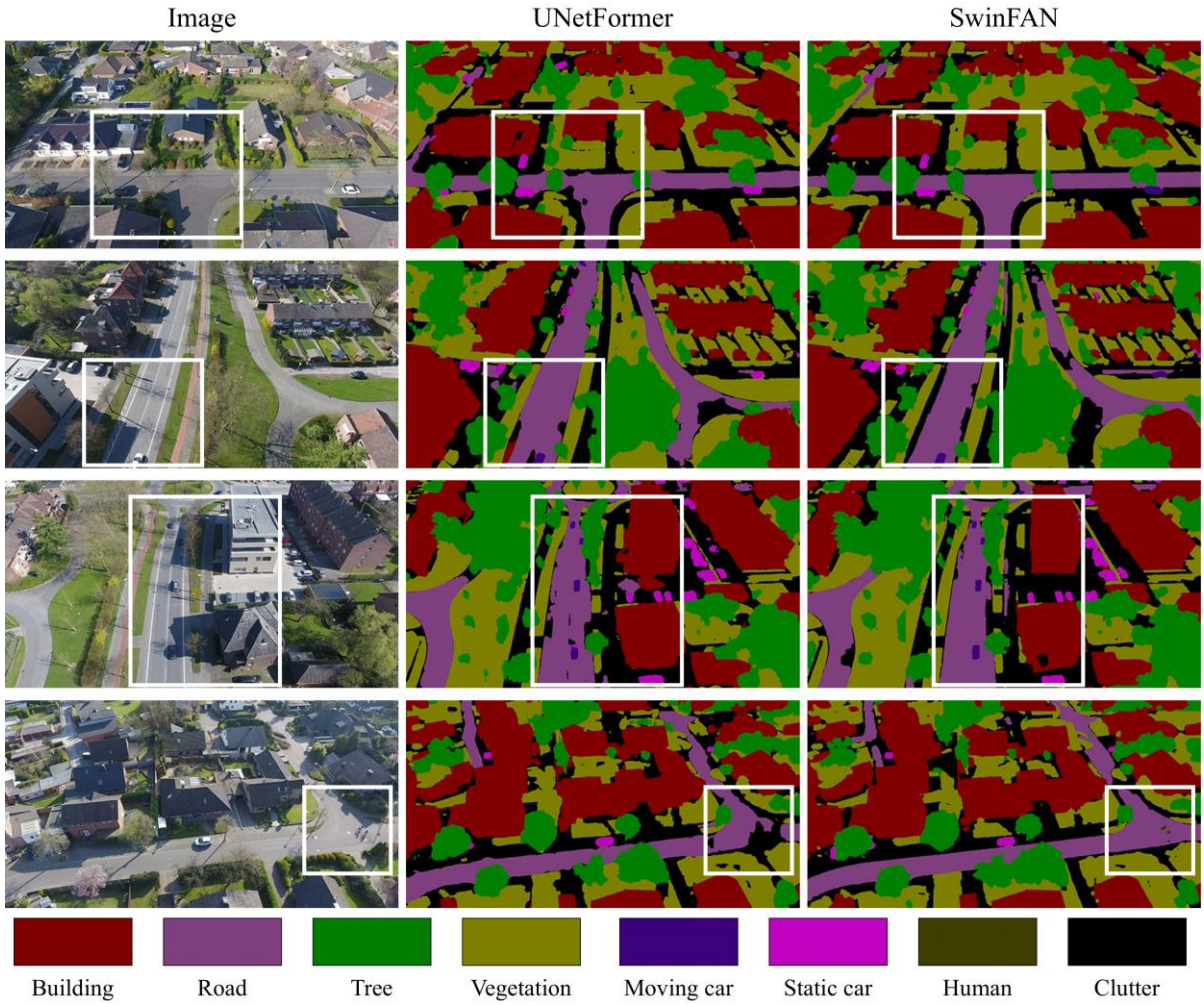

Figura 7. Analiză calitativă comparativă a performanței de segmentare semantică pentru UNetFormer și SwinFAN pe setul de test UAVid (Germania), pentru secvențele 29, 30 și 38. Cu chenarul alb sunt evidențiate îmbunătățirile în detecția claselor stradă, mașină și persoană.

Tabel 3. Evaluarea performanței pe setul de test UAVid, cu evidențierea celor mai bune rezultate cu caractere aldine și a celor de pe locul al doilea prin subliniere.

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**

Răspunsuri la Întrebări Vizuale în Teledetecţie

Răspunsuri la Întrebări Vizuale post-inundaţie în teledetecţie

Introducem reţeaua DiViNeCaps (DistilBERT Vision Network with CapsuleNet classifier) pentru Răspunsuri la Întrebări Vizuale (VQA) în evaluarea daunelor produse de inundaţii, cu accent pe întrebările de tip numărare şi pe maparea fiecărei perechi întrebare–imagine către un spaţiu de răspuns mixt, care include atât răspunsuri categoriale, cât şi valori numerice (numărări). Fluxul propus (prezentat în **Figura 8**) utilizează DistilBERT pentru codificarea întrebării şi Vision Mamba pentru calculul embedding-urilor imaginii, astfel încât să fie păstrată informaţia vizuală relevantă pentru numărare fină. Cele două modalităţi sunt apoi integrate printr-un mecanism de fuziune bazat pe atenţie, care aplică self-attention peste caracteristicile textuale, atenţie cu poartă (gated attention) peste caracteristicile vizuale şi cross-attention bidirecţional pentru alinierea semanticii întrebării cu conţinutul imaginii. Predicţia este realizată printr-un cap de tip Capsule Network, care operează pe reprezentarea fuzionată pentru a susţine, într-o formulare unificată, atât predicţia structurată a categoriilor, cât şi regresia numărării. Antrenarea este ghidată de un obiectiv comun, care combină cross-entropy pentru răspunsurile categoriale cu RMSE pentru ieşirile numerice, în acord cu vocabularul de răspunsuri din FloodNet [9].

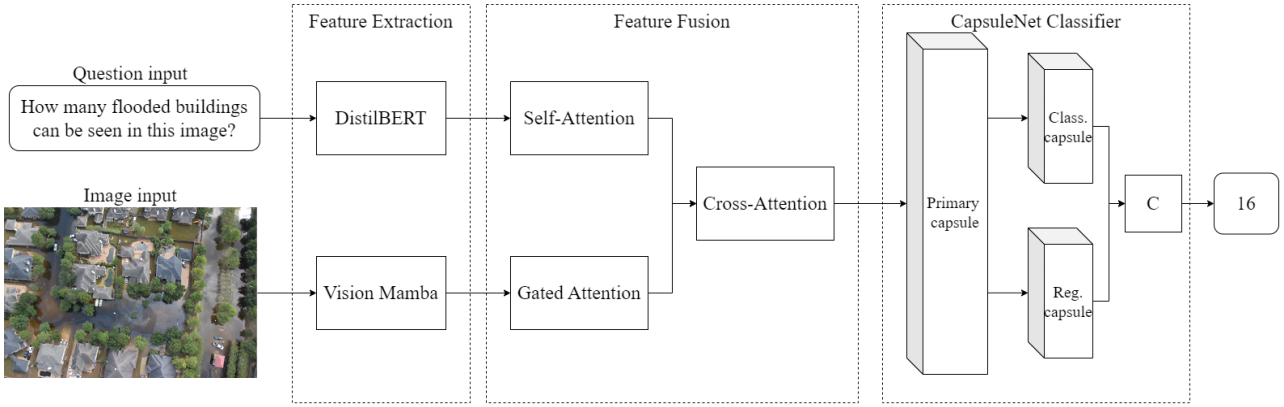

Figura 8. Prezentare generală a arhitecturii propuse DiViNeCaps (DistilBERT Vision Network cu clasificator CapsuleNet) pentru Răspunsuri la Întrebări Vizuale pe setul de date FloodNet.

DiViNeCaps este evaluat pe benchmark-ul FloodNet pentru Răspunsuri la Întrebări Vizuale în teledetecţie, în scenarii post-inundaţie. Metoda este comparată cu un set variat de baseline-uri VQA, de la arhitecturi timpurii de tip LSTM + VGG16 şi modele bazate pe atenţie (SAN [10], MFB cu Co-Attention [11]), până la abordări recente cu modele pre-antrenate pentru limbaj şi viziune, care folosesc encodere de tip RoBERTa sau DistilBERT şi backbone-uri vizuale ConvNeXt, ResNet-152 sau CLIP, cu fuziune prin concatenare, înmulţire element-cu-element sau atenţie de tip MFB. În raport cu aceste baseline-uri, DiViNeCaps combină codificarea întrebării cu DistilBERT şi caracteristici vizuale extrase cu Vision Mamba, o schemă de fuziune bazată pe atenţie (self-attention, atenţie cu poartă şi cross-attention) şi un cap de predicţie de tip CapsuleNet, antrenat cu un obiectiv comun care îmbină cross-entropy pentru răspunsurile categoriale cu RMSE pentru regresia numărării, astfel încât să reflecte spaţiul mixt de răspunsuri din FloodNet. Conform rezultatelor raportate în **Tabel 4**, modelul propus obţine cea mai bună acurateţe globală (98,84%) şi se clasează pe primul loc pentru toate categoriile de întrebări raportate, incluzând întrebările de numărare (40,02% Simple Count şi 40,78% Complex Count) şi întrebările de tip clasificare (98,92% Yes/No, 98,73% Image Condition şi

98,55% Road Condition). Rezultatele calitative din **Figura 9** susţin tendinţele cantitative şi indică faptul că principalele cazuri de eşec apar la întrebări care implică numărări foarte mari de clădiri, unde predicţiile devin mai puţin stabile. Comparativ cu baseline-uri recente puternice, precum CNX-mul şi Grad-Cam, DiViNeCaps îmbunătăţeşte numărarea prin componenta sa dedicată de regresie, în timp ce cel mai apropiat competitor ca performanţă globală, CLIP-mul-taskwise, nu raportează rezultate pentru sarcina de numărare.

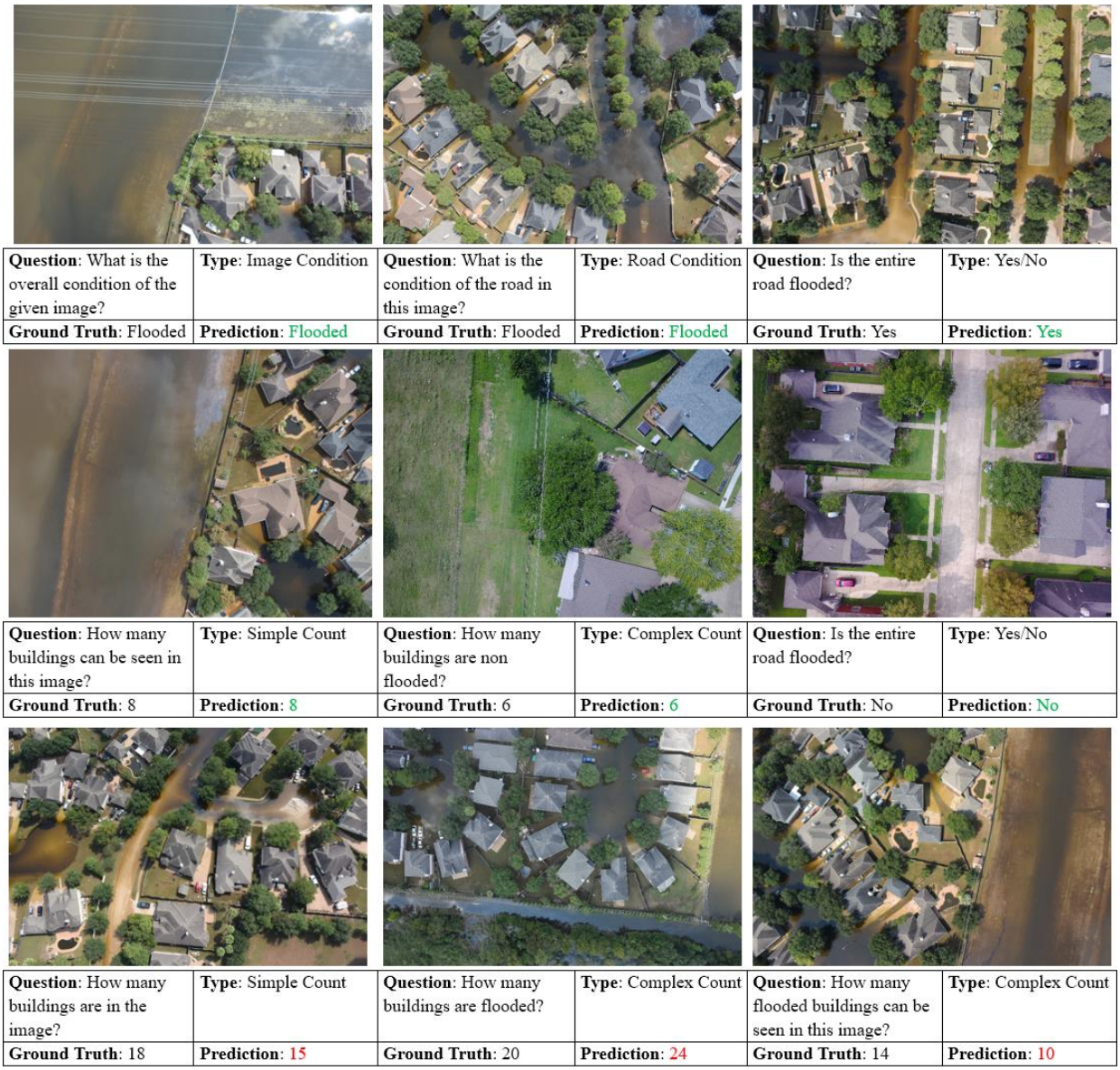

Figura 9. Rezultatele rețelei propuse pe setul de date FloodNet, primele două rânduri ilustrând predicții corecte pentru diverse tipuri de întrebări (Image Condition, Road Condition, Yes/No, Simple Count, Complex Count). Rândul de jos evidențiază situații în care modelul a întâmpinat dificultăți, în special la întrebările de tip Simple Count și Complex Count.

Tabel 4. Comparație de performanță între diverse rețele pentru Răspunsuri la Întrebări Vizuale pe setul de date FloodNet, pe mai multe tipuri de întrebări, ordonată după acuratețea globală. Valorile scrise cu caractere aldine și cele subliniate indică, pentru fiecare categorie, cel mai bun rezultat, respectiv al doilea rezultat.

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**

Numărare robustă pentru Răspunsuri la Întrebări Vizuale post-dezastru

Propunem arhitectura DeVANet (**Figura 10**) pentru Răspunsuri la Întrebări Vizuale (VQA) post-dezastru în imagini de teledetecție, cu un accent principal pe îmbunătățirea acurateții numărării obiectelor printr-o fuziune multimodală eficientă. DeVANet utilizează DeBERTa pentru codificarea întrebării și un backbone vizual Vision Mamba, extins cu un modul de Atenție Local–Globală (LGA), pentru a păstra atât evidența la scară fină, cât și indicii de context mai larg necesare numărării în scene aglomerate, așa cum este prezentat în **Figura 11**. Reprezentarea imaginii este construită prin ghidarea caracteristicilor Vision Mamba cu harta de caracteristici local–globală, astfel încât să fie puse în evidență regiunile constant informative. Caracteristicile textuale și vizuale sunt apoi integrate printr-un modul de fuziune bidirecțională, care combină self-attention specific fiecărei modalități cu cross-attention bidirecțional și co-attention, pentru a susține raționamentul comun. Pentru predicție, DeVANet folosește două capete MLP specifice sarcinii: unul pentru răspunsuri categoriale și unul pentru estimarea numărării, selectate în funcție de tipul întrebării. Modelul este antrenat cu un obiectiv comun care combină cross-entropy ponderată pentru clasificarea dezechilibrată cu o pierdere de tip Negative Binomial pentru a modela datele de numărare supra-dispersate.

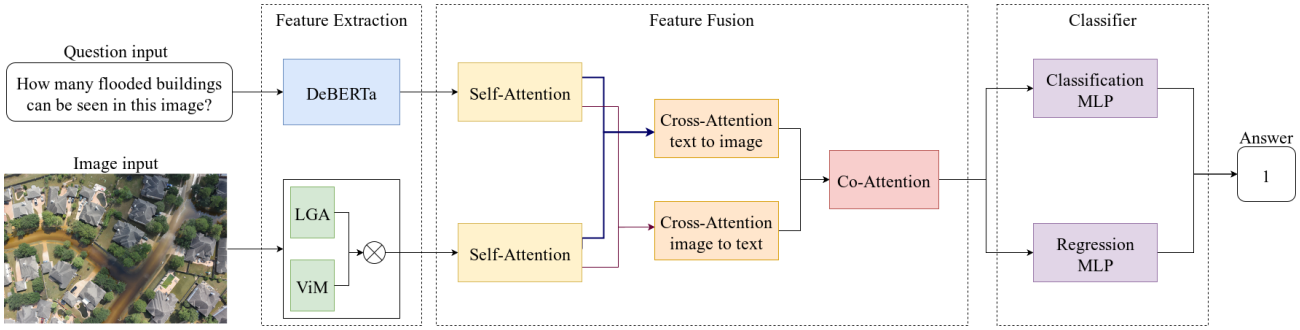

Figura 10. Prezentare generală a arhitecturii DeVANet. Sistemul primește două intrări, o întrebare și o imagine, care sunt procesate de DeBERTa și LGA-ViM pentru extragerea caracteristicilor textuale și vizuale. Aceste caracteristici sunt integrate printr-un modul de fuziune bidirecțională, apoi sunt transmise către un predictor dual de tip MLP pentru a produce răspunsul final.

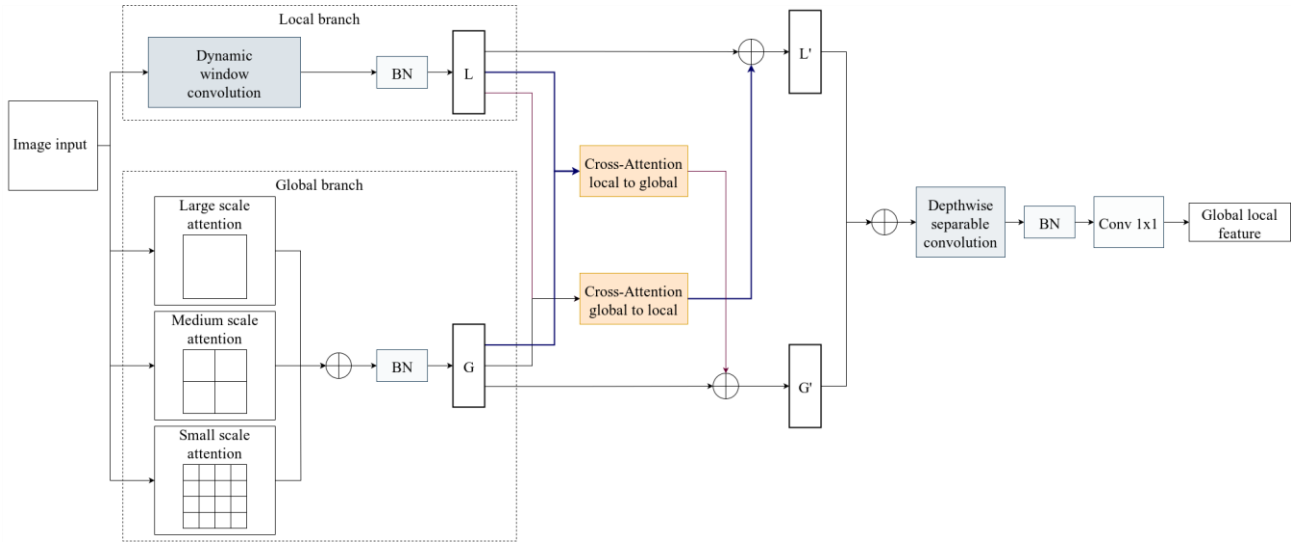

Figura 11. Arhitectura LGA, evidențiind componentele ramurii locale și ale ramurii globale.

Evaluăm DeVANet pe seturile de date post-dezastru FloodNet şi RescueNet [12]. Metoda este comparată cu baseline-uri multimodale reprezentative din VQA pe imagini naturale şi din VQA în teledetecţie, incluzând MCAN [13], VisualBERT [14], ViLBERT [15], UNITER [16], RSVQA [17], SHRNet [18], SAM-VQA [19] şi RSAdapter [20]. Acestea acoperă modele co-attention non-transformer, arhitecturi cu transformeri cu un singur flux sau cu două fluxuri, precum şi adaptări specifice teledetecţiei.

Pe FloodNet, DeVANet obţine cea mai mare acurateţe globală (98,93%) şi îmbunătăţeşte substanţial performanţa la numărare (59,72% Simple Count, 54,48% Complex Count) faţă de cele mai puternice baseline-uri, RSAdapter şi UNITER, conform **Tabel 5**. Analiza erorilor de numărare confirmă, în plus, o robusteţe mai bună atât în regimurile cu număr mic, cât şi cu număr mare, unde DeVANet obţine cele mai mici valori RMSE/MAE şi erori logaritmice reduse (LRMSE/LMAE), ceea ce indică o acurateţe relativă mai bună pe măsură ce creşte numărul de obiecte (**Tabel 6**).

Pe RescueNet, care include scene mai dense şi o diversitate mai mare de sarcini, DeVANet se clasează din nou pe primul loc la nivel global (88,47%), depăşind RSAdapter cu 2,24% şi UNITER cu 3,74%. Cele mai mari câştiguri apar la numărare (60,11% Simple Count, 55,32% Complex Count) şi se observă îmbunătăţiri consecvente şi pentru sarcinile legate de daune şi pentru sarcinile de raţionament (**Tabel 7**). Metricile de eroare pentru numărare pe RescueNet indică acelaşi trend de stabilitate la valori mari ale numărării, DeVANet obţinând cele mai mici erori pentru RMSE/MAE şi pentru variantele lor logaritmice (**Tabel 8**).

O comparaţie de eficienţă pe RescueNet arată că DeVANet atinge cea mai bună acurateţe la un cost computaţional competitiv, comparabil cu baseline-urile bazate pe transformeri. Studiile extinse de ablaţie pe RescueNet evidenţiază principalii factori care contribuie la performanţă: intrări la rezoluţia 512×512 oferă cel mai bun compromis acurateţe–cost; DeBERTa este cel mai performant encoder lingvistic; backbone-ul vizual propus LGA–ViM produce cele mai mari îmbunătăţiri pentru sarcinile dificile; utilizarea a trei scări LGA echilibrează acurateţea şi eficienţa; Bidirectional Attention Fusion este cea mai eficientă strategie de fuziune; iar un predictor dual-MLP antrenat cu pierdere Negative Binomial asigură cea mai bună performanţă globală, în special pentru numărarea complexă şi pentru alte categorii dezechilibrate.

Tabel 5. Evaluarea acurateții DeVANet pe setul de date FloodNet.

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**

Tabel 6. Metrici ale erorii de numărare pe FloodNet pentru numere mici (<10) și mari (≥10) de obiecte.

	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**

Tabel 7. Evaluarea acurateții DeVANet pe setul de date RescueNet.

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**

Tabel 8. Metrici ale erorii de numărare pe RescueNet pentru numere mici (<10) și mari (≥10) de obiecte.

	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**

Detectarea schimbărilor în teledetecție

Detectarea schimbărilor sensibilă la regiuni

Introducem SFCNet (Semantic Field-aware Change Network), o arhitectură pentru detectarea schimbărilor, concepută pentru peisaje agricole structurate, în care schimbările urmează, de regulă, limite coerente ale parcelelor, și nu apar ca pixeli izolați. Așa cum este ilustrat în **Figura 12**, SFCNet utilizează un flux de procesare sensibil la regiuni: extrage caracteristici convoluționale multi-scală din intrări bi-temporale, le agregă în tokeni de regiune aliniați la parcele printr-un Semantic Region Tokenizer (SRT), ghidat de măști predefinite ale regiunilor

de cultură, și realizează raționament temporal la nivel de regiune cu un Field-Aware Region Encoder (FARE), bazat pe grupare de tip slot a tokenilor, pentru a surprinde tipare temporale comune între parcele. Embedding-urile obținute, îmbogățite cu context, sunt apoi decodate înapoi în reprezentări spațiale dense printr-un Contextual Crop Decoder (CCD), permițând reconstruirea la nivel de pixel cu păstrarea structurii parcelelor. Un cap de predicție bazat pe diferență (scădere) produce harta finală a schimbărilor, iar antrenarea se realizează cu un obiectiv de tip binary cross-entropy.

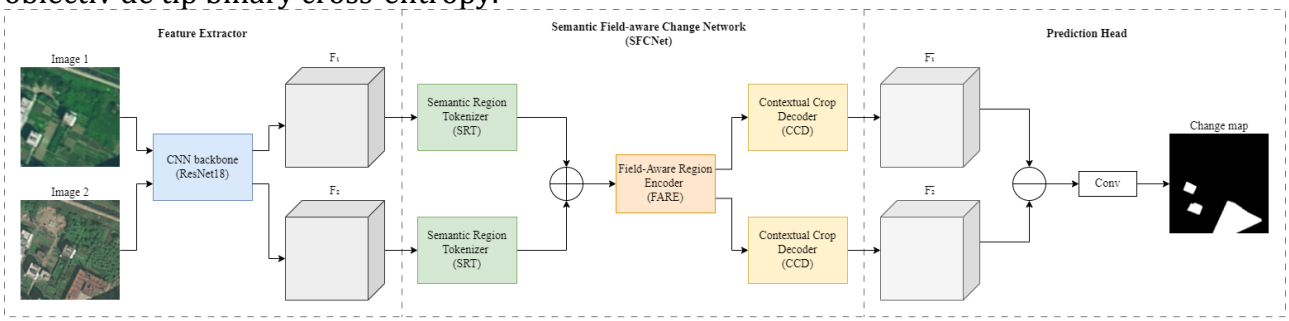

Figura 12. Prezentare generală a arhitecturii propuse SFCNet pentru detectarea schimbărilor în regiuni agricole structurate. Modelul include un backbone CNN partajat pentru extragerea caracteristicilor bi-temporale, un Semantic Region Tokenizer (SRT), un Field-Aware Region Encoder (FARE) și două decodoare Contextual Crop Decoder (CCD).

Evaluăm SFCNet pe benchmark-ul CLCD pentru detectarea fină a schimbărilor în terenuri agricole [21] și îl comparăm cu șapte baseline-uri reprezentative bazate pe CNN, modele hibride și modele bazate pe transformeri (SiamUNet-Diff [22], A2Net [23], BIT [24], ChangeFormer [25], ICIF-Net [26], DMINet [27] și RDP-Net [28]). Conform rezultatelor din **Tabel 9**, SFCNet obține cele mai bune performanțe la toate metricile, atingând 52,47% IoU și 67,23% F1, cu precizie și recall echilibrate (67,85% și 68,37%), depășind cel mai puternic baseline, BIT, cu 3,31% IoU. Comparațiile calitative din **Figura 13** arată că SFCNet produce contururi mai curate și mai puține detecții false, reducând atât supra-segmentarea observată la baseline-urile bazate pe transformeri, cât și ratarea structurilor subțiri, tipică modelelor bazate pe CNN. În final, analiza de eficiență indică faptul că SFCNet obține cel mai mare IoU și cea mai bună acuratețe, menținând în același timp un cost computațional competitiv: depășește arhitecturi mai grele, precum ChangeFormer, și rămâne comparabil ca latență cu baseline-urile mai ușoare.

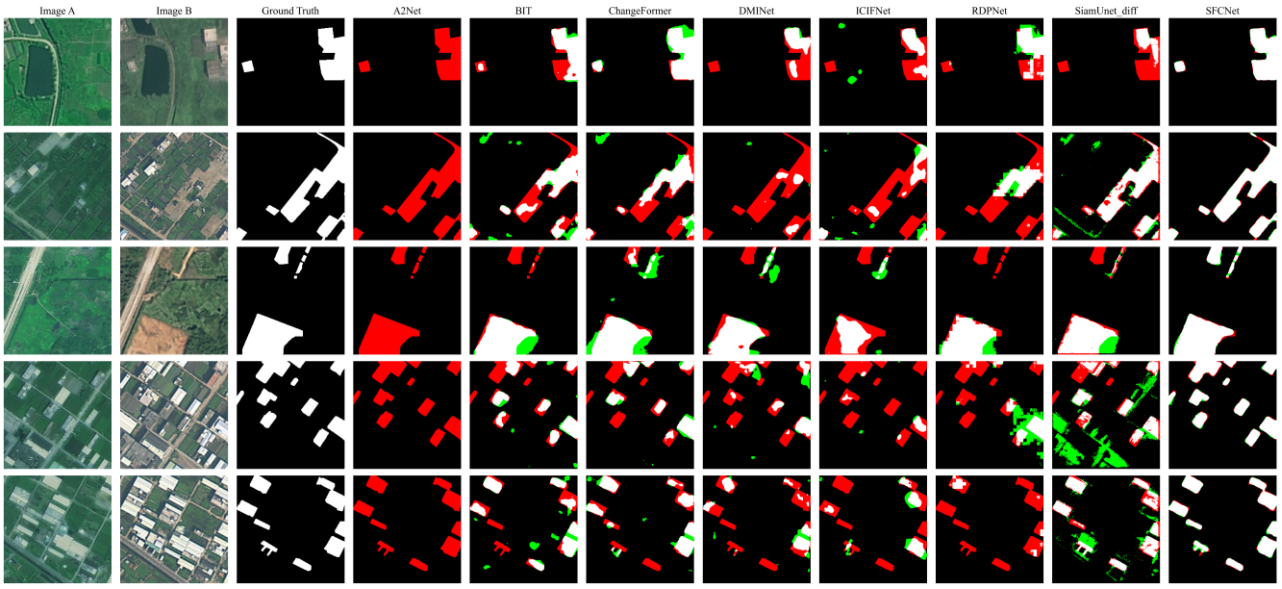

Figura 13. Comparație calitativă a rezultatelor de detectare a schimbărilor pe setul de date CLCD. Alb: adevărat pozitiv; roșu: fals negativ; verde: fals pozitiv; negru: adevărat negativ.

Tabel 9. Comparație cantitativă a modelelor de detectare a schimbărilor pe setul de date CLCD.

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**

Detectarea schimbărilor clădirilor

Propunem RADNet (Region-Aware Dual-path Network), o arhitectură de detectare a schimbărilor bazată pe tokeni, cu mecanisme global–local, destinată imaginilor urbane de înaltă rezoluţie. Metoda se concentrează pe schimbările la nivelul clădirilor şi urmăreşte reducerea erorilor frecvente generate de aglomerarea structurală, variaţiile de punct de vedere şi schimbările de iluminare. Aşa cum este prezentat în **Figura 14**, RADNet extrage caracteristici bi-temporale multi-scală cu un backbone convoluţional partajat şi le transformă în tokeni aliniaţi la clădiri printr-un Instance-Consistent Tokenizer (ICT), care estimează regiuni de instanţă „soft" pe baza indiciilor de atenţie şi a continuităţii spaţiale, permiţând o tokenizare semantic coerentă fără a depinde de măşti externe de instanţă, conform **Figura 15**. Tokenii rezultaţi sunt rafinaţi de un Global-Local Attention Encoder (GLAE), care combină interacţiuni globale între tokeni cu evidenţă geometrică locală, pentru a păstra atât contextul pe distanţe lungi, cât şi detaliul la nivel de contur. Un Cross-Temporal Cross-Scale Decoder (CTCSD) agregă informaţia între nivelurile piramidei de caracteristici, menţinând însă separate embedding-urile specifice fiecărui moment temporal, astfel încât să fie posibilă compararea la nivel de instanţă. În final, un Correlation-Based Prediction Head (CBPH) reconstruieşte reprezentări dense pornind de la embedding-urile de instanţă şi estimează probabilitatea schimbării prin diferenţiere bazată pe similaritate, obţinând hărţi ale schimbărilor coerente spaţial şi reducând influenţa variaţiilor de aspect care nu corespund unor modificări structurale reale.

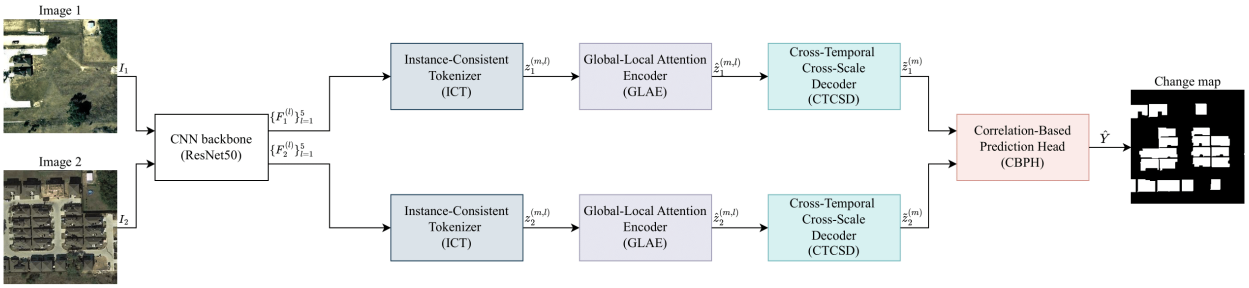

Figura 14. Arhitectura RADNet. Modelul procesează două imagini de intrare printr-un backbone ResNet50 partajat pentru a extrage caracteristici multi-scală, care sunt apoi tokenizate, îmbogățite cu informație contextuală și aliniate temporal printr-o succesiune de module sensibile la instanțe. În final, capul de predicție bazat pe corelații (Correlation-Based Prediction Head – CBPH) generează o hartă densă a probabilității de schimbare.

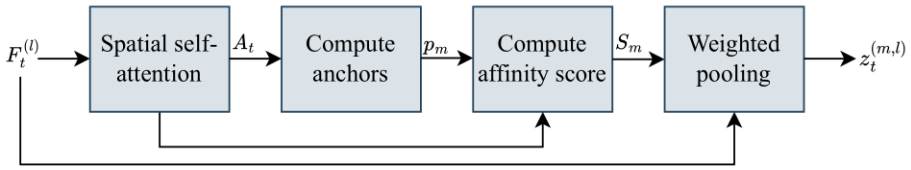

Figura 15. Instance-Consistent Tokenizer (ICT). Răspunsurile de atenție sunt utilizate pentru a identifica puncte-ancoră de tip „vârf" (peak anchors), din care se derivă hărți de afinitate spațială pentru a construi măști „soft". Aceste măști sunt apoi folosite pentru a extrage tokeni la nivel de instanță prin agregare ponderată (weighted pooling).

Evaluăm RADNet pe trei benchmark-uri publice pentru detectarea schimbărilor clădirilor (EGY-BCD [29], LEVIR-CD [30] şi SYSU-CD [31]) şi îl comparăm cu baseline-uri reprezentative bazate pe CNN, modele hibride şi modele bazate pe transformeri (SiamUNet-Diff, A2Net, BIT, CASP-Mb0 [32], ChangeFormer, ICIF-Net, DMINet şi RDP-Net).

Pe EGY-BCD, **Tabel 10** arată că RADNet obţine cele mai bune rezultate la toate metricile (67,51% IoU, 81,27% F1), depăşind BIT şi CASP-Mb0, menţinând în acelaşi timp cel mai bun echilibru între precizie şi recall. Pe LEVIR-CD, **Tabel 11** indică faptul că RADNet se clasează din nou pe primul loc (82,34% IoU, 90,26% F1, 99,05% acurateţe), confirmând îmbunătăţiri consistente pentru detectarea schimbărilor clădirilor la scară mare. Studiile de ablaţie evidenţiază, în plus, un compromis favorabil acurateţe–eficienţă faţă de arhitecturile puternic bazate pe transformeri (în special un număr mai mic de FLOPs şi un timp de antrenare considerabil redus comparativ cu ChangeFormer), păstrând performanţe competitive şi la inferenţă.

Pe SYSU-CD, **Tabel 12** arată că RADNet atinge cel mai mare IoU (65,42%) şi F1 (78,12%), precum şi cel mai bun recall (82,38%), ceea ce indică o sensibilitate crescută la tipare variate de schimbare în scene eterogene. Rezultatele calitative din **Figura 16** confirmă tendinţele cantitative la nivelul tuturor seturilor de date, evidenţiind contururi mai curate ale clădirilor, mai puţine detecţii false şi hărţi de schimbare mai puţin fragmentate.

În final, ablaţiile pe LEVIR-CD arată contribuţii complementare ale componentelor: tokenizarea consistentă la nivel de instanţă (scăderea cea mai mare apare la eliminarea acesteia), codificarea cu atenţie global–locală, decodarea temporală cross-scale şi predicţia bazată pe corelaţii. Aceste rezultate susţin rolul alinierii tokenilor şi al raţionamentului bazat pe corelaţii în localizarea robustă a schimbărilor clădirilor.

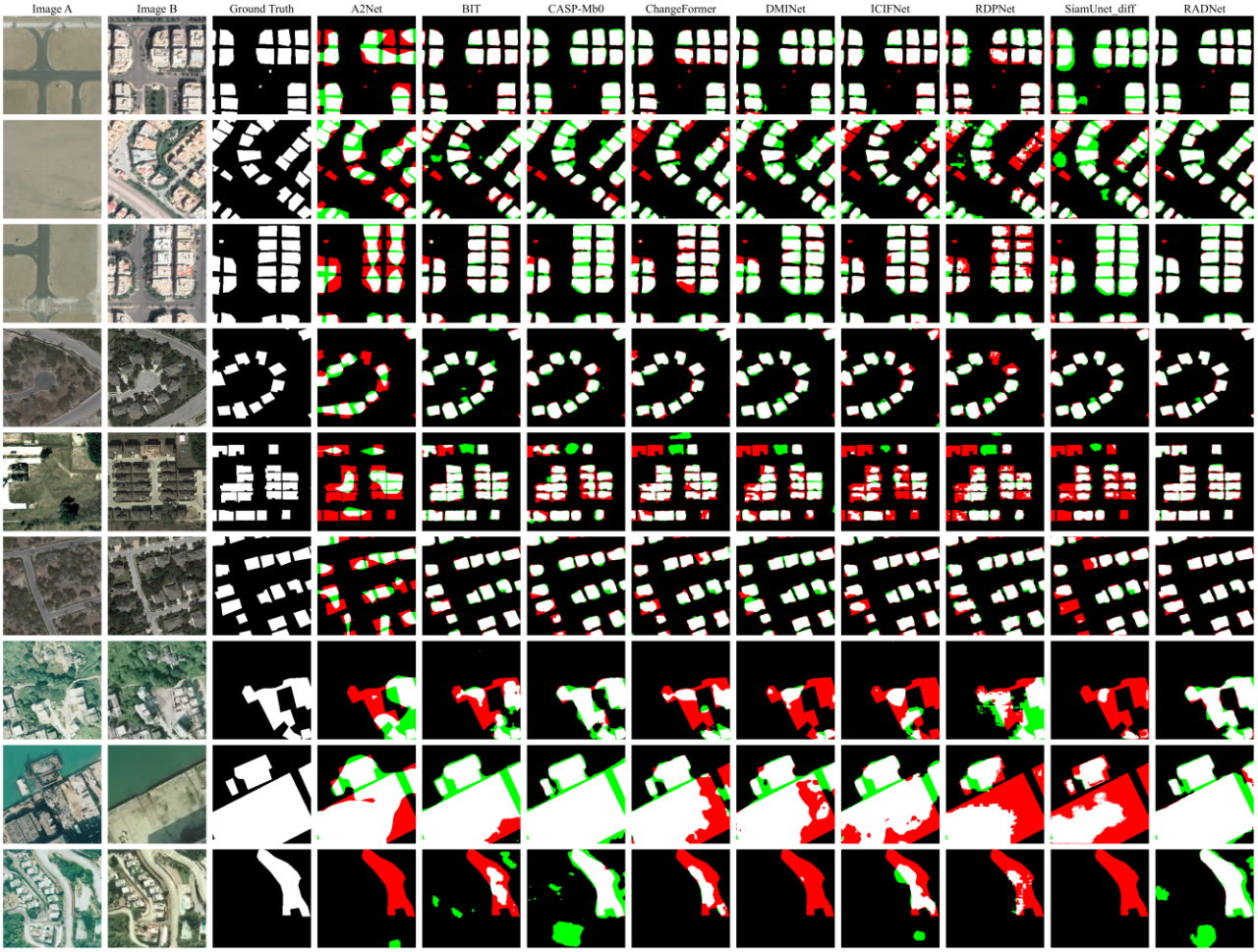

Figura 16. Comparație vizuală a rezultatelor de detectare a schimbărilor. Rândurile 1–3 prezintă exemple din setul de date EGY-BCD, rândurile 4–6 sunt din setul de date LEVIR-CD, iar rândurile 7–9 sunt din setul de date SYSU-CD. Alb: adevărat pozitiv; roșu: fals negativ; verde: fals pozitiv; negru: adevărat negativ.

Tabel 10. Rezultate cantitative pe setul de date EGY-BCD.

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**

Tabel 11. Rezultate cantitative pe setul de date LEVIR-CD.

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**

Tabel 12. Rezultate cantitative pe setul de date SYSU-CD.

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**

Direcții de cercetare viitoare

Identificăm direcții de dezvoltare viitoare pe trei axe complementare: date, modelare și implementare operațională. Din perspectiva datelor, o prioritate este extinderea generării sintetice controlabile către o diversitate ecologică mai bogată și către efecte de achiziție mai realiste (de exemplu, fenologie sezonieră, umbre, ceață atmosferică, estompare datorată mișcării și zgomot de senzor), precum și consolidarea legăturii dintre imagistica simulată și cea reală prin adaptare de domeniu și calibrare între domenii. Din perspectiva modelării, pașii următori includ învățarea sensibilă la incertitudine (de exemplu, aproximări bayesiene și estimarea calibrată a încrederii) pentru a susține decizii dependente de risc, împreună cu integrarea mai explicită a constrângerilor geometrice și a indiciilor de contur pentru segmentare și detectarea schimbărilor, precum și îmbunătățirea numărării multimodale prin obiective distribuționale și raționament condiționat pe regiuni. Pentru analiza temporală, o direcție firească este extinderea detectării schimbărilor de la intrări bi-temporale la secvențe mai lungi, permițând caracterizarea schimbărilor în funcție de persistență și estimarea tendințelor, în locul unor hărți binare ale schimbării. În final, în vederea utilizării în aplicații reale, cercetarea viitoare trebuie să pună accent pe eficiență și fiabilitate: backbone-uri ușoare pentru inferență la marginea rețelei (edge inference) pe platforme UAV, testarea robusteții în condiții nefavorabile și protocoale de evaluare de tip human-in-the-loop, care să cuantifice modul în care explicațiile modelului și scorurile de încredere influențează deciziile de inspecție și navigație.

Contribuții personale

Această teză urmărește obiectivul general de a avansa metodele de percepție vizuală și înțelegere a scenei care susțin luarea deciziilor relevante pentru control și navigație în contexte aeriene și de teledetecție. Contribuțiile personale se aliniază acestui obiectiv prin dezvoltarea unor resurse de date și de evaluare care reflectă constrângeri operaționale, prin propunerea unor arhitecturi adaptate structurii imaginilor aeriene și supervizării multimodale, precum și prin furnizarea unor dovezi experimentale sistematice, bazate pe comparații, studii de ablație și analize ale cazurilor de eșec, pentru sarcini reprezentative.

Pentru obiectivul de îmbunătățire a înțelegerii semantice în imagini UAV și aeriene, teza oferă o perspectivă structurată asupra modului în care disponibilitatea datelor, calitatea etichetelor și variabilitatea controlabilă influențează performanța segmentării și capacitatea de generalizare. În această direcție, lucrarea pune accent pe generarea de date prin simulare ca

mecanism practic de obținere a imaginilor dens adnotate la scară mare, în condiții de achiziție controlate (altitudine, punct de vedere și vreme). Resursele rezultate și analizele asociate sunt utilizate pentru a studia comportamentul modelelor în prezența schimbărilor de domeniu și pentru a clarifica factorii de scenă care influențează cel mai mult robustețea segmentării pe clase în medii centrate pe păduri și vegetație.

Pentru obiectivul de consolidare a raționamentului multimodal în teledetecție, teza dezvoltă un flux de lucru pentru Răspunsuri la Întrebări Vizuale (VQA) specializat pentru evaluarea post-dezastru, cu un accent particular pe acuratețea numărării și pe fiabilitatea răspunsurilor în scene eterogene. Contribuțiile includ o formulare multimodală coerentă, care combină reprezentări lingvistice cu caracteristici vizuale, și tratarea explicită a numărării ca problemă de predicție sensibilă la distribuție. Prin alegeri de proiectare țintite și studii controlate, teza identifică mecanisme practice care reduc erorile tipice de numărare (subnumărarea în regiuni aglomerate și supra-numărarea indusă de texturi repetitive), menținând totodată arhitectura compatibilă cu rezoluțiile specifice imaginilor de teledetecție și cu limitările seturilor de date.

Pentru obiectivul de îmbunătățire a raționamentului temporal și a detectării fine a schimbărilor, teza introduce un cadru de detectare a schimbărilor fără encoder–decoder, care modelează direct interacțiunile dintre caracteristicile extrase la momente diferite și pune accent pe raționament sensibil la localitate pentru schimbări mici și subtile. Lucrarea contribuie cu componente modulare pentru fuziune temporală și modelarea discrepanțelor, împreună cu o metodologie de evaluare care izolează impactul fiecărei alegeri de proiectare prin studii de ablație. Dincolo de scorurile agregate, teza oferă analize calitative și orientate pe erori, pentru a caracteriza cazuri tipice de eșec (false pozitive induse de iluminare și ambiguități la contururi) și pentru a motiva decizii de proiectare care echilibrează sensibilitatea și precizia în contexte operaționale.

În rezumat, teza aduce următoarele contribuții originale:
Segmentare semantică pentru înțelegerea scenelor UAV și aeriene.

1. Un set de date sintetic UAV pentru segmentare semantică în medii forestiere, proiectat cu condiții de achiziție controlabile și adnotări dense la nivel de pixel, pentru antrenare și evaluare reproductibile în scenarii forestiere.
2. O rețea neuronală spațio-temporală slab supervizată pentru segmentarea videoclipurilor UAV, bazată pe propagarea etichetelor și rafinare temporală printr-un modul ST-GRU, incluzând o variantă GRU orientată spre eficiență pentru utilizare practică.
3. O arhitectură de segmentare semantică bazată pe transformeri (SwinFAN), care consolidează integrarea global–locală a caracteristicilor în imagini aeriene de înaltă rezoluție prin atenție la nivel de decodor și rafinarea reprezentărilor.
4. Îmbunătățiri metodologice în decodorul SwinFAN pentru creșterea modelării contextului global și a preciziei contururilor în segmentarea UAV::
    a. Guided Focal–Axial Attention (GFA) pentru fuziune global–locală a contextului, care integrează atenția focală pentru evidențierea regiunilor relevante la mai multe scări cu atenția axială pentru captarea eficientă a dependențelor pe distanțe mari de-a lungul axelor spațiale, susținând segmentarea precisă atât a structurilor mari, cât și a obiectelor mici în imagini UAV de înaltă rezoluție.
    b. Attention-Based Feature Refinement Head (AFRH) pentru îmbunătățirea contururilor și a detaliilor, rafinând ieșirile decodorului prin self-attention și operații convoluționale pentru delimitarea mai bună a marginilor și

recunoaşterea claselor mici sau slab reprezentate.

5. Un protocol de validare care combină evaluarea pe benchmark-uri, studii de ablaţie şi evaluare prin transfer, permiţând o analiză structurată a utilităţii setului de date şi a comportamentului modelului în condiţii de schimbare de domeniu.

Răspunsuri la Întrebări Vizuale în teledetecţie, cu accent pe evaluarea post-dezastru şi numărare.

1. O analiză a metodelor de răspunsur la întrebări vizuale în teledetecţie, cu accent pe dificultatea numărării în prezenţa variaţiei de scară, a aglomerării şi a distribuţiilor variate ale răspunsurilor.
2. DiViNeCaps, o arhitectură pentru VQA post-inundaţie, care integrează reprezentările text–imagine prin fuziune multimodală şi un modul de clasificare structurată, evaluată pe FloodNet.
3. DeVANet, o reţea VQA orientată spre numărare în teledetecţie, care integrează embedding-uri multi-scală ale imaginii, fuziune multimodală a caracteristicilor şi predicţie cu două head-uri pentru modelarea numărării discrete şi continue.
4. Îmbunătăţiri metodologice introduse în DeVANet pentru creşterea robusteţii la numărare în condiţii de variaţie de scară şi distribuţii dezechilibrate ale numărării:
    a. Local-Global Attention Vision Mamba (LGA-ViM) pentru embedding-uri multi-scală, care integrează atenţia local–globală cu Vision Mamba prin înmulţire element-cu-element pentru a întări reprezentările pe mai multe scări spaţiale în imagini aeriene de înaltă rezoluţie.
    b. Bidirectional Fusion Attention (BFA) pentru integrarea multimodală, combinând self-attention în fiecare modalitate cu cross-attention bidirecţional şi co-attention pentru a îmbunătăţi alinierea dintre întrebări şi regiunile relevante din imagine.
    c. O formulare a funcţiei de pierdere sensibilă la distribuţie pentru numărări dezechilibrate şi supra-dispersate, care combină cross-entropy ponderată pentru clasificare discretă cu pierderea de tip negative binomial pentru regresia numărării, reducând instabilitatea optimizării şi îmbunătăţind acurateţea la numărări mari.
5. O evaluare experimentală extensivă pe FloodNet şi RescueNet, susţinută de studii de ablaţie care izolează impactul backbone-ului vizual, al mecanismului de fuziune, al head-urilor de predicţie şi al funcţiei de pierdere.

Detectarea fină a schimbărilor, cu tokenizare aliniată la structură şi raţionament temporal

1. O analiză structurată a metodelor pentru detectarea schimbărilor, care motivează tokenizarea aliniată la structură şi raţionamentul temporal în domenii în care obiectele şi regiunile coerente determină semantica schimbării.
2. SFCNet, o reţea pentru scene agricole structurate, incluzând un tokenizer semantic pe regiuni, un encoder temporal la nivel de regiune şi o strategie de decodare regiune–pixel pentru predicţia densă a hărţii de schimbare.
3. RADNet, o arhitectură orientată pe instanţe pentru detectarea schimbărilor clădirilor, incluzând tokenizare consistentă la nivel de instanţă, codificare cu atenţie global–locală şi decodare sensibilă la contururi pentru localizare precisă.
4. Îmbunătăţiri metodologice introduse în RADNet pentru aliniere semantică la nivel de instanţă şi raţionament contextual în scene urbane dense:
    a. Instance-Consistent Tokenizer (ICT) pentru aliniere semantică la nivel de

clădire, care generează tokeni de regiune aliniați semantic cu clădiri individuale sau structuri coerente, folosind indicii interne de atenție pentru a deduce limitele regiunilor direct din hărțile de caracteristici, permițând raționament la nivel de instanță fără măști de segmentare de referință la inferență.

b. Global-Local Attention Encoder (GLAE) pentru raționament contextual și geometric, care rafinează tokenii la nivel de instanță prin modelarea simultană a relațiilor semantice globale între regiuni (prin multihead self-attention) și a consistenței geometrice locale (prin atenție convoluțională), susținând detectarea schimbărilor la clădiri în condiții de densitate urbană ridicată.

5. O metodologie unificată de evaluare și ablație pe benchmark-uri de detectare a schimbărilor, care cuantifică rolul tokenizării, al fuziunii temporale, al strategiei de decodare și al proiectării capului de predicție.

Diseminare

Publicații în reviste ISI (4 ca prim autor):

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, Horia 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**

Prezentări la conferințe internaționale:

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.

Prezentări la workshop-uri:

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.

Cercetarea prezentată în această teză a fost realizată în cadrul mai multor proiecte ale Image Processing and Pattern Recognition Research Center, coordonat de Prof. Dr. Ing. Sergiu Nedevschi. Contribuţiile incluse în teză au fost dezvoltate în cadrul următoarelor proiecte:

- SEPCA – „Integrated Semantic Visual Perception and Control for Autonomous Systems", grant finanţat de Ministerul Educaţiei şi Cercetării din România, cod PN-III-P4-ID-PCCF-2016-0180.
- DeepPerception – „Deep Learning Based 3-D Perception for Autonomous Driving", grant finanţat de Ministerul Educaţiei şi Cercetării din România, cod PN-III-P4-PCE-2021-1134.
- Distributed Sensemaking – proiect finanţat de Lockheed Martin Australia.

Bibliografie

[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.