ACADEMIA ROMÂNĂ

Şcoala de Studii Avansate a Academiei Române

Institutul de Matematică ”Simion Stoilow”

REZUMATUL TEZEI DE DOCTORAT

Înspre Învățarea Nesupervizată a Consensului dintre Multiple Sarcini în Spațiu și Timp pentru Înțelegerea Scenelor Aeriene

CONDUCĂTOR DE DOCTORAT:

Prof. Dr. Marius Leordeanu

DOCTORAND:

Ing. Alina-Elena Marcu

2024

Cuprins

ABSTRACT

Această teză avansează înțelegerea scenelor aeriene prin abordări inovatoare de învățare profundă, concentrându-se pe sarcini precum segmentarea semantică, estimarea adâncimii (în metri) și învățarea multiplelor sarcini cu supraveghere umană limitată. Lucrarea introduce o nouă rețea neurală convoluțională cu două ramuri de procesare în paralel (dual) pentru segmentarea obiectelor, atingând performanțe de vârf în segmentarea clădirilor și drumurilor. Cercetarea noastră propune noi seturi de date și metode pentru detectarea drumurilor, rețele convoluționale eficiente pentru aplicații cu răspuns în timp real la bordul vehiculelor aeriene fără pilot (VAFP) și estimarea nesupervizată a adâncimii folosind zboruri reale cu drone. Pentru a reduce necesarul de adnotări manuale în cazul segmentării scenelor în video, această teză propune o metodă automată de propagare a etichetelor, și introduce setul de date, Ruralscapes, pentru segmentarea semantică a scenei folosind date din lumea reală. Lucrarea noastră culminează cu Dronescapes, setul de date conceput pentru învățarea multiplelor sarcini în scene complexe din lumea reală, și abordări de învățare semi-supervizată folosind un graf de rețele neuronale și un model hipergraf ce folosește multiple reprezentări pentru înțelegerea holistică a scenei cu supraveghere umană foarte limitată. Prin contribuțiile aduse comunității de Vedere Computațională (eng. Computer Vision), Teledetecție (eng. Remote Sensing) și Robotică (eng. Robotics), cercetarea noastră deschide calea pentru dezvoltarea de drone autonome capabile să înțeleagă scene complexe din lumea reală.

Cuvinte cheie – *Înțelegerea Scenelor Aeriene, Vehicule Aeriene fără Pilot (VAFP), Roboți Aerieni, Teledetecție, Segmentare Semantică, Local-Global, Estimarea Nesupervizată a Adâncimii Metrice, Învățare prin Consens, Hipergrafuri Multi-sarcină, Consens Multi-sarcină, Învățare Profundă, Inteligență Artificială, Robotică*

Capitolul 1

INTRODUCERE

Domeniul Inteligenței Artificiale (IA) a cunoscut progrese remarcabile atribuite avansurilor în algoritmi, disponibilității unor cantități vaste de date și creșterii exponențiale a puterii de calcul, în special a unităților de procesare grafică [1]. Robotica se concentrează pe dezvoltarea de agenți fizici (roboți) capabili să interacționeze cu lumea reală, făcând legătura între inteligența digitală și acțiunea tangibilă și a fost puternic influențată de progresele în domeniul IA.

Descrierea succintă a lui Peter Corke surprinde esența a ceea ce constituie un *robot* [2] – *"o mașină orientată spre obiective care poate percepe, planifica și acționa"*. Această definiție subliniază componentele cheie ale sistemelor robotice: percepția, luarea deciziilor și interacțiunea cu lumea fizică. Dintre aceste componente, percepția, în special sub forma înțelegerii scenei, joacă un rol crucial în a permite roboților să navigheze și să interacționeze eficient cu mediul în care activează, prezentând propriul set de provocări unice. Complexitatea înțelegerii scenei provine din mai mulți factori. În primul rând, lumea vizuală este în mod inerent diversă și dinamică, necesitând ca sistemele să generalizeze peste o vastă gamă de scenarii posibile. În al doilea rând, înțelegerea unei scene trece dincolo de simpla recunoaștere a obiectelor; implică înțelegerea relațiilor spațio-temporale, a indiciilor contextuale și include chiar predicția intenției (în viitor), peste care adăugăm necesitatea obținerii unui răspuns în timp real.

Această teză introduce metode pentru obținerea unei înțelegeri holistice a scenei folosind mai multe reprezentări vizuale ale lumii. Această abordare implică soluționarea mai multor sarcini, chiar simultan, pentru a obține informații complementare și interpretări vizuale diverse, toate în contextul Roboticii și al aplicațiilor din lumea reală. Luăm în considerare atât structura spațială tridimensională a lumii reale, cât și evoluția temporală, tinzând spre o înțelegere semantică care îmbunătățește capacitățile de navigare. Adăugând un alt strat de complexitate, explorăm aceste concepte dintr-o perspectivă aeriană, care prezintă provocări și oportunități unice. Pentru a ne atinge obiectivele, folosim consensul dintre multiple surse de informații și combinăm metode de învățare adâncă din date. Înțelegem importanța și dificultatea obținerii de date de înaltă calitate pentru a spori eficiența acestor algoritmi și ne concentrăm în principal pe dezvoltarea de abordări care maximizează utilizarea senzorilor în timp ce minimizează intervenția umană, abordând astfel mai eficient complexitățile lumii reale.

Cercetarea noastră a evoluat de la abordarea individuală a sarcinilor la dezvoltarea unor abordări inovatoare de combinare a mai multor sarcini, pentru înțelegerea holistică a scenei. Lucrarea face legătura între învățarea supervizată și cea nesupervizată prin exploatarea consensului care apare între multiple interpretări vizuale cu minimă supraveghere umană. Ne concentrăm pe îmbunătățirea înțelegerii scenei în aria roboților aerieni deoarece aceaștia prezintă provocări unice care necesită soluții inovatoare. **Principalele contribuții** aduse de munca noastră în acesta teză sunt rezumate mai jos:

Învățarea consensului local-global în mai multe etape – Călătoria noastră a început cu imagini satelitare, unde am introdus învățarea consens local-global și în mai multe etape pentru segmentarea aeriană. Această abordare a demonstrat puterea combinării aspectului local al obiectelor cu informațiile contextuale globale pentru două sarcini importante în domeniul Teledetecției, segmentarea clădirilor și a drumurilor. Am introdus, de asemenea, rețele neurale în multiple etape și pentru multiple sarcini capabile de segmentare și geolocalizare în același timp. Într-o manieră similară, în mai mult etape, arătăm și îmbunătățirea segmentării drumurilor de lățimi variate și în medii îndepărtate dificile, cu performanțe de vârf pe seturi de date relavante în domeniu.

Rețele convoluționale eficiente cu răspuns în timp real pe VAFP – Cercetarea folosind imagini din satelit la cea folosind date preluate de VAFP a fost o tranziție naturală și necesară, determinată de raritatea seturilor de date cuprinzătoare pentru VAFP și de provocările unice prezentate de acest domeniu. Recunoscând constrângerile computaționale ale VAFP, am dezvoltat modele eficiente adecvate pentru implementarea la bord, cum ar fi SafeUAVNets. Aceste arhitecturi constituie baza pentru toate contribuțiile noastre ulterioare, dovedind adaptabilitatea și eficacitatea lor pentru multiple sarcini.

Abordări hibride (analitice și învățate) – Am combinat abordări analitice și geometrice cu tehnici de învățare profundă din date date, așa cum se vede în metoda noastră de distilare a adâncimii în metri și UFODepth pentru a anula limitările reciproce și o estimare eficientă adâncimii din imagini, într-un mod nesupervizat.

Învățare cu intervenție umană minimală – Am dezvoltat tehnici noi de învățare semi-supervizată, inclusiv metoda noastră SegProp care folosește consensul spațio-temporal pentru propagarea automată a adnotărilor și modelele noastre bazate pe grafuri și hipergrafuri care necesită un set limitat de data adnotate pentru a produce rezultate convingătoare și oferă un avantaj în analiza imaginilor aeriene, unde datele etichetate sunt adesea rare și costisitoare.

Învățarea consensului dintre multiple sarcini – Cercetarea noastră a introdus, de asemenea, mecanisme inovatoare de învățare a multiplelor sarcini evoluând de la modelul bazat pe un graf de rețele neurale la modelul mai complex, de tip hipergraf. Aceste abordări exploatează interdependențele dintre diverse interpretări ale scenei ducând la o înțelegere mai robustă a scenei. Găsind consensul dintre multiple sarcini, am demonstrat îmbunătățirea performanței tuturor sarcinilor simultan cu mai puțin de 2% date etichetate manual.

Seturi de date complexe pentru VAFP – Pentru a susține și valida cercetarea noastră, am introdus noi seturi de date precum *Ruralscapes* pentru segmentarea semantică a scenei, *seturi de date de odometrie* pentru estimarea nesupervizată a adâncimii în metri și *Dronescapes* pentru învățarea multiplelor sarcini din zboruri reale preluate de VAFP. Aceste contribuții oferă resurse valoroase pentru comunitatea de cercetare și stabilesc noi standarde pentru evaluarea metodelor de înțelegere a scenelor aeriene.

Capitolul 2

ÎNVĂŢAREA ÎN MULTIPLE ETAPE A CONSENSULUI LOCAL-GLOBAL PENTRU SEGMENTARE AERIANĂ

2.1 Învățarea Consensului Local-Global

Studiem importanța folosirii contextului în segmentarea eficientă a obiectelor din imagini satelitare care oferă o vedere de sus în jos, sunt realizate în condiții de iluminare slabă și la rezoluție scăzută. Aspectul local al obiectelor în imaginile aeriene este adesea degradat din cauza ocluziunilor, iluminării și umbrelor. În astfel de cazuri, indiciile contextuale oferă informații semantice care îmbunătățesc recunoașterea obiectelor. Această lucrare propune o abordare cu ramuri duale de procesare integrate în rețele neurale convoluționale profunde care combină aspectul *local* al obiectului cu informații

*Această secțiune se bazează pe lucrarea – **Alina Marcu**, and Marius Leordeanu. "Object contra context: Dual local-global semantic segmentation in aerial images." In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence. 2017. [3]*

globale recuperate dintr-o regiune mai largă. Astfel, obiectul este văzut atât ca o entitate separată din perspectiva propriului aspect, cât și ca parte a unei scene mai mari care acționează complementar și conține implicit informații despre acesta.

Am studiat rolul informațiilor locale și globale separat (Figura 2.1). Un model antrenat doar cu informație locală sub forma unor regiuni de mici dimensiuni (*L-Seg*) este capabil să identifice cu acuratețe forma obiectului. Cu toate acestea, are o rată ridicată de predicții fals-pozitive. Pe de altă parte, un model antrenat exclusiv cu informație globală (*G-Seg*) pierde detalii importante, dar răspunde cu încredere în zonele cu densitate rezidențială ridicată și reușește să elimine clasificările eronate ale modelului anterior.

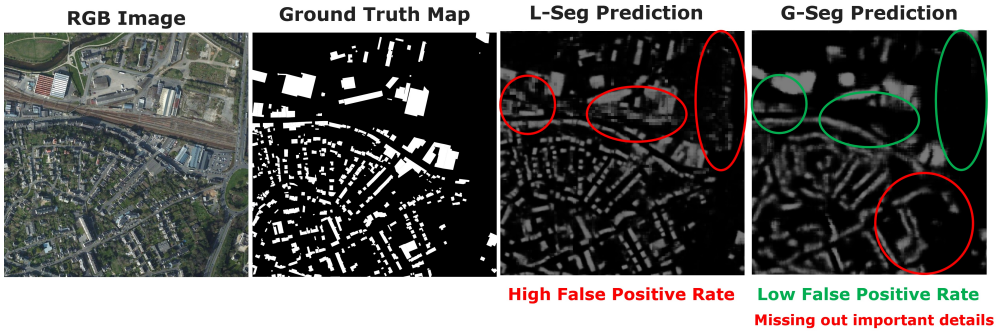


FIGURA 2.1: Modelul antrenat exclusiv cu informație locală identifică formele obiectelor, dar produc multe răspunsuri fals-pozitive (*L-Seg*). Modelul global pierde detalii, dar excelează în zone cu densitate rezindențială ridicată, reducând clasificările eronate (*G-Seg*). Abordarea noastră valorifică rolurile complementare ale informațiilor locale și globale.

Pe baza acestor observații, dezvoltăm două arhitecturi noi cu ramuri duale de procesare, combinând mai întâi modelul L-Seg și modelul G-Seg într-o singură rețea profundă local-globală, numită *LG-Seg* (Figura 2.2 (Stânga)). Cele două căi procesează informațiile în paralel, luând ca intrare regiuni din imagine de dimensiuni diferite. Fluxul local se concentrează pe capturarea detaliilor fine ale aspectului obiectului, în timp ce fluxul global adună un context semantic mai larg din jurul obiectului. Deși complementare, ambele fluxuri combinate (prin straturi complet conectate) ajung la consens printr-o procedură de învățare de la un capăt la altul pentru îmbunătățirea performanței segmentării. Spre deosebire de alții, am ales intenționat două tipuri diferite de arhitecturi cu dimensiuni diferite ale regiunilor din imagine ca intrare, pentru a încuraja învățarea de reprezentări diferite de-a lungul celor două căi.

Modelul LG-Seg are o arhitectură combinată cu două fluxuri, o rețea de tip VGG-Net adaptată [4] și o rețea de tip AlexNet adaptată [5] care se unesc în ultimele două straturi

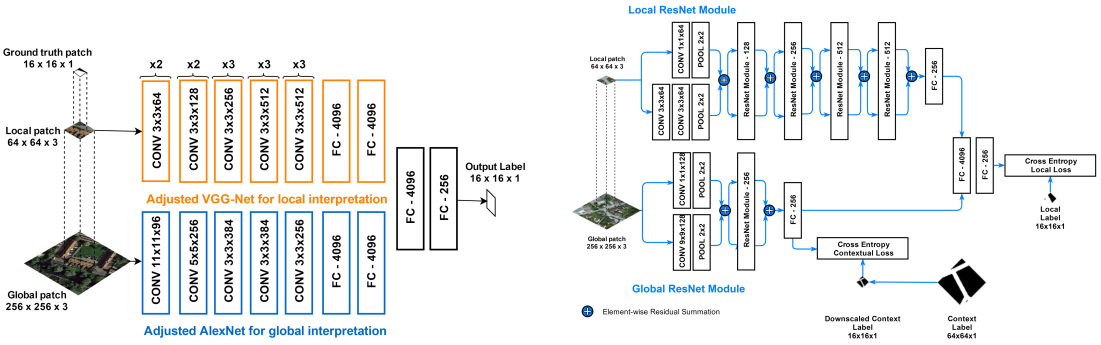


FIGURA 2.2: (**Stânga**) Arhitectura noastră LG-Seg (≈133M parametri) combină informații locale și globale într-o singură rețea convoluțională cu două fluxuri cu dimensiuni diferite ale regiunilor de intrare, încurajând reprezentări diverse de-a lungul celor două căi. (**Dreapta**) LG-Seg-ResNet-IL folosește blocuri reziduale și o funcție de cost intermediară pe ramura globală, îmbunătățind înțelegerea contextuală cu o eficiență mai mare decât LG-Seg și doar ≈23M parametri.

complet conectate, care ajung în total la ≈133M parametri. Având în vedere cerințele computaționale substanțiale ale modelului LG-Seg, folosim conexiuni reziduale [6], deoarece acestea sunt capabile să ocolească niveluri suplimentare de adâncime și astfel să combine simultan căi superficiale și profunde într-o rețea neurală multi-ramură unificată cu filtre de dimensiuni diferite care acționează în paralel, cu un număr redus de parametri, numit *LG-Seg-ResNet*. Pe lângă această arhitectură, introducem și *LG-Seg-ResNet-IL* (Figura 2.2 (Dreapta)) în care am adăugat o constrângere intermediară suplimentară pentru calea globală pentru a impune învățarea contextului global. Astfel, sperăm să îmbunătățim timpul de antrenament și, de asemenea, calitatea segmentării în locurile unde contextul contează mai mult.

TABELA 2.1: Evaluare cantitativă a modelelor noastre pe diferite seturi de date. De la stânga la dreapta în ordine, prezentăm rezultate pe setul de date Massachusetts Buildings și pe seturile noastre de date European Buildings și European Roads.

Metodă	F-măsură	Metodă	F-măsură	Metodă	F-măsură
Mnih et al. [7]	92.11	G-Seg	62.71		
Saito et al. [8]	92.30	L-Seg	82.66	LG-Seg	70.46
LG-Seg	94.23	LG-Seg	**84.20**	LG-Seg-ResNet	72.07
LG-Seg-ResNet-IL	**94.30**	LG-Seg-ResNet-IL	83.87	LG-Seg-ResNet-IL	**73.42**

Abordăm segmentarea obiectelor, în special, obiecte cu structuri regulate, cum ar fi **clădirile**, și cele cu forme variabile și continue, cum ar fi **drumurile**. De asemenea, propunem două noi seturi de date, mult mai mari decât cel mai cunoscut set de date la acea vreme, Massachusetts Buildings [7]. Experimentele au demonstrat următoarele. În primul rând, obținem rezultate de top pe setul de date public cu ambele arhitecturi propuse (Tabelul 2.1 (Stânga)). În al doilea rând, complexitatea seturilor de date propuse este

reflectată și de scăderea performanței de la 94% la 84% pentru segmentarea clădirilor pe setul de date European Buildings propus (Tabelul 2.1 (Mijloc)). În cele din urmă, putem observa că în cazul obiectelor cu structuri complexe, continue și variate, contextul este cu adevărat important, deoarece arhitectura cu module reziduale și context modelat explicit obține cele mai bune rezultate pentru problema segmentării drumurilor, cu o creștere mai mare a performanței, de 3% comparativ cu modelul LG-Seg inițial (Tabelul 2.1 (Dreapta)).

2.2 Rețele Neurale în Multiple Etape pentru Multiple Sarcini

În timp ce rețelele noastre anterioare local-globale au demonstrat importanța valorificării atât a informațiilor locale, cât și a celor contextuale pentru sarcini precum segmentarea clădirilor și drumurilor, acestea aveau limitarea de a trata fiecare sarcină independent. Introducem o nouă rețea neurală cu multiple etape pentru multiple sarcini (MSMT), care soluționează limitarea anterioară și abordează simultan două sarcini cruciale aeriene: segmentarea semantică și geolocalizarea din imagini. Arhitectura propusă (Figura 2.3) folosește o ramură comună de codificare (*UniEncoder*), urmată de ramuri de decodificare separate (regresie sau segmentare) pentru fiecare sarcină (*LocDecoder*). Această alegere de design este inspirată de lucrarea noastră anterioară, care a arătat că procesarea diferitelor tipuri de informații prin fluxuri separate, mai ales când ramurile rezolvă două scopuri diferite, este benefică. Aspectul în mai multe etape al rețelei permite rafinarea progresivă a rezultatelor. Din punct de vedere structural, arhitectura noastră folosește module de tip encodor-decodor (similare cu U-Net [10] dar utilizând convoluții dilatate cu rate progresiv crescătoare la nivelul central) la fiecare etapă.

Modelul nostru MSMT-Stage-1 stabilește un nou reper de performanță pe setul de date Massachusetts Buildings, depășind semnificativ metodele anterioare (Tabelul 2.2

*Această secțiune se bazează pe lucrarea – **Alina Marcu**, Dragos Costea, Emil Slusanschi, and Marius Leordeanu. "A multi-stage multi-task neural network for aerial scene interpretation and geolocalization." arXiv preprint arXiv:1804.01322 (2018). [9]*

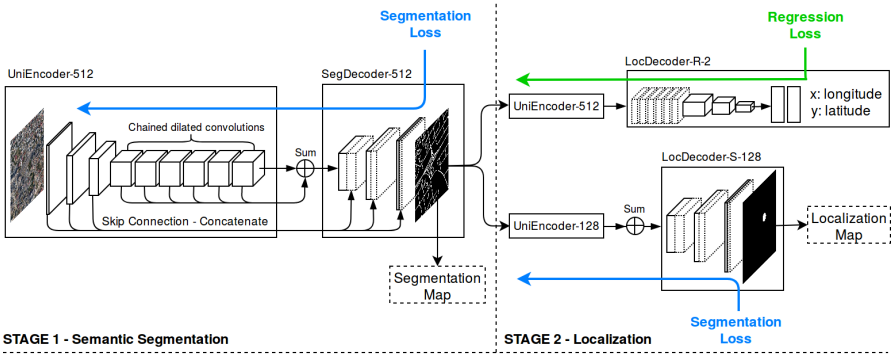


FIGURA 2.3: Arhitectura propusă în multiple etape pentru multiple sarcini (MSMT) pentru segmentare semantică și geolocalizare în imagini aeriene.

(Stânga) raportează metrica F1-score cu un factor de relaxare de 3 pixeli [11]). Abordarea noastră obține aceste rezultate folosind predicția de la un singur model, în contrast cu metode precum [12] care folosesc ansambluri de rețele. Același lucru poate fi observat pentru setul de date Inria [13] pe setul de testare (Tabelul 2.2 (Dreapta) măsurând IoU), depășind alți competitori (rezultate verificate pe clasamentul oficial).

TABELA 2.2: (**Stânga**) Comparație cu metodele de segmentare a clădirilor de ultimă generație pe setul de date Massachusetts Buildings [7]. (**Dreapta**) Comparație cu metodele de ultimă generație pentru sarcina de segmentare a clădirilor din imagini aeriene pe setul de date Inria [13].

Metodă		F1-Score
Deeplab	[14]	89.7
Mnih et al.	[7]	91.5
Saito et al.	[8]	92.3
U-Net	[10]	94.1
Saito et al.	[12]	94.3
Marcu et al.	[3]	94.3
Hamaguchi et al.	[15]	94.3
MSMT-Stage-1		**96.04**

Metodă		Austin	Chicago	Kitsap Co.	West Tyrol	Vienna	General
MLP [13]	IoU	61.20	61.30	51.50	57.95	72.13	64.67
	Acuratețe	94.20	90.43	98.92	96.66	91.87	94.42
Mask R-CNN [16]	IoU	65.63	48.07	54.38	70.84	64.40	59.53
	Acuratețe	94.09	85.56	97.32	**98.14**	87.40	92.49
SegNet MT-Loss[17]	IoU	**76.76**	67.06	**73.30**	66.91	76.68	73.00
	Acuratețe	93.21	**99.25**	97.84	91.71	**96.61**	95.73
MSMT-Stage-1	IoU	75.39	**67.93**	66.35	**74.07**	**77.12**	**73.31**
	Acuratețe	**95.99**	92.02	**99.24**	97.78	92.49	**96.06**

Pentru sarcina de geolocalizare, am colectat propriul nostru set de date din locații eșantionate aleatoriu acoperind o zonă de dimensiunea unui oraș. După cum au demonstrat Costea et al. [18], drumurile servesc ca o amprentă unică a unei zone urbane, prin urmare am decis să folosim segmentări de drumuri (antrenate folosind MSMT-Stage-1). A doua etapă a rețelei noastre ia ca intrare rezultatul segmentării drumului și învață să îl mapeze la o locație specifică folosind două ramuri. Ramura de regresie produce coordonatele de longitudine și latitudine, în timp ce ramura de segmentare modelează localizarea ca o problemă de segmentare, generând o hartă cu locații potențiale marcate ca puncte. În experimentele noastre, am demonstrat că, în timp ce ramura de segmentare depășește în general ramura de regresie, aceasta poate produce uneori locații multiple sau poate eșua în identificarea vreunei (imagine goală). În astfel de cazuri, ne

bazăm pe rezultatul dat de regresie. Pentru a crește precizia localizării ca un pas final de rafinare, folosim algoritmul Iterative Closest Point (ICP) [19], pentru a alinia segmentarea drumului cu drumurile din OpenStreetMap [20] de la locația prezisă. În Figura 2.4 prezentăm, de la stânga la dreapta, în ordine, imaginea RGB de intrare, localizarea punctului la nivel de oraș generată de MSMT-Stage-2-LocDecoder-S-128, drumurile prezise (verde) din imaginea RGB peste drumurile reale de la locația punctului (alb), înainte și după aliniere.

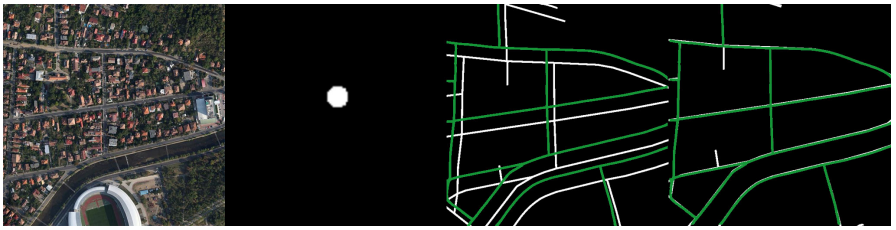


FIGURA 2.4: Rezultate calitative pentru geolocalizare folosind ramura MSMT-Stage-2-LocDecoder-S-128, harta de segmentare a drumurilor folosind rețeaua MSMT-Stage-1 și drumurile recuperate din OpenStreetMap, înainte și după aliniere.

2.3 Ansamblu de Rețele Neurale Adânci în Multiple Etape

Bazându-ne pe arhitectura noastră anterioară (MSMT) detaliată în secțiunea anterioară, abordăm problema generării unei hărți rutiere din imagini satelitare. Această secțiune prezintă abordarea noastră submisă în cadrul competiției DeepGlobe [22], unde am obținut o îmbunătățire semnificativă de peste 4% comparativ cu concurentul de pe locul doi. O prezentare generală a abordării noastre se regăsește în Figura 2.5. Prima etapă se bazează pe arhitectura MSMT-Stage-1 (aceeași arhitectură de tip U-Net cu convoluții dilatate la nivelul central) folosită pentru a crea un ansamblu de mai multe astfel de rețele, fiecare folosind rate de dilatare diferite (și un număr diferit de parametri). Această abordare permite modelului să învețe diverse aspecte ale scenei, îmbunătățind segmentarea binară a drumurilor. Ratele de dilatare variate permit rețelei să capteze caracteristici la mai multe scale, ceea ce este crucial pentru identificarea precisă a drumurilor de lățimi diferite și în diverse contexte. Antrenăm următoarele tipuri de arhitecturi (cu

*Această secțiune se bazează pe lucrarea – **Alina Marcu***, Dragos Costea*, Emil Slusanschi, and Marius Leordeanu. "Roadmap generation using a multi-stage ensemble of deep neural networks with smoothing-based optimization." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 220-224. 2018. [21]*

rate de dilatare progresive în paranteze) pentru segmentarea drumurilor și formăm două ansambluri din ele (Ansamblul 1 - sumă, Ansamblul 2 - prin învățare (etapa 2)): (1) **Dilatare maximă 32** (1, 2, 4, 8, 16, 32), (2) **Dilatare maximă 48** (1, 2, 4, 8, 16, 32, 48) și (3) **Dilatare maximă 64** (1, 2, 4, 8, 16, 32, 48, 64). A doua etapă introduce un proces de rafinare nou, ce implică antrenarea unei rețele Dilatare maximă 32 pentru să îmbunătățirea adițională a segmentările drumurilor într-o manieră strict supervizată.

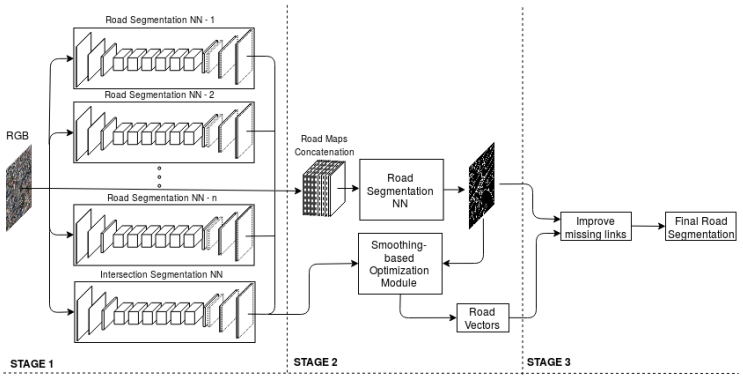


FIGURA 2.5: Prezentare generală a metodei noastre în trei etape pentru îmbunătățirea generării hărții rutiere.

Inovația cheie aici este fuziunea mai multor hărți de drumuri alături de imaginea RGB și, de asemenea, segmentarea intersecțiilor prezise de o rețea diferită antrenată pentru a îmbunătăți conexiunile drumurilor. Acest lucru permite modelului să învețe un consens între aceste diverse reprezentări, rezultând o segmentare a drumurilor mai precisă și consistentă. Această etapă poate fi văzută ca o extensie a pasului de rafinare a geolocalizării în arhitectura noastră MSMT, dar este acum aplicată specific segmentării drumurilor. Etapa finală este inspirată din lucrarea lui Costea et al. [21]. Acest pas aplică o procedură de optimizare drumurilor segmentate, tratându-le ca un graf. Scopul este de a adăuga legăturile lipsă bazate pe structura grafului inferată, îmbunătățind calitatea generală a segmentării. Această etapă completează abordarea noastră de învățare profundă cu tehnici clasice de Vedere Computțională, cu un potențial de a completa rupturile cauzate de rețeaua neurală. Arătăm o îmbunătățire consistentă pe setul de test DeepGlobe în ambele runde ale competiției folosind arhitecturile noastre antrenate în iterații semi-supervizate (cu predicții pe datele de testare folosite în a doua iterație) (Tabelul 2.3 (Stânga)) și după cum era de așteptat, modelul cu cea mai bună performanță fiind ansamblul învățat (Tabelul 2.3 (Dreapta)), depășind modelul de referință cu $\approx 4\%$.

TABELA 2.3: (**Stânga**) (Runda 1) – Rezultatele segmentării drumurilor pe setul oficial de evaluare de 1.243 de imagini pentru care nu au fost furnizate date de referință în timpul competiției pentru a evita înșelăciunea și supraajustarea. (**Dreapta**) (Runda 2) – Rezultatele segmentării drumurilor pe setul oficial de testare de 1.101 imagini. Rezultatele au fost raportate după adăugarea legăturilor lipsă. Rezultatele (raportate în procente) au fost furnizate de site-ul oficial al competiției pentru ambele runde.

Model	Iterație	IoU Evaluare
Dilatare maximă 32	1	59.24
	2	59.75
Dilatare maximă 48	1	60.39
	2	**60.58**

Model	IoU Evaluare
Referință [22]	54.5
Ansamblul 1	57.88
Ansamblul 2	**58.62**

2.4 Rețele Convoluționale Eficiente pentru Înțelegerea Scenei Aeriene

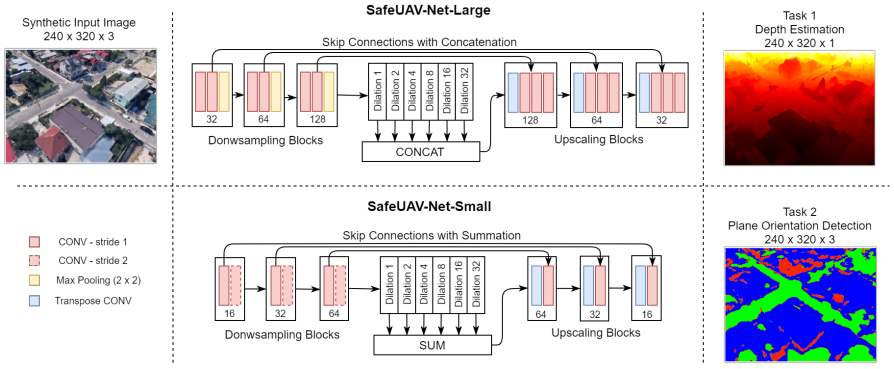


FIGURA 2.6: SafeUAVNets, rețele eficiente propuse pentru procesarea directă la bord a VAFP, sunt antrenate pentru estimarea adâncimii și predicția orientării planului.

Până acum ne-am concentrat în principal pe analiza imaginilor satelitare, care oferă o vedere de sus a suprafeței Pământului. În timp ce sateliții oferă acoperire globală de la altitudini foarte mari, vehiculele aeriene fără pilot (VAFP) au șase grade de libertate, fiind capabile să capteze atât vederi de sus, cât și vederi în perspectivă la altitudini mult mai mici comparativ cu sateliții. Această flexibilitate permite VAFP să-și ajusteze altitudinea pentru rezoluție și acoperire optimă, să colecteze date la scale variabile, de la priviri de ansamblu la detalii și să furnizeze imagini în timp real, la cerere, pentru aplicații critice. Această secțiune prezintă primele noastre eforturi în dezvoltarea unor soluții

*Această secțiune se bazează pe lucrarea – **Alina Marcu**, Dragos Costea, Vlad Licaret, Mihai Pirvu, Emil Slusanschi, and Marius Leordeanu. "SafeUAV: Learning to estimate depth and safe landing areas for UAVs from synthetic data." In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0-0. 2018. [23]*

eficiente care utilizează perspectiva vizuală pentru înțelegerea semantică și structurală a scenei din reconstrucții 3D ale mediilor reale, rulând totodată pe dispozitive cu resurse limitate. Inspirați de arhitecturile noastre anterioare, am dezvoltat două noi variante ale rețelei neurale MSMT-Stage-1 (sau Dilatare maximă 32). Prima, numită *SafeUAVNet-Large*, rulează la $\approx$ 35 FPS pe Nvidia Jetson TX2. A doua, numită *SafeUAVNet-Small*, este o versiune simplificată și rulează la $\approx$ 130 FPS pe același dispozitiv integrat pentru o imagine de $240 \times 320 \times 3$. Ambele arhitecturi sunt reprezentate în Figura 2.6.

Extragem date din Google Earth [24] și reconstruim scenele în 3D. Obținem imagini RGB și hărți dense de adâncime, cu valori la nivel de pixel în metri, din perspectiva unei drone care zboară la altitudini joase și la un unghi de 45 de grade. Pe baza normalelor suprafeței, se construiește un set de adnotări semantice pentru a delimita zonele sigure de cele nesigure, care pot fi folosite ulterior în procedura de aterizare sigură a unei drone. Evaluarea cantitativă a demonstrat eficiența celor două arhitecturi comparativ cu arhitecturi bine-cunoscute de predicție densă, atât pentru sarcina de segmentare a zonelor sigure, nesigure sau înclinate (altele) în scenă (Tabelul 2.4 (Stânga)), cât și pentru estimarea adâncimii în metri dintr-o imagine (Tabelul 2.4 (Dreapta)).

Am examinat, de asemenea, performanța modelelor antrenate pe date sintetice atunci când sunt aplicate în scenarii din lumea reală. În ciuda asemănărilor vizuale, am constatat diferențe semnificative de performanță între datele sintetice și cele reale. Acest lucru evidențiază necesitatea antrenării pe date din lumea reală pentru a asigura eficacitatea modelului în scenarii similare.

TABELA 2.4: Evaluare cantitativă a SafeUAVNets pentru (**Stânga**) Detecția zonelor sigure, modelată ca o problemă de segmentare a planurilor orizontale, verticale și înclinate într-o scenă. Rezultatele sunt raportate în procente, valorile cu caractere îngroșate fiind cele mai bune. (**Dreapta**) Sarcina de estimare a adâncimii din imagini aeriene. Numerele mai mici sunt mai bune.

Model	mAcc.	mPrec.	mRec.	mIoU
U-Net [10]	72.9	56.0	50.5	35.6
DeepLabv3+ [25]	84.0	75.3	73.9	59.7
SafeUAVNet-Small	82.3	72.8	69.3	55.1
SafeUAVNet-Large	**84.6**	**76.1**	**74.8**	**60.7**

Model	RMSE	Metri
U-Net [10]	0.041	9.63
DeepLabv3+ [25]	0.034	8.49
SafeUAVNet-Small	0.031	7.22
SafeUAVNet-Large	**0.026**	**6.09**

Capitolul 3

ESTIMAREA NESUPERVIZATĂ A ADÂNCIMII ÎN METRI FOLOSIND CONSENS

Acest capitol abordează o problemă esențială pentru sistemele aeriene autonome: estimarea precisă și eficientă a adâncimii în metri din imagini în perspectivă. În plus, comparativ cu metodele anterioare propuse în această teză, facem un pas uriaș către utilizarea eficientă a indiciilor temporale din video și extragem adâncimea în metri într-un mod complet nesupervizat. Abordările tradiționale s-au bazat fie pe metode pur geometrice, care pot fi precise, dar nu sunt robuste, fie pe abordări bazate pe învățare, care adesea se luptă cu generalizarea și păstrarea scării metrice. În acest capitol propunem o serie de contribuții inovatoare care combină sinergic metode analitice cu abordări de învățare profundă. Pentru a aborda lipsa de date relevante din lumea reală, introducem două seturi de date semnificative - un set de date de zbor continuu preluat direct de VAFP de 20 de minute care acoperă două stațiuni montane din România și un set de date extins de 33 de minute care cuprinde o varietate de scene, inclusiv urbane, rurale și Delta Dunării. Aceste seturi de date constituie repere cruciale pentru evaluarea metodelor de estimare a adâncimii în metri în scenarii realiste.

3.1 Distilarea Adâncimii: Formarea Consensului dintre Cinematică, Flux Optic și Învățarea Profundă

Considerăm scenariul unui zbor într-un perimetru dat, cu accent pe învățarea nesupervizată a adâncimii cu scopul înțelegerii structurii scenei aeriene. Introducem un model hibrid care combină o abordare analitică, folosind dometrie, cu tehnici de învățare profundă nesupervizată. În timp ce calea analitică este precisă matematic, îi lipsește robustețea în prezența zgomotului și eșuează numeric în zonele de expansiune focală. Pe de altă parte, abordarea nesupervizată bazată pe date este robustă, dar nu la fel de precisă și, mai important, nu este metrică. Folosim paradigma ”Profesor-Student” pentru a distila cunoștințele [27] ”Profesorului” într-un ”Student” mai compact. Prin formarea unui ansamblu din cele două căi (Profesor) și distilarea acestora într-o singură rețea (Student), reușim să îmbunătățim atât precizia, cât și să reducem cerințele computaționale.

Această metodă distilează consensul dintre geometria scenei, poziția camerei, cinematica și imaginea RGB într-o singură rețea neurală, obținând atât eficiență computațională, cât și precizie ridicată. Este important de menționat că această abordare este concepută pentru implementare pe dispozitive integrate, făcând-o potrivită pentru aplicații ale VAFP din lumea reală (deoarece se bazează pe arhitecturile SafeUAVNets). Prezentarea generală a abordării noastre se regăsește în Figura 3.1 (Stânga) în care combinăm mai multe căi complementare pentru estimarea precisă a adâncimiii în metri. De-a lungul unei căi, estimăm adâncimea non-metrică într-un mod nesupervizat (D_{Unsup}). De-a lungul altei căi, folosim odometria și fluxul optic pentru a estima adâncimea exactă, metrică ($D_{OdoFlow}$). $D_{OdoFlow}$ este utilizat pentru a scala D_{Unsup} și a-l face metric. Formăm ansamblul dintre cele două (Profesor) și îl utilizăm pentru a distila un model Student. De-a lungul unei a treia căi, reconstruim adâncimea scenei prin Structure-from-Motion [28]

*Această secțiune se bazează pe lucrarea – Mihai Pirvu, Victor Robu, Vlad Licaret, Dragos Costea, **Alina Marcu**, Emil Slusanschi, Rahul Sukthankar, and Marius Leordeanu. ”Depth distillation: unsupervised metric depth estimation for UAVs by finding consensus between kinematics, optical flow and deep learning.” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3215-3223. 2021. [26]*

(D_{SfM}), care are o scară metrică prin alinierea traiectoriei sale prezise cu traiectoria metrică din GPS. D_{SfM} reprezintă referința (deși nu este perfect) și este utilizat doar pentru evaluare.

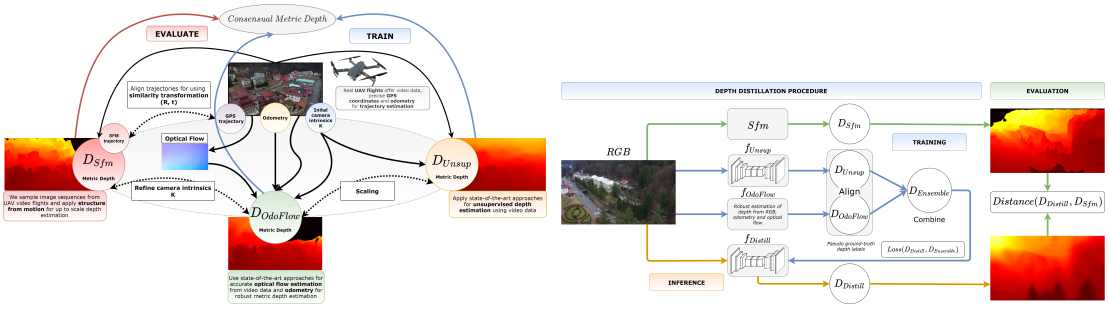


FIGURA 3.1: (**Stânga**) Prezentare generală a abordării noastre, combinând mai multe căi complementare pentru estimarea precisă a profunzimii metrice. (**Dreapta**) Procedura noastră de distilare a profunzimii pentru estimarea precisă a profunzimii metrice.

Introducem un nou set de date cu două zboruri continue în două stațiuni montane, **Slănic** și **Herculane**. Împărțim Slănic în două subseturi care nu se suprapun, unul utilizat pentru *antrenarea* rețelei neurale de distilare și celălalt pentru *testare*, măsurând cât de bine ar performa un VAFP în aceeași scenă într-un nou zbor. Herculane este folosit doar pentru testare, pentru a estima capacitățile de generalizare ale soluțiilor propuse, în scene noi diferite, nevăzute în timpul antrenării.

În Figura 3.1 (Dreapta) prezentăm procedura noastră de distilare a adâncimii metrice. Combinăm două metode (nesupervizată și analitică) într-un singur rezultat și îl evaluăm în raport cu reconstrucția SfM. În analiza noastră experimentală (Tabelul 3.1), am demonstrat că Studentul distilat poate îmbunătăți semnificativ performanța față de Profesor pe videoclipul de test din prima scenă, rămânând competitiv în a doua scenă, nevăzută în timpul antrenării.

TABELA 3.1: Erori absolute și relative în zonele unde atât $D_{OdoFlow}$, cât și D_{Unsup} sunt valide. Observăm că aceste zone produc rezultate mai stabile. Herculane are erori mai mari în principal din cauza înălțimii mai mari a setului de date. Studenții distilați (SafeUAVNets) au cele mai bune rezultate pe setul de test Slănic. Numerele mai mici sunt mai bune și cele mai bune sunt **îngroșate**.

	Slănic		**Herculane**	
	Metric (m)	Relativ (%)	Metric (m)	Relativ (%)
D_{Unsup}	21,06	15,31	31,61	16,60
$D_{OdoFlow}$	19,56	14,39	**24,97**	**12,72**
$D_{Ensemble}$	19,03	13,81	27,10	13,79
SafeUAVNet-Large	16,66	13,41	37,43	22,41
SafeUAVNet-Small	**16,11**	**12,90**	37,42	22,95

3.2 UFODepth: Învățarea Nesupervizată prin Optimizarea Odometriei folosind Flux Optic pentru Estimarea Adâncimii în Metri

Bazându-ne pe conceptul anterior de distilare a adâncimii, introducem o metodă de învățare nesupervizată pentru estimarea adâncimii în metri care încorporează o nouă procedură de optimizare a odometriei bazată pe fluxul optic. Această secțiune propune o îmbunătățire a precizia estimării analitice și demonstrează capacități superioare de generalizare folosind date video și odometrie din zboruri VAFP din lumea reală.

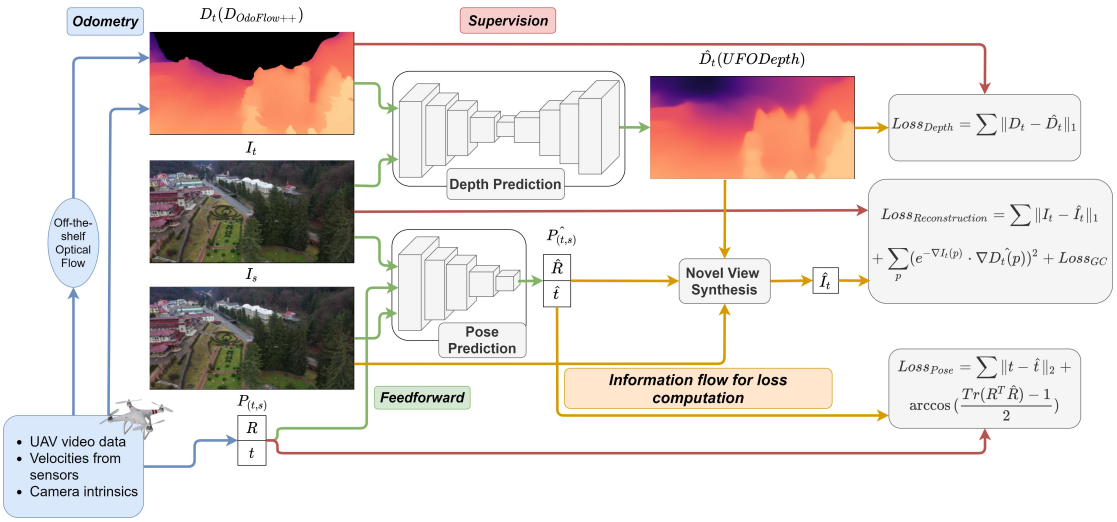


FIGURA 3.2: Prezentare generală a abordării noastre (UFODepth). Combinăm trei tipuri de funcții de cost cu o formulare matematică îmbunătățită, ceea ce rezultă într-o estimare mai robustă a adâncimii în scene variate, menținând în același timp inferența în timp real.

Corectăm măsurătorile odometrice zgomotoase prin optimizarea alinierii dintre fluxul optic reorientat și viteza liniară proiectată în imagine. Apoi, detaliăm o metodă analitică de estimare a adâncimii bazată pe fluxul optic și vitezele corectate ale camerei ($D_{OdoFlow++}$). Ulterior, harta de adâncimea și vitezele camerei obținute analitic sunt

Această secțiune se bazează pe lucrarea – Vlad Licaret, Victor Robu*, **Alina Marcu***, Dragos Costea, Emil Slusanschi, Rahul Sukthankar, and Marius Leordeanu."UFODepth: Unsupervised learning with flow-based odometry optimization for metric depth estimation." In 2022 International Conference on Robotics and Automation (ICRA), pp. 6526-6532. IEEE, 2022. [29]*

folosite ca funcții de cost pentru antrenarea noii noastre arhitecturi de învățare nesupervizată pentru estimarea adâncimii metrice (*UFODepth* prezentat în Figura 3.2). Experimentăm extensiv cu un set de date colectat din lumea reală, pe care îl extindem semnificativ adăugând scene complet noi. Extinderea setului de date constă în aproximativ 33 de minute de timp de zbor cu informații de odometrie și GPS. Acesta acoperă o varietate de scene cum ar fi urbane, rurale și din Delta Dunării (Figura 3.3).

FIGURA 3.3: Eșantioane din setul de date din toate secvențele video din setul de date extins. Cadre eșantion (sus) împreună cu traiectoriile estimate suprapuse peste reconstrucția SfM (jos). Arătăm eșantioane din lucrările anterioare Slănic, Herculane [26], alături de scenele nou introduse - Oveselu, Olănești, Chilia.

Depășim cu marje semnificative diferite abordări de ultimă generație, variind de la soluții analitice și nesupervizate la arhitecturi bazate pe arhitecturi complexe care necesită calcul intensiv și pre-antrenare (Tabelul 3.2). Chiar dacă unele detalii mici sunt pierdute, metoda noastră oferă o adâncime consistentă temporal și precisă. Deși vizual metode precum DPT sau BMD oferă o separare mai bună dintre fundal și prim-plan, acestea nu sunt la fel de potrivite pentru sarcini practice de estimare a adâncimii din cauza cerințelor lor computaționale intensive. Algoritmul nostru ar putea fi implementat pe dispozitive integrate, fiind un bun candidat pentru cazuri de utilizare practică în Robotică, cum ar fi evitarea obstacolelor și aterizarea sigură pentru VAFP.

TABELA 3.2: Erori medii absolute și relative pe întreaga hartă în comparație cu harta de adâncime de referință D_{SfM}. Metrica este reprezentată în metri (m) și Relativă ca procent (%). Deoarece raportăm erorile, pentru ambele metrici, numerele mai mici sunt mai bune. Metodele care nu oferă estimări metrice sunt scalate folosind $D_{OdoFlow++}$ pentru o comparație corectă. "Overall" denotă media peste toate scenele.

Metoda	Slănic		Chilia		Olănești		Herculane		Oveselu		Overall	
	Metric	Relativ	Metric	Relativ	Metric	Relativ	Metric	Relativ	Metric	Relativ	Metric	Relativ
D_{Unsup} [30]	25,00	15,4	44,52	23,4	25,85	18,4	34,40	16,0	31,10	22,3	32,17	19,1
$D_{Ensemble}$ [26]	24,83	14,9	37,93	18,4	22,46	**15,2**	34,28	16,6	33,15	22,1	30,53	17,44
SafeUAVNet-Small [26]	26,34	16,7	46,11	23,7	26,76	19,5	41,30	19,8	32,76	23,8	34,65	20,7
DPT [31]	34,33	22,8	**23,87**	**13,9**	26,36	20,1	30,48	14,8	28,57	26,8	28,72	19,7
BMD [32]	42,09	33,1	46,83	36,5	33,75	31,4	80,44	44,1	38,83	41,4	48,4	37,3
PackNet [33]	34,36	21,4	43,82	25,4	31,34	22,9	42,64	20,1	33,41	25,2	37,11	23,0
UFODepth-RGB	**21,52**	**14,4**	49,90	27,6	25,28	18,5	32,52	16,2	30,80	23,0	32,0	19,9
UFODepth	22,36	14,9	33,56	17,0	**21,98**	15,4	**26,45**	**13,0**	**26,73**	**19,4**	**26,21**	**15,9**

Capitolul 4

EXPLOATAREA CONSENSULUI TEMPORAL PENTRU SEGMENTAREA SEMANTICĂ A SCENEI

În contextul segmentării semantice a scenei în video, este nepractică adnotarea manuală a fiecărui cadru în mod independent, mai ales având în vedere că există relativ puține schimbări de la un cadru la altul și o rată mare de cadre capture. Prin urmare, capacitatea de a efectua o adnotare automată a întregii scene ar fi extrem de valoroasă. În această direcție, introducem *SegProp*, o nouă metodă iterativă bazată pe fluxul optic adnotând dens fiecare pixel al fiecărui cadru pe baza conexiunilor adiacente în timp. SegProp folosește cadre adnotate rar, alături de mișcarea în timp modelată prin lanțuri de flux optic și propagă etichete semantice înspre cadrele intermediare neadnotate. Această metodă exploatează coerența spațială și temporală în videoclipuri pentru a îmbunătăți propagarea adnotărilor. Astfel, SegProp reduce semnificativ efortul de adnotare în video fără a sacrifica calitatea adnotărilor generate.

*Acest capitol se bazează pe lucrarea – **Alina Marcu**, Vlad Licaret, Dragos Costea, and Marius Leordeanu. "Semantics through time: Semi-supervised segmentation of aerial videos with iterative label propagation." In Proceedings of the Asian Conference on Computer Vision. 2020. [34]*

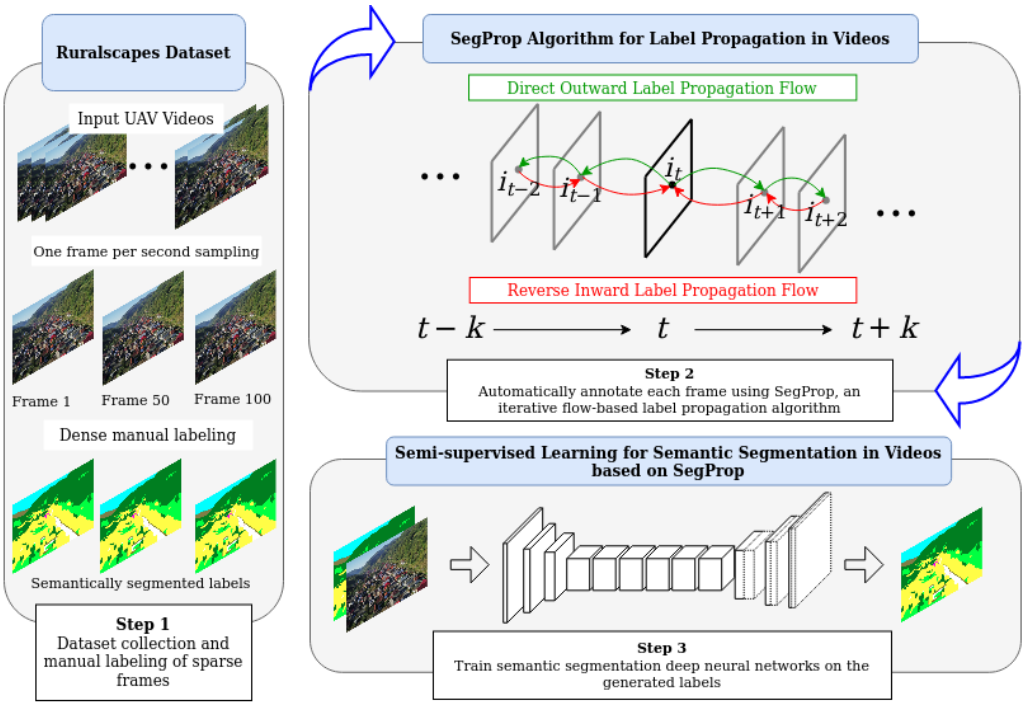


FIGURA 4.1: **SegProp** – metoda propusă pentru propagarea automată a adnotărilor semantice în contextul segmentării semi-supervizate în videoclipuri aeriene. Fiecare pas descrie contribuțiile pe care le aducem (1) introducem setul de date Ruralscapes, (2) propunem metoda de propagare a adnotărilor sporadice SegProp și (3) aratăm eficiența în scenarii de învățare semi-supervizată.

Una dintre motivațiile din spatele dezvoltării SegProp este raritatea seturilor de date video aeriene reale din perspectiva VAFP. Seturile de date existente adesea duc lipsă de adnotări dense necesare pentru antrenarea unor modele de segmentare semantică robuste. Pentru a umple acest gol, introducem *Ruralscapes*, un nou set de date care cuprinde imagini de înaltă rezoluție (4K) cu adnotări manuale dense la fiecare 50 de cadre, fiind cel mai mare set de date disponibil pentru problema segmentării semantice a scenei din zboruri reale aeriene. Prezentăm contribuțiile noastre principale în Figura 4.1. În **Pasul 1** eșantionăm videoclipurile, la intervale regulate (de ex. un cadru, la fiecare secundă). Cadrele rezultate sunt apoi etichetate manual (*Ruralscapes*). În **Pasul 2** propagăm automat etichetele către cadrele neetichetate rămase folosind algoritmul nostru *SegProp* - bazat pe votul clasei, la nivel de pixel, conform fluxurilor de propagare a etichetelor înspre și dinspre cadrul curent și un cadru adnotat. Fluxurile de propagare ar putea fi bazate pe flux optic (implicit), transformare omografică sau altă metodă de propagare, așa cum se arată în experimentele noastre (Tabelul 4.1). SegProp propagă iterativ votul de segmentare a clasei până la convergență, îmbunătățind performanța peste iterații. În **Pasul 3** amestecăm toate adnotările generate cu adnotările manuale de referință pentru

a antrena rețele neurale convoluționale adânci de ultimă generație pentru segmentare semantică și îmbunătățim semnificativ performanța în cazuri video-urilor nevăzute.

TABELA 4.1: Rezultatele propagării automate a adnotărilor. Zhu et al. [35] (care are doar 1 iterație) vs. SegProp pornind fie de la [35], fie de la cadrele de referință algoritmul nostru cu îmbunătățiri constante peste mai multe iterații.

Metode \ Iterații		1	2	3	4	5	6	7	+ Filt.
Zhu et al. [35]	mF1	.846	-	-	-	-	-	-	-
	mIOU	.747	-	-	-	-	-	-	-
SegProp de la [35]	mF1	.846	.874	.877	.885	.888	.891	.893	.896
	mIOU	.747	.785	.790	.801	.805	.810	.813	.818
SegProp	mF1	.884	.894	.896	**.897**	**.897**	**.897**	**.897**	**.903**
	mIOU	.801	.817	.819	**.821**	**.821**	**.821**	**.821**	**.829**

SegProp demonstrează o performanță remarcabilă în adnotarea automată a celor 98% din cadrele din setul de date Ruralscapes, atingând o acuratețe care depășește 90% în F1-score. Această acuratețe depășește semnificativ pe cea a metodelor existente de propagare a etichetelor de ultimă generație (Zhu et al. [35]). Mai mult, modularitatea SegProp permite integrarea altor metode în cadrul buclei sale iterative de propagare a etichetelor, rezultând într-o creștere suplimentară a performanței față de etichetele de start sau de referință (Tabelul 4.1). Pe lângă capacitățile sale de propagare a etichetelor, SegProp este testat într-un cadru de învățare semi-supervizată (Tabelul 4.2). Aici, antrenăm mai multe arhitecturi cunoscute de segmentare semantică pe cadrele etichetate automat de SegProp și evaluăm performanța lor pe videoclipuri noi, nevăzute. Rezultatele arată în mod consistent o îmbunătățire substanțială față de modelele antrenate într-o manieră pur supervizată. Acest lucru evidențiază potențialul SegProp de a îmbunătăți eficiența antrenării CNN-urilor pentru segmentare semantică cu date adnotate limitate.

TABELA 4.2: Rezultate cantitative după antrenarea rețelelor neuronale pe etichetele generate. Raportăm media F1-score peste toate videoclipurile din setul de testare, pentru fiecare clasă: (1) - teren, (2) - pădure, (3) - rezidențial, (4) - căpiță, (5) - drum, (6) - biserică, (7) - mașină, (8) - apă, (9) - cer, (10) - deal, (11) - persoană, (12) - gard și media peste toate clasele. Cele mai bune rezultate pentru fiecare clasă și fiecare model antrenat sunt îngroșate. Rezultatele arată clar o creștere semnificativă a performanței față de baza de referință, atunci când se antrenează cu SegProp (SP). Marcăm cu ✗cazul pur supervizat și cu ✓cel în care augmentăm etichetele cu SP și antrenăm cu amestecul.

Metode	SP	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	Toate
Unet	✗	.681	.497	.834	.000	.000	.000	.000	.000	**.967**	.000	.000	.000	.248
[10]	✓	**.757**	**.544**	**.838**	.000	**.556**	**.672**	.000	.000	.900	**.454**	.000	.000	**.393**
DeepLab	✗	.500	.416	.745	.000	.220	.073	.000	.000	.909	.242	.000	.000	.259
v3+ [25]	✓	**.570**	**.452**	**.776**	**.022**	**.369**	**.122**	**.007**	.000	**.926**	**.272**	**.004**	**.043**	**.297**
SafeUAV	✗	.713	.475	.757	.000	.371	.640	.000	.000	.953	**.260**	.000	.003	.348
Net [23]	✓	**.783**	**.488**	**.836**	**.364**	**.552**	**.748**	**.031**	**.428**	**.973**	.176	**.481**	**.610**	**.515**

Capitolul 5

GRAFURI/HIPERGRAFURI DE ÎNVĂŢARE A CONSENSULUI MULTIPLELOR SARCINI

Eforturile noastre de până acum s-au concentrat pe îmbunătățirea unei singure sarcini la un moment dat, dar obiectivul nostru final este să rezolvăm mai multe simultan. Motivația noastră provine din înțelegerea faptului că lumea reală nu se limitează la a fi văzută prin prisma imaginilor RGB; în schimb, avem diverse metode de a interpreta ceea ce vedem, și având la dispoziție mai mulți senzori, are sens să îi folosim pentru a ne îmbunătăți percepția și înțelegerea. Lumea noastră este interconectată, cu unele sarcini strâns legate, și ne propunem să utilizăm informații similare din surse diferite. Acest lucru ne-a condus la următoarele întrebări cheie: Cum putem valorifica eficient aceste interconexiuni? Cum putem folosi multiple indicii vizuale pentru a prezice altele mai complexe care nu pot fi derivate ușor? Putem dezvolta un cadru pentru integrarea diverselor surse de date și reprezentări, pentru a crea o înțelegere holistică a scenei, fără a fi nevoie de efort uman suplimentar? În acest capitol, prezentăm eforturile noastre în a oferi răspunsuri acestor întrebări deschise.

5.1 Înțelegerea Scenei prin Multiple Sarcini și Formarea Consensului Într-un Graf Neural

Abordăm problema învățării mai multor probleme vizuale ale lumii prin învățarea consensului într-un graf de rețele neurale. Ca răspuns la întrebările deschise menționate anterior, propunem Neural Graph Consensus (NGC) care integrează diverse interpretări ale scenelor dinamice într-un graf neural unificat, cuprinzând structura 3D, poziția, mișcarea și segmentarea semantică a obiectelor și activităților în spațiu și timp. Fiecare nod al grafului este o reprezentare a scenei, în timp ce fiecare muchie este o rețea adâncă care transformă un strat dintr-un nod într-altul dintr-un nod diferit. În cadrul acestui graf, mai multe căi care converg către un nod acționează ca profesori colectivi prin acorduri consensuale, ghidând rețelele individuale de pe muchii către același nod. Această abordare de antrenare auto-supervizată se dovedește a fi eficientă chiar și în contextul învățării nesupervizate, cu date neetichetate.

Folosim o procedură de inițializare pentru graf prin antrenarea fiecărei muchii în mod independent, într-o manieră supervizată. Ulterior, muchiile sunt antrenate folosind pseudo-adnotări generate din consensul multiplelor căi care ajung la un anumit nod. Utilizăm paradigma bine-cunoscută ”Profesor-Student” pentru învățare continuă. Căile direcționate către un nod funcționează ca un ansamblu ”Profesor” pentru fiecare muchie, oferind semnale de supraveghere cu confidență ridicată atunci când există consens. Procesul este repetat pe parcursul mai multor iterații, în care fiecare muchie devine un ”Student” și, de asemenea, parte a unui ansamblu diferit ”Profesor” pentru antrenarea altor studenți. Prin optimizarea consensului dintre diferite căi, graful atinge consistență și robustețe, chiar și în absența datelor etichetate.

Analiza experimentală este efectuată pe un set de date sintetic, dar realist, care replică îndeaproape zborurile reale ale VAFP extrase din simulatorul CARLA [37] (Figura 5.1).

*Această secțiune se bazează pe lucrarea – Marius Leordeanu, Mihai Pirvu, Dragos Costea, **Alina Marcu**, Emil Slusanschi, and Rahul Sukthankar. ”Semi-supervised learning for multi-task scene understanding by neural graph consensus.” In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 3, pp. 1882-1892. 2021. [36]*

Acest simulator produce date de înaltă calitate, cu multiple reprezentări și permite validarea corespunzătoare a metodei propuse. Căile simulează o traiectorie a dronei, cu mici variații aleatorii în toate unghiurile. În timp ce traiectoria de antrenare este un zbor tradițional urmând o structură ”grid”, pentru testare traiectoria urmărește să capteze cât mai multe puncte de vedere posibile, pentru a crește complexitatea.

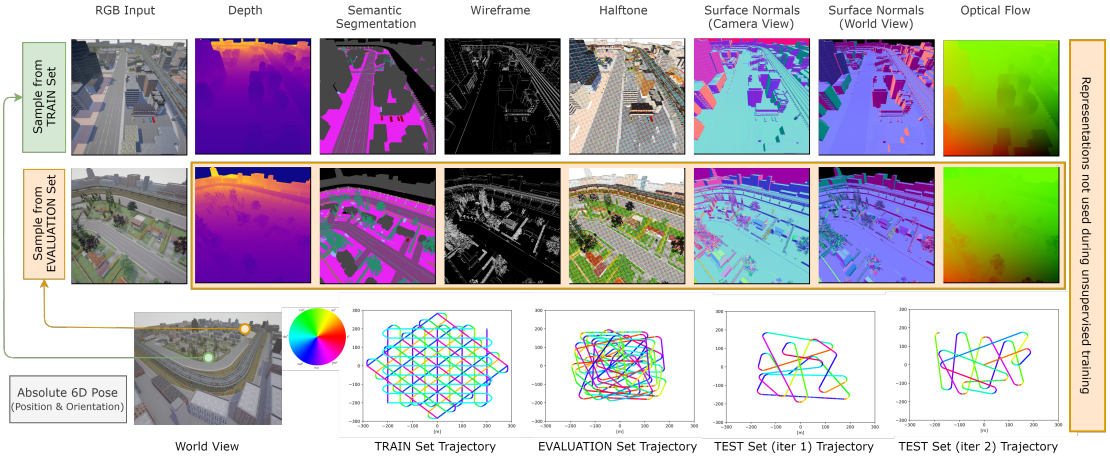


FIGURA 5.1: Exemple din setul de date sintetic (antrenare, primul rând, evaluare, al doilea rând) pe care l-am colectat folosind mediul virtual CARLA.

Analiza noastră experimentală (Tabelul 5.1), bazată pe o structură de graf pre-configurată, ceea ce înseamnă că am aplicat o abordare sub-optimală pentru a determina cel mai bun set de muchii pentru fiecare problemă abordată, a demonstrat eficacitatea metodei în îmbunătățirea rezultatelor pentru șase sarcini complexe pe parcursul a două iterații de învățare, nu doar la nivelul ansamblului mediu NGC, ci și la nivelul unei singure muchii cu doar $\approx$ 1.1M parametri [23] pentru fiecare.

TABELA 5.1: Rezultate cantitative pentru metoda NGC propusă pe 6 reprezentări, pe parcursul a 2 iterații de învățare nesupervizată. Arătăm cele mai bune rezultate pentru ansamblurile de profesori NGC (**îngroșate**) și studenții de pe o singură muchie (**îngroșate**). Se pot observa îmbunătățirile iterative constante.

Reprezentare	Metric de Evaluare	Iterația 0	Iterația 1		Iterația 2	
		EdgeNet	NGC	Distil. EdgeNet	NGC	Distil. EdgeNet
Adâncime	L1 ↓ (metri)	4.98	3.48	4.28	**3.29**	**3.95**
	Pixeli ↑ (%)	-	79.30	60.66	**79.69**	**61.90**
Normale de Suprafață (C)	L1 ↓ (grade)	8.48	7.79	8.28	**7.45**	**7.67**
	Pixeli ↑ (%)	-	74.18	53.59	**74.61**	**53.94**
Normale de Suprafață (W)	L1 ↓ (grade)	11.88	8.82	10.75	**8.52**	**8.67**
	Pixeli ↑ (%)	-	79.95	57.88	**81.12**	**61.14**
Segmentare Semantică	Acuratețe ↑ (%)	90.01	91.81	90.19	**92.45**	**92.83**
	mIOU ↑ (%)	48.40	49.78	49.80	**52.58**	**51.59**
	Pixeli ↑ (%)	-	79.46	69.62	**81.49**	**71.95**
Wireframe	Acuratețe ↑ (%)	96.17	96.55	96.54	**96.61**	**96.55**
	Pixeli ↑ (%)	-	77.71	72.57	**78.02**	**73.46**
Poziție	L2 ↓ (metri)	25.75	15.53	20.02	**12.07**	**15.55**
Orientare	L1 ↓ (grade)	3.84	2.50	3.39	**2.20**	**3.00**

5.2 Hipergrafuri de Multiple Sarcini pentru Înțelegerea Aeriană

Bazându-ne pe conceptul NGC, introducem o structură nouă de hipergraf pentru învățarea de multiple sarcini și demonstrăm eficacitatea acesteia în scenarii mai dificile și reale cu și mai puțină supervizare umană decât înainte. Aplicăm modelul nostru în două domenii distincte (Figura 5.2): 1) scene complexe din lumea reală capturate în setul de date Dronescapes, pe care îl introducem, o colecție de video-uri din lumea reală înregistrată cu ajutorul VAFP și 2) setul de date NASA NEO [40], un set de date cu observații ale Pământului care acoperă 22 de ani. Setul de date Dronescapes, cu reprezentările sale multiple, este ideal pentru învățarea a multiple sarcini, în timp ce setul de date NASA NEO prezintă provocări precum date lipsă și schimbări de distribuție temporală.

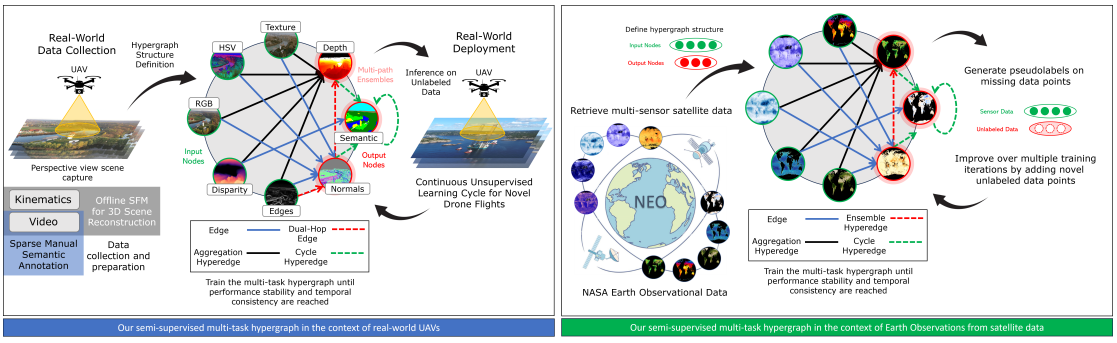


FIGURA 5.2: Prezentare generală a hipergrafului nostru pentru rezolvarea a multiple probleme vizuale într-un mod semi-supervizat pentru aplicații în multiple domenii, folosind zboruri VAFP din lumea reală sau observații ale Pământului.

Acest model extinde relațiile pereche din NGC, încorporând conexiuni de ordin superior, prin mai multe tipuri de hipermuchii (Figura 5.3) care captează interdependențe mai complexe. Similar cu NGC, în hipergraful nostru, fiecare nod este un nivel de interpretare a scenei. Unitățile de procesare de bază ale hipergrafului sunt **legături neurale directe (DNL)** care reprezintă muchia RGB→Sarcină. Legăturile neurale care conectează un nod de intrare la un nod de ieșire sunt muchii simple (E), în timp ce celelalte,

*Această secțiune se bazează pe lucrarea –**Alina Marcu**, Mihai Pirvu, Dragos Costea, Emanuela Haller, Emil Slusanschi, Ahmed Nabil Belbachir, Rahul Sukthankar, and Marius Leordeanu. "Self-supervised hypergraphs for learning multiple world interpretations." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 983-992. 2023. [38] și o mică parte din*
*Mihai Pirvu, **Alina Marcu**, Maria Alexandra Dobrescu, Ahmed Nabil Belbachir, and Marius Leordeanu. "Multi-Task Hypergraphs for Semi-supervised Learning using Earth Observations." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3404-3414. 2023. [39]*

care conectează mai multe noduri la un nod de ieșire formează hipermuchii complexe. Fiecare hipermuchie este modelată folosind o rețea neurală eficientă [23] (SafeUAVNet-Small). Lucrările anterioare folosesc doar muchii, în timp ce noi introducem diferite tipuri de hipermuchii pentru a capta relații mai complexe între straturi/noduri diferite. Având noduri de intrare: A, B, C, și ieșiri D și E, formăm patru tipuri de hipermuchii de complexități diferite: muchii pereche (E), muchii cu două salturi (DH-E), hipermuchii de tip ansamblu (EH), hipermuchii de agregare (AH) și hipermuchii ciclice (CH). Am demonstrat pe multiple sarcini că aceste hipermuchii depășesc cu mult performanța muchiilor simple (Figura 5.4 (Stânga)).

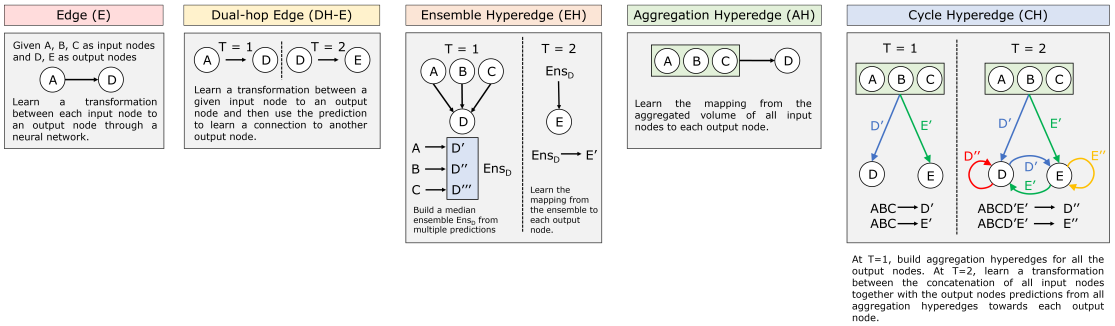


FIGURA 5.3: Diferența dintre diferite tipuri de muchii și hipermuchii în hipergraf.

	Type	Train Unlabeled (iter 2)			Train Unlabeled (iter 3)		
		(1)	(2)	(3)	(1)	(2)	(3)
Edges	E: rgb	42.85	5.04	10.37	32.79	21.66	12.40
	E: hsv	41.70	4.69	10.54	33.51	19.90	12.48
	E: softedges	32.47	6.26	11.56	27.28	18.61	13.53
	E: softseg	30.71	5.97	11.14	24.68	22.70	12.76
	E: ufo	20.77	7.19	11.69	16.93	17.55	12.89
	DH-E: sseg	-	6.25	11.39	-	19.00	12.93
	DH-E: depth	29.24	-	12.22	24.11	-	13.79
	DH-E: norm	30.56	6.17	-	26.35	21.15	-
	mean	32.61	5.94	11.27	26.52	**20.08**	12.97
Hyperedges	AH	41.80	5.33	10.37	33.63	23.96	12.24
	AH-ufo	41.96	5.16	10.78	33.82	21.10	12.72
	CH	44.63	4.93	10.32	36.92	20.36	12.23
	mean	**42.80**	**5.14**	**10.49**	**34.79**	21.81	**12.40**

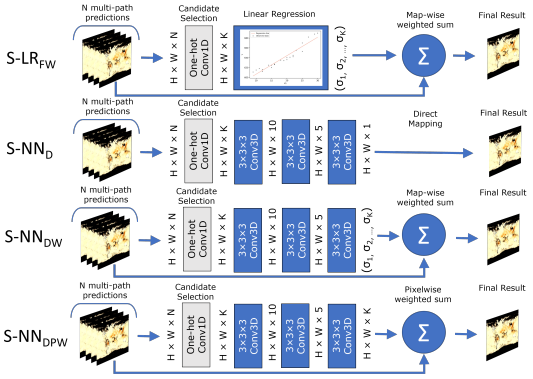


FIGURA 5.4: (**Stânga**) Evaluarea muchiilor și hipermuchiilor pe setul de date Dronescapes (Figura 5.5 (Stânga)), pentru multiple sarcini: 1 - segmentare semantică (sseg); 2 - estimarea adâncimii (depth); 3 - normale de suprafață (norm). Raportăm IoU mediu (% - valorile mai mari sunt mai bune (↑) pentru sarcina de segmentare semantică și eroarea L1 * 100 (mai mică este mai bine (↓)) pentru estimarea adâncimii și a normalelor. Rezultatele îngroșate evidențiază câștigul mediu de performanță al antrenării hipermuchiilor față de muchii. Semnificația seturilor de antrenare este în Figura 5.5 (Dreapta). (**Dreapta**) Arhitecturi propuse pentru ansamblurile învățate.

Mai multe căi pot ajunge la același nod pentru a forma ansambluri din care obținem pseudo-adnotări robuste și valorificăm puterea învățării semi-supervizate în hipergraf pe parcursul mai multor iterații adăugând date noi neetichetate. Diferit de metodele anterioare de formare a ansamblurilor (agregare prin medie, mediană sau bazată pe

distanță) introducem patru tipuri de ansambluri, toate cu un model inițial învățat de selecție a candidaților, care păstrează doar candidații relevanți înainte de a-i combina: S-LR$_{FW}$ învață o pondere fixă per candidat. S-NN$_{DW}$ produce dinamic o pondere per candidat, în funcție de intrare, în timp ce S-NN$_{DPW}$ produce dinamic o pondere pentru fiecare pixel al fiecărui candidat. În loc să combine liniar candidații, S-NN$_{D}$ învață o mapare non-lineară directă de la candidați la ieșire (Figura 5.4 (Dreapta)). Toate sunt învățate de la un capăt la altul.

TABELA 5.2: Comparație cu metodele anterioare bazate pe grafuri pentru rezolvarea mai multor sarcini – (**Stânga**) Pe setul de date Dronescapes, arătăm îmbunătățiri considerabile prin adăugarea hipermuchiilor propuse (notate cu HE în tabel) peste lucrările existente care folosesc doar muchii în structura lor de graf. (**Dreapta**) Pe setul de date NEO, evaluăm pe setul de test diferite modele ansamblu, pentru fiecare nod de ieșire (1) - *AOD*, (2) - *CM*, (3) - *FIRE*, (4) - *LAI*, (5) - $LSTD_{AN}$, (6) - $LSTN_{AN}$, (7) - *WV*. Cele mai bune numere sunt îngroșate, în timp ce cele subliniate sunt imediat următoarele.

Metodă	IoU(↑)			
	Barsana	Comana	Norvegia	Media
NGC [36] (Media)	41.53	40.75	27.38	36.55
NGC (Media) + HE	42.61	42.17	27.96	37.58
NGC [36] (Mediana)	39.25	37.41	27.01	34.56
NGC (Mediana) + HE	44.34	38.99	22.63	35.32
CShift [41] (Media)	43.91	42.13	29.68	38.57
CShift (Media) + HE	44.71	43.88	30.09	39.56
CShift [41] (Mediana)	43.30	40.62	29.51	37.81
CShift (Mediana) + HE	46.27	43.67	29.09	39.68
S-LR$_{FW}$	46.51	**45.59**	**30.17**	**40.76**
S-NN$_{D}$	45.53	42.92	28.37	38.94
S-NN$_{DW}$	45.48	43.25	26.36	38.36
S-NN$_{DPW}$	**48.21**	44.85	28.94	40.67

Tip Ans.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	ARPI (↑)
NGC [36]	0.44	12.97	12.15	5.44	-50.2	-39.0	2.21	-8.01
CShift [41]	1.86	13.07	8.66	6.64	-43.5	-28.0	2.65	-5.52
S-Media	3.31	14.71	**12.42**	8.97	-9.50	-2.48	5.93	4.77
S-LR$_{FW}$	4.85	11.15	8.62	8.45	0.21	4.90	5.84	6.29
S-NN$_{D}$	**6.67**	13.20	10.56	**9.68**	0.69	3.57	**7.86**	7.46
S-NN$_{DW}$	4.79	**14.80**	12.39	9.50	0.25	**4.91**	6.03	**7.52**
S-NN$_{DPW}$	5.74	6.51	11.21	8.26	1.33	5.62	4.98	6.24

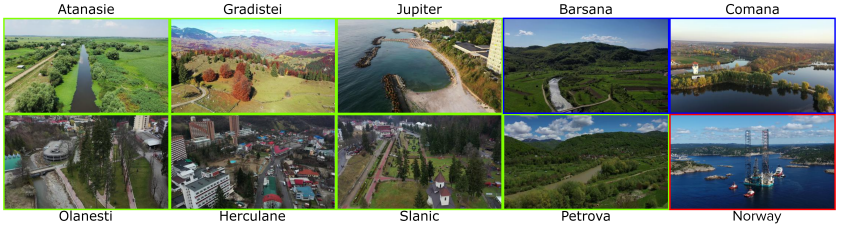
Evaluăm performanța ansamblurilor noastre învățate pe setul de date Dronescapes propus, pentru sarcina de segmentare semantică (Tabelul 5.2 (Stânga)) și pe toate nodurile de ieșire din setul de date NEO (Tabelul 5.2 (Dreapta)). Pe Dronescapes, aducem o îmbunătățire a performanței prin adăugarea hipermuchiilor și, de asemenea, prin permiterea ansamblurilor să învețe. Atât modelele NGC [36] cât și CShift [41] folosesc doar muchii și modele de ansamblu non-parametrice relativ simple la noduri (NGC - medie simplă și CShift - medie ponderată non-parametrică la nivel de pixel). Experimentele noastre arată că învățarea modelelor de ansamblu parametrice, chiar și a unuia linear simplu, îmbunătățește semnificativ (peste 2% în medie) rezultatele față de lucrările publicate anterior. Performanța este raportată pe Train Unlabeled (iter 3), care include doar scenele de test. Arătăm un tipar similar de îmbunătățire pe setul de date NEO, unde toate ansamblurile învățate cu selecție oferă o îmbunătățire relativă pozitivă comparativ cu metodele anterioare fără selecție (NGC și CShift).

TABELA 5.3: Performanța învățării iterative pe setul de date Dronescapes, la nivelul legăturilor neurale directe pentru fiecare sarcină. Evaluarea (raportată în medie) a fost făcută pe scenele de test.

	Tip	Semantic		Adâncime		Normale	
		IoU (↑)	Cons. (↑)	L1 (↓)	Cons. (↑)	L1 (↓)	Cons. (↑)
DNL	supervizat	25.04	88.85	-	-	-	-
	(iterația 1)	32.79	94.04	21.66	5.89	12.40	98.32
	(iterația 2)	37.26	95.72	17.34	7.06	11.93	98.87
	(iterația 3)	**40.31**	**98.13**	**16.64**	**30.26**	**11.71**	**99.30**

De asemenea, arătăm că valorificarea pseudo-adnotărilor generate din predicțiile ansamblului pe parcursul mai multor iterații de învățare îmbunătățește continuu performanța modelului, chiar și la nivelul celei mai simple muchii (DNL) (Tabelul 5.3) cu îmbunătățiri în termeni de acuratețe peste multiple sarcini și la nivelul consistenței temporale.

Introducem *Dronescapes*, un nou set de date video preluat de VAFP la scară largă cu scene diverse și adnotări generate automat pentru multiple sarcini (Figura 5.5 (Stânga)). Toate secvențele video includ informații GPS, viteze liniare și unghiulare și unghiuri absolute ale camerei (cu excepția scenei din afara distribuției din Norvegia). Lungimea totală a videoului este de aproximativ 50 de minute, cu cadre 4K și odometrie furnizată la 10 Hz. Colectăm un total de 10 scene foarte variate pe care le împărțim în 7 scene de antrenare și 3 scene de test (detalii despre împărțirea setului de date în Figura 5.5 (Dreapta)). Există o variație mare în distribuțiile spațiale ale claselor între diferitele scene Dronescapes, care variază de la rurale (Atanasie, Gradistei, Petrova, Barsana, Comana), la urbane (Olanesti, Herculane, Slanic) și de coastă (Jupiter, Norvegia), fiind în același timp și geografic depărtate.



Scene Name	Train Labeled	Train Unlabeled (iter 1)	Train Unlabeled (iter 2)	Train Unlabeled (iter 3)
Atanasie	76	4501 (76)	4500 (75)	-
Gradistei	18	726 (18)	484 (12)	-
Herculane	12	484 (12)	363 (9)	-
Jupiter	21	847 (21)	605 (15)	-
Olanesti	18	726 (18)	484 (12)	-
Petrova	12	484 (12)	363 (9)	-
Slanic	76	4501 (76)	4500 (75)	-
Barsana	-	-	-	1452 (36)
Comana	-	-	-	1210 (30)
Norway	-	-	-	2941 (50)
TOTAL	233	12269 (233)	11299 (207)	5603 (116)

FIGURA 5.5: (**Stânga**) Dronescapes – Scenele încadrate cu **verde** reprezintă scene de antrenare pentru care avem acces la o mică fracțiune de adnotări manuale în timpul antrenării. Celelalte reprezintă scene de test nevăzute cu distribuții semantice care sunt mai aproape de setul de antrenare (în **albastru**) sau în afara distribuției (**roșu**). (**Dreapta**) Împărțirea setului de date Dronescapes.

Metodele dezvoltate în această secțiune au implicații largi pentru diverse aplicații, de la navigația autonomă a VAFP la monitorizarea pe termen lung a mediului și studiile privind schimbările climatice.

Capitolul 6

CONCLUZII

Studiul nostru abordează provocările esențiale în înțelegerea scenelor aeriene, oferind atât soluții practice, cât și contribuții teoretice relevante. Conștientizăm că dificultățile legate de înțelegerea scenelor aeriene sunt departe de a fi complet rezolvate. Drept urmare, la final, propunem mai multe direcții pentru cercetările viitoare. Aceste direcții includ tranziția de la reprezentările 2D la cele 3D, valorificarea datelor sintetice și a tehnicilor de adaptare la domenii reale, integrarea unei game mai largi de senzori, îmbunătățirea înțelegerii semantice prin învățarea continuă de noi clase, explorarea învățării adaptive și continue, precum și exploatarea unor arhitecturi avansate.

De asemenea, în final adresăm considerații etice. Am respectat reglementările de zbor cu drone în România, prioritizând siguranța prin efectuarea zborurilor în zone slab populate și la altitudini mai mari. Preocupările legate de confidențialitate au fost atenuate atât prin obținerea consimțământului verbal, acolo unde a fost posibil, cât și informarea indivizilor despre scopul de cercetare al zborurilor. Încheiem cu reflecții asupra direcției generale de cercetare în IA, subliniind responsabilitatea experților de a modela percepția și înțelegerea publică a domeniului, accentuând importanța menținerii siguranței IA, desfășurării cercetării interdisciplinare și promovării discuțiilor informate pentru a construi încredere și a ajunge la un consens global privind IA ca instrument pentru progresul și în beneficiul umanității.

Bibliografie

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[2] Peter Corke. *Robotics, Vision and Control: fundamental algorithms in Python*, volume 146. Springer Nature, 2023.

[3] Alina Marcu and Marius Leordeanu. Object contra context: Dual local-global semantic segmentation in aerial images. *AAAI Workshops*, 2017.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Volodymyr Mnih. *Machine learning for aerial image labeling*. PhD thesis, University of Toronto (Canada), 2013.

[8] Shunta Saito and Yoshimitsu Aoki. Building and road detection from large aerial imagery. In *IS&T/SPIE Electronic Imaging*, pages 94050K–94050K. International Society for Optics and Photonics, 2015.

[9] Alina Marcu, Dragos Costea, Emil Slusanschi, and Marius Leordeanu. A multi-stage multi-task neural network for aerial scene interpretation and geolocalization. *arXiv preprint arXiv:1804.01322*, 2018.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[11] Christian Wiedemann, Christian Heipke, Helmut Mayer, and Olivier Jamet. Empirical evaluation of automatically extracted road axes. *Empirical Evaluation Techniques in Computer Vision*, pages 172–187, 1998.

[12] Shunta Saito, Takayoshi Yamashita, and Yoshimitsu Aoki. Multiple object extraction from aerial imagery with convolutional neural networks. *Electronic Imaging*, 2016(10):1–9, 2016.

[13] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, 2017.

[14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[15] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. *arXiv preprint arXiv:1709.00179*, 2017.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[17] Benjamin Bischke, Patrick Helber, Joachim Folz, Damian Borth, and Andreas Dengel. Multi-task learning for segmentation of building footprints with deep neural networks. *arXiv preprint arXiv:1709.05932*, 2017.

[18] Dragos Costea and Marius Leordeanu. Aerial image geolocalization from recognition and matching of roads and intersections. In *BMVC*, 2016.

[19] P.J. Besl and N.D. McKay. A method for registration of 3D shapes. *PAMI*, 14(2): 239–256, 1992.

[20] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017. Accessed: July, 2024.

[21] Dragos Costea, Alina Marcu, Emil Slusanschi, and Marius Leordeanu. Roadmap generation using a multi-stage ensemble of deep neural networks with smoothing-based optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[22] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *arXiv preprint arXiv:1805.06561*, 2018.

[23] Alina Marcu, Dragos Costea, Vlad Licaret, Mihai Pîrvu, Emil Slusanschi, and Marius Leordeanu. Safeuav: Learning to estimate depth and safe landing areas for uavs from synthetic data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[24] Google. Google Earth, 2018. URL `https://www.google.com/earth/`. Available at `https://www.google.com/earth/`, version 7.3.0.

[25] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[26] Mihai Pirvu, Victor Robu, Vlad Licaret, Dragos Costea, Alina Marcu, Emil Slusanschi, Rahul Sukthankar, and Marius Leordeanu. Depth distillation: unsupervised metric depth estimation for uavs by finding consensus between kinematics, optical flow and deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3215–3223, 2021.

[27] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL `http://arxiv.org/abs/1503.02531`.

[28] AliceVision. Meshroom: A 3D reconstruction software., 2018. URL `https://github.com/alicevision/meshroom`.

[29] Vlad Licăret, Victor Robu, Alina Marcu, Dragoş Costea, Emil Sluşanschi, Rahul Sukthankar, and Marius Leordeanu. Ufo depth: Unsupervised learning with flow-based odometry optimization for metric depth estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6526–6532. IEEE, 2022.

[30] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in Neural Information Processing Systems*, pages 35–45, 2019.

[31] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021.

[32] S. Mahdi, H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proc. CVPR*, 2021.

[33] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[34] Alina Marcu, Vlad Licaret, Dragos Costea, and Marius Leordeanu. Semantics through time: Semi-supervised segmentation of aerial videos with iterative label

propagation. In *Asian Conference on Computer Vision (ACCV), 2020*, pages 2881–2890, 2020.

[35] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019.

[36] Marius Leordeanu, Mihai Cristian Pîrvu, Dragos Costea, Alina E Marcu, Emil Slusanschi, and Rahul Sukthankar. Semi-supervised learning for multi-task scene understanding by neural graph consensus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1882–1892, 2021.

[37] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. *arXiv preprint arXiv:1711.03938*, 2017.

[38] Alina Marcu, Mihai Pirvu, Dragos Costea, Emanuela Haller, Emil Slusanschi, Ahmed Nabil Belbachir, Rahul Sukthankar, and Marius Leordeanu. Self-supervised hypergraphs for learning multiple world interpretations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 983–992, October 2023.

[39] Mihai Pirvu, Alina Marcu, Maria Alexandra Dobrescu, Ahmed Nabil Belbachir, and Marius Leordeanu. Multi-task hypergraphs for semi-supervised learning using earth observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3404–3414, October 2023.

[40] Earth Observations NASA. Nasa earth observations dataset, 2020. URL `https://neo.gsfc.nasa.gov/`. [Online; accessed: 10.06.2024].

[41] Emanuela Haller, Elena Burceanu, and Marius Leordeanu. Self-supervised learning in multi-task graphs through iterative consensus shift. *BMVC*, 2021.