Construirea lumii cu o dronă: crearea eficientă a scenelor aeriene prin localizare, recunoaștere și sinteză de perspective virtuale

Costea Dragos

1 Introducere

Această secțiune prezintă o privire de ansamblu a tezei, ce explorează potențialul navigației bazate pe viziune și înțelegerea scenei, concentrându-se pe analiza și reconstrucția imaginilor aeriene. Principalele întrebări abordate sunt:

- Putem naviga într-un mediu folosind doar input-ul camerei?
- Cât de bine trebuie să înțelegem mediul pentru o navigare eficientă?
- Cum putem valorifica și îmbunătăți datele existente ale hărților?
- Poate fi efectuată toată procesarea de bază pe o dronă?

Cercetarea este împărțită în patru domenii principale: Localizare, Cartografiere, Reconstrucția Scenei și Scene Interactive.

1.1 Localizare

Localizarea este crucială pentru navigarea autonomă, mai ales în mediile urbane unde semnalele GPS pot fi nesigure. Localizarea bazată pe viziune oferă o alternativă robustă, inspirându-se din capacitățile vizuale umane. Provocarea constă în dezvoltarea unor algoritmi eficienți care pot rula pe bugete de calcul limitate, tipice pentru hardware-ul dronelor.

Lucrările recente în localizarea imaginilor aeriene au arătat rezultate promițătoare. De exemplu, [1] a propus o metodă pentru geolocalizarea de la sol la aer prin învățarea unui spațiu de încorporare comun pentru imaginile terestre și aeriene. [2] a extins această idee la geolocalizarea imaginilor pe zone largi folosind imagini aeriene de referință.

1.2 Cartografiere

Bazându-se pe localizare, cartografierea robustă a drumurilor este esențială pentru navigare și poate oferi o bază pentru utilizarea intersecțiilor ca repere de localizare. Scopul este de a reduce decalajul dintre datele existente ale hărților și informațiile vizuale în timp real.

Primele abordări pentru detectarea drumurilor în imaginile aeriene se bazau pe caracteristici proiectate manual [3–5]. Tehnicile de învățare profundă au dus la îmbunătățiri semnificative, cu [6] și [7] printre primii care au aplicat rețele neuronale convoluționale (CNN-uri) la această sarcină.

Abordările mai recente s-au concentrat pe valorificarea contextului și a informațiilor multi-scală. [8] a propus o rețea cu două fluxuri care procesează atât contextul local, cât și cel global, demonstrând o eficacitate deosebită în gestionarea scenelor urbane complexe.

1.3 Reconstrucția Scenei, Navigare și Reprezentare

Acest domeniu se concentrează pe crearea de replici precise ale lumii pentru simulări avansate, conservarea patrimoniului și experiențe îmbunătățite ale utilizatorilor. Explorează sinergia dintre metodele geometrice și analitice pentru sinteza de noi vizualizări și investighează reprezentări eficiente pentru sisteme de navigare bazate pe viziune, cu costuri reduse.

Estimarea adâncimii din imagini monoculare a cunoscut progrese semnificative, determinate de avansurile în învățarea profundă. [9] a demonstrat fezabilitatea antrenării CNN-urilor pentru a prezice adâncimea din imagini singulare. Tehnicile de învățare nesupervizată și auto-supervizată, precum cele introduse de [10] și [11], au câștigat popularitate deoarece elimină necesitatea datelor costisitoare de adevăr de teren pentru adâncime.

Domeniul reconstrucției 3D a cunoscut o schimbare de paradigmă odată cu introducerea Neural Radiance Fields (NeRF) de către [12]. NeRF reprezintă scenele ca funcții volumetrice continue și a demonstrat rezultate impresionante în sinteza de noi vizualizări. Lucrări ulterioare, precum Instant-NGP [13], s-au concentrat pe îmbunătățirea eficienței și scalabilității.

Pentru reconstrucția aeriană la scară largă, tehnicile tradiționale de fotogrammetrie rămân larg utilizate. Pipeline-urile Structure-from-Motion (SfM), precum COLMAP [14], pot reconstrui scene 3D din colecții mari de imagini neordonate. Cu toate acestea, aceste metode se confruntă adesea cu dificultăți în ceea ce privește scara și complexitatea reconstrucțiilor la scara orașului.

1.4 Scene Interactive

Ultima zonă abordează provocările generării de vizualizări pentru replici imperfecte și integrarea personajelor virtuale în medii reconstruite. De asemenea, explorează potențialul artei generative în interacțiunea om-mașină, în special în transmiterea emoțiilor.

În contextul interacțiunii om-IA, a existat un interes crescând pentru generarea de răspunsuri emoționale adecvate de la agenți virtuali. [15] au explorat utilizarea modelelor computaționale de emoție pentru a controla comportamentul personajelor virtuale. [16] a investigat abordări bazate pe date pentru a genera expresii faciale realiste pentru agenți virtuali în timp real.

1.5 Motivație și Aplicații

Motivațiile din spatele acestei cercetări includ:

- Dezvoltarea de sisteme robuste de localizare bazate pe viziune pentru dispozitive aeriene pentru a îmbunătăți siguranța în zonele urbane.
- Demonstrarea fezabilității navigației bazate exclusiv pe viziune, bazându-se pe succesele în domenii de nișă precum cursele de drone [17] și mediile simulate [18].
- Permiterea actualizărilor automate la scară largă ale hărților prin combinarea localizării precise și a recunoașterii vizuale.
- Îmbunătățirea calității reconstrucției 3D prin exploatarea informațiilor complementare din metodele clasice și neurale.
- Crearea de noi vizualizări plauzibile pentru a umple golurile de informații în scenele reconstruite.
- Proiectarea de algoritmi eficienți potriviți pentru utilizare embeddedă, asigurând că agenții autonomi pot opera în siguranță chiar și în cazul unor întreruperi de rețea.

Lucrări Conexe Domeniul analizei imaginilor aeriene a evoluat de la abordările timpurii care foloseau caracteristici proiectate manual [3–5, 19, 20] la tehnicile moderne de învățare profundă. S-au făcut progrese semnificative în segmentarea semantică, cu metode precum cele propuse de [8] și [21] îmbunătățind acuratețea în scene urbane complexe.

Cercetarea în geolocalizare a progresat de la metodele tradiționale de potrivire a caracteristicilor [22, 23] la abordările de învățare profundă care învață reprezentări robuste [1, 2]. În contextul navigației UAV, localizarea bazată pe viziune a fost explorată ca o alternativă sau complement la GPS [24].

Estimarea adâncimii și reconstrucția 3D au cunoscut progrese semnificative, de la abordările timpurii bazate pe învățare [9] la tehnicile mai recente nesupervizate [10, 11]. Introducerea Neural Radiance Fields (NeRF) [12] a revoluționat sinteza de noi vizualizări, cu lucrări ulterioare concentrându-se pe îmbunătățirea eficienței și scalabilității [13].

Estimarea zonelor sigure de aterizare pentru UAV-uri a evoluat de la metode bazate pe caracteristici proiectate manual [25] la abordări de învățare profundă [26]. Lucrări recente au explorat utilizarea datelor sintetice pentru a aborda disponibilitatea limitată a datelor de antrenament etichetate [27].

În domeniul interacțiunii om-IA, cercetarea a progresat de la abordările tradiționale de recunoaștere a emoțiilor [28] la metode de învățare profundă [29]. Lucrările privind generarea de răspunsuri emoționale adecvate de la agenți virtuali [15] și generarea de expresii faciale în timp real [16] au deschis noi căi pentru interfețe naturale și intuitive.

1.6 Concluzie

Această cercetare se află la intersecția dintre viziunea computerizată, robotică și interacțiunea om-calculator. Se bazează pe fundații existente, introducând în același timp tehnici noi pentru a avansa stadiul actual al tehnologiei în aceste domenii interconectate. Lucrarea are implicații potențiale pentru navigarea autonomă, planificarea urbană, modelarea 3D și dezvoltarea unor sisteme AI mai intuitive.

Prin abordarea acestor provocări, cercetarea își propune să contribuie la dezvoltarea unor sisteme autonome mai robuste, eficiente și inteligente, capabile să înțeleagă și să interacționeze cu medii complexe din lumea reală. Natura interdisciplinară a lucrării evidențiază potențialul pentru polenizarea încrucișată a ideilor între diferite subdomenii ale viziunii computerizate și roboticii.

2 Localizare din drumuri și intersecții în imagini aeriene

2.1 Introducere

Analiza imaginilor aeriene are aplicații importante în cartografierea automată, planificarea urbană, monitorizarea mediului și ajutorul în caz de dezastre. Acest capitol abordează sarcina de geolocalizare automată a imaginilor aeriene prin recunoașterea și potrivirea drumurilor și intersecțiilor. Propunem un pipeline nou pentru geolocalizare, de la detectarea drumurilor și intersecțiilor până la identificarea regiunii geografice prin potrivirea intersecțiilor detectate cu cele etichetate manual din OpenStreetMap (OSM). Aceasta este urmată de alinierea geometrică între drumurile detectate și adnotările OSM. Testăm pe un set de date de imagini aeriene din două orașe europene și folosim OSM pentru adnotări de drumuri ca adevăr de referință. Experimentele arată o localizare precisă atunci când se antrenează pe un oraș și se testează pe celălalt, chiar și cu imagini aeriene de calitate relativ slabă. De asemenea, demonstrăm că alinierea între drumurile detectate și adnotările OSM poate îmbunătăți calitatea detectării drumurilor.

2.2 Lucrări conexe

Detectarea drumurilor în imagini aeriene a fost abordată folosind caracteristici proiectate manual [3, 4] și mai recent rețele neuronale convoluționale [6, 7]. Unele metode încearcă să corecteze vectorii de drumuri nealiniați prin alinierea lor cu imagini aeriene [30]. Geolocalizarea pentru UAV-uri folosind caracteristici rare proiectate manual a fost propusă [24]. Alte abordări fuzionează intrarea camerei cu date GPS și IMU [31, 32]. Geolocalizarea imaginilor de la sol folosind perechi de imagini aeriene a fost, de asemenea, explorată [1, 2].

2.3 Abordare

Metoda noastră are mai multe etape:

Clasificarea pixel cu pixel a drumurilor folosind o rețea CNN cu flux dual local-global [33]. Detectarea intersecțiilor pe baza drumurilor detectate. Potrivirea intersecțiilor detectate cu un set de date stocat de intersecții OSM folosind descriptori învățați. Aliniere geometrică pentru îmbunătățirea localizării și îmbunătățirea detectării drumurilor.

Folosim intersecțiile ca ancore pentru localizare, deoarece sunt rare, eficiente din punct de vedere computațional și tind să aibă modele unice de drumuri înconjurătoare utile pentru recunoaștere.

2.3.1 Detectarea drumurilor și intersecțiilor

Pentru detectarea drumurilor, folosim o rețea CNN de ultimă generație cu flux dual local-global [33] care combină aspectul local și informațiile contextuale mai largi. Pentru detectarea intersecțiilor, antrenăm un AlexNet ajustat care ia ca intrare imaginea RGB și harta estimată a drumurilor.

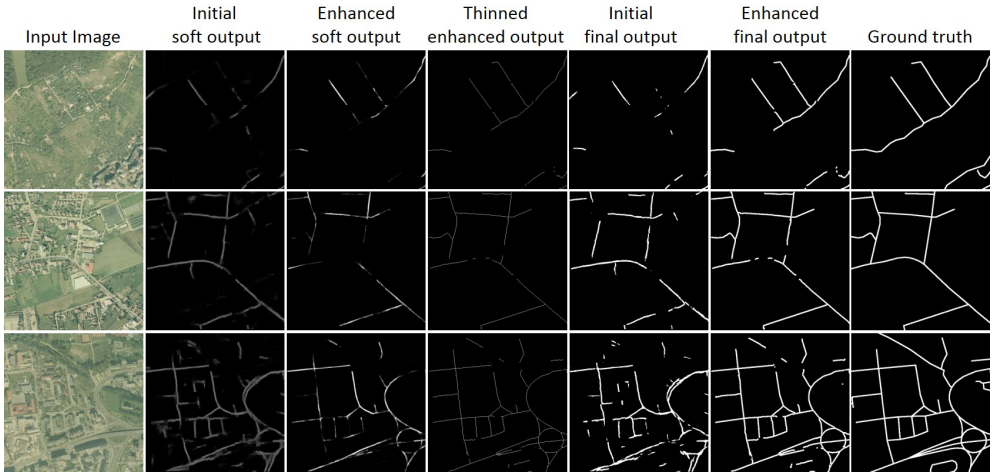

Figura 1: Îmbunătățirea detectării drumurilor prin recunoașterea regiunii și alinierea geometrică cu drumurile OSM. Procedura noastră îmbunătățește hărțile de drumuri detectate și ar putea corecta etichetele OSM.

Tabela 1: Erori de localizare înainte și după aliniere (în metri)

Metoda	Înainte		După	
	Media	Mediana	Media	Mediana
1NN pix2pix	470,50	53,90	57,97	1,60
1NN MSMT Stage 1	34,11	31,65	454,79	1,40
MSMT cu LocDecoder-R-2	88,92	53,90	57,97	1,60
MSMT cu LocDecoder-S-128	**9,03**	**6,85**	**1,89**	**0,75**
MSMT cu LocCombined	26,93	7,27	18,42	0,78

2.3.2 Potrivirea intersecțiilor și localizarea

Reprezentăm fiecare intersecție printr-un descriptor învățat astfel încât intersecțiile identice din drumurile detectate și OSM să aibă descriptori similari. Ajustăm fin rețeaua de detectare a intersecțiilor pentru a ajusta distanțele în spațiul descriptorilor și a îmbunătăți performanța potrivirii. Localizarea este rafinată prin alinierea geometrică între drumurile estimate și drumurile OSM în regiunile centrate la intersecțiile potrivite. Folosim o abordare de potrivire a grafului bipartit pentru a găsi corespondențe între intersecțiile detectate și OSM.

2.3.3 Aliniere geometrică și îmbunătățirea drumurilor

Folosim algoritmul Iterative Closest Point pentru a alinia segmentarea drumului cu drumurile OSM la locația prezisă. Am dezvoltat o versiune simplificată care estimează doar translația, presupunând că imaginile sunt aliniate cu punctele cardinale. Pentru a îmbunătăți detectarea drumurilor, aplicăm o dilatare ușoară pe harta estimată a drumurilor, o înmulțim cu harta OSM aliniată, netezim cu un filtru gaussian și subțiem folosind supresie non-maximă.

2.4 Experimente

Am colectat imagini aeriene din două orașe europene (A și B) aliniate cu hărți rutiere OSM. Orașul A are 4027 de imagini de 512x512 pixeli folosite pentru antrenament, în timp ce orașul B are 3177 de imagini folosite pentru testare. Rezoluția spațială este de 1m/pixel. Pentru detectarea drumurilor și intersecțiilor, obținem o măsură F de 82,22% pe European Roads Dataset [34] și 99,88% pe setul nostru de date de localizare cu relaxare de 3 pixeli. Pentru localizare, 96,84% din locațiile de test au o eroare <20m fără aliniere. După aliniere, 94,56% sunt în limita a 2,5m și 97,58% în limita a 5m față de adevărul de referință, comparabil cu precizia GPS comercială [35].

2.5 Concluzii

Am prezentat un sistem complet pentru geolocalizare din imagini aeriene fără GPS. Pipeline-ul nostru include metode eficiente pentru detectarea drumurilor și intersecțiilor, recunoașterea intersecțiilor cu aliniere geometrică pentru localizare precisă și îmbunătățirea detectării drumurilor. Abordarea ar putea fi folosită ca o alternativă la GPS sau în conjuncție cu GPS pentru aplicații care necesită procesare offline sau în timp real. Lucrările viitoare s-ar putea concentra pe îmbunătățirea vitezei de detectare, extinderea spațiului de căutare la mai multe orașe și adaptarea pipeline-ului pentru utilizare pe timp de noapte.

3 Detectarea drumurilor și clădirilor în imagini aeriene

3.1 Introducere

Recunoașterea drumurilor și intersecțiilor în imagini aeriene este o problemă dificilă în domeniul viziunii computerizate, cu aplicații în localizarea și navigarea UAV-urilor. În timp ce abordările recente de învățare profundă au îmbunătățit segmentarea la nivel de pixel, noi susținem că drumurile ar trebui recunoscute la nivelul semantic superior al grafurilor rutiere. Prezentăm o metodă în două etape: 1) Detectarea drumurilor și intersecțiilor cu o nouă rețea generativă adversarială cu două salturi (DH-GAN) pentru segmentarea la nivel de pixel. 2) Găsirea celui mai bun graf rutier de acoperire folosind optimizarea bazată pe netezire (SBO). De asemenea, prezentăm o îmbunătățire multi-etapă a acestui proces.

3.2 Detectarea drumurilor cu DH-GAN și SBO

Arhitectura noastră DH-GAN constă din două GAN-uri condiționale: primul generează segmentări de drumuri, în timp ce al doilea generează intersecții folosind atât intrarea RGB originală, cât și segmentarea drumului de la primul GAN (Figura 2). Arhitectura completă este antrenată end-to-end. Reprezentăm hărțile rutiere ca grafuri $G = (V, E)$, unde V sunt noduri cu poziții (x_i, y_i) și E sunt

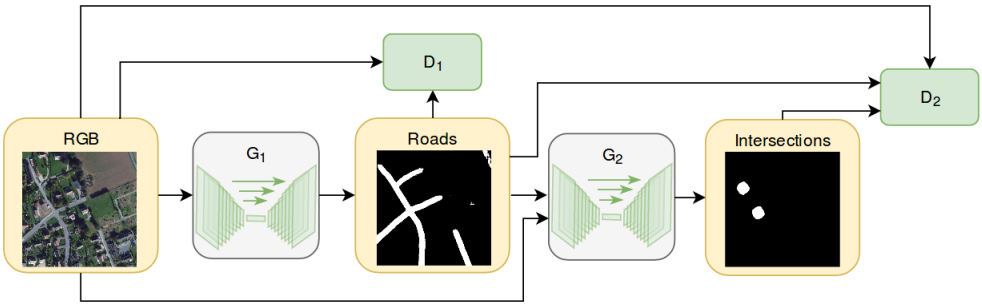

Figura 2: Arhitectura propusă DH-GAN.

muchii cu segmente de linie asociate. Definim costul muchiei $c(i, j)$ ca distanța medie până la cel mai apropiat punct de drum în segmentare, și scorul general al grafului $S(V)$ ca intersecția peste uniune între graful dilatat și harta la nivel de pixel. Abordarea noastră SBO optimizează locațiile nodurilor pentru a maximiza $S(V)$, începând de la intersecțiile găsite de DH-GAN și adăugând iterativ puncte mijlocii. Aplicăm, de asemenea, o bază de referință de eșantionare greedy pentru comparație. Evaluăm pe European Road Dataset [33], cu 200 de imagini de antrenare, 20 de validare și 50 de testare. Tabelele 2-3 arată rezultatele noastre în comparație cu bazele de referință. DH-GAN depășește alte metode în acuratețea la nivel de pixel. Abordările bazate pe grafuri (DH-GAN+Greedy și DH-GAN+SBO) sacrifică o parte din acuratețe pentru economii semnificative de stocare (Tabelul 4). Rezultatele calitative sunt prezentate în Figura 3.

3.3 Abordarea ansamblului multi-etapă

Propunem o metodă în trei etape pentru extragerea drumurilor:

Metodă	F-score
GAN [36]	77,70%
LG-Seg-ResNet-IL [37]	81,06%
U-net [38]	79,79%
DH-GAN	**84,05%**
DH-GAN + Greedy	80,81%
DH-GAN + SBO	81,74%

Tabela 2: Rezultatele detectării drumurilor. Mai mare este mai bine.

Tip de etichetare	Metodă	F-score
OSM	GAN [36]	54,89%
	DH-GAN	**63,01%**
	DH-GAN + Greedy	31,81%
	DH-GAN + SBO	59,79%
Independent	GAN [36]	64,42%
	DH-GAN	82,65%
	DH-GAN + Greedy	43,42%
	DH-GAN + SBO	**86,00%**

Tabela 3: Rezultatele detectării intersecțiilor. Mai mare este mai bine.

Metodă	Vârfuri	Muchii
LG-Seg-ResNet-IL [37]	782*	-
DH-GAN	1345*	-
DH-GAN + Greedy	23	21
DH-GAN + SBO	**19**	**17**
OSM (ground truth)	29	31

Tabela 4: Cost mediu de stocare. Mai mic este mai bine. *Vârfuri obținute prin subțierea drumurilor la o lățime de un singur pixel.

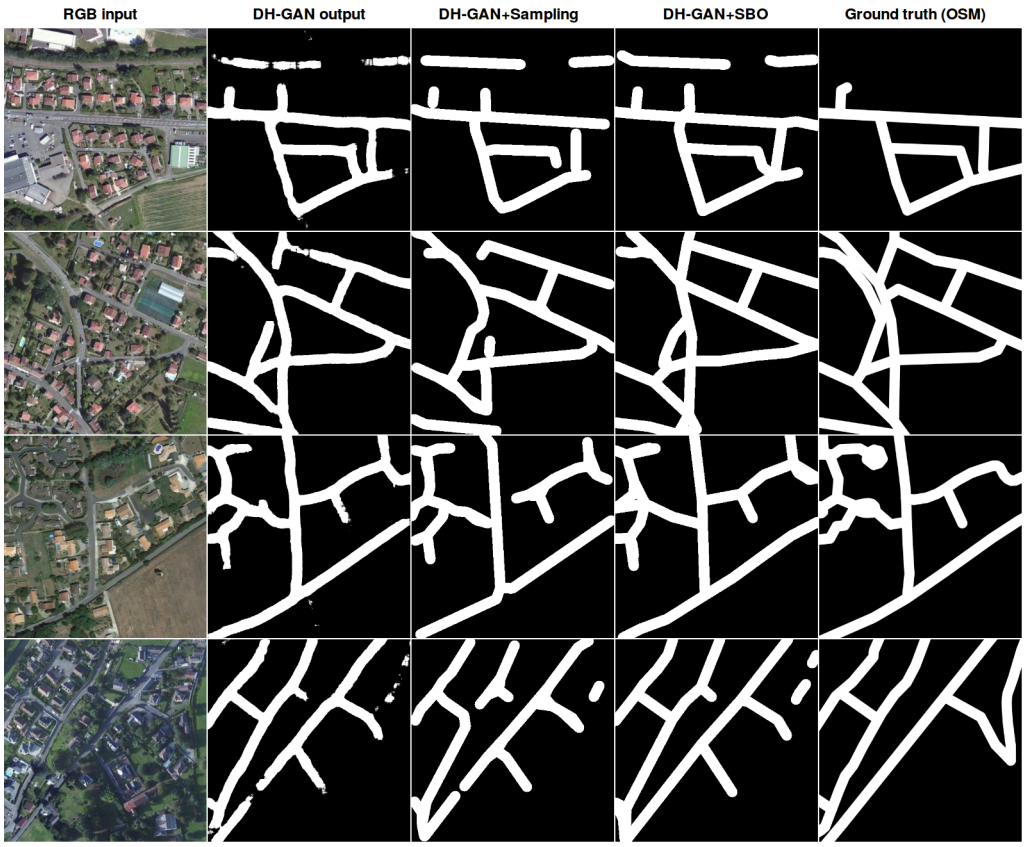

Figura 3: Rezultate calitative pentru detectarea drumurilor și generarea hărților folosind grafuri.

Tabela 5: Rezultatele segmentării drumurilor pe setul nostru de antrenare/validare.

Model	Antrenare		Validare	
	IoU	F1	IoU	F1
Dilatare max 32	0,6432	0,7824	0,6483	0,7883
Dilatare max 48	0,6577	0,7913	0,6601	0,7957
Dilatare max 64	**0,6591**	**0,7919**	**0,6640**	**0,7966**

Antrenarea mai multor rețele de tip U-net cu rate de dilatare diferite pentru segmentarea drumurilor și intersecțiilor. Combinarea predicțiilor parțiale cu intrarea RGB într-o altă rețea pentru o segmentare îmbunătățită. Generarea vectorilor de drum folosind SBO. Adăugarea legăturilor lipsă folosind atât segmentarea, cât și vectorii de drum.

Folosim setul de date DeepGlobe [39] cu 6226 de imagini de antrenare, 1243 de validare și 1101 de testare. Variantele noastre U-net folosesc convoluții dilatate înlănțuite cu rate maxime de dilatare diferite (32, 48, 64). De asemenea, antrenăm o rețea pe drumuri cu lățime constantă. Rezultatele sunt prezentate în Tabelele 5-**??**. Am constatat că antrenarea pe drumuri mai groase a îmbunătățit performanța (Tabelul 6).

3.4 Concluzii și Lucrări Viitoare

Am prezentat două abordări pentru detectarea drumurilor în imagini aeriene:

DH-GAN cu SBO pentru extragerea eficientă a grafurilor rutiere, combinând învățarea profundă și optimizarea grafurilor. O abordare ansamblă multi-etapă folosind rate de dilatare multiple și rafinarea vectorilor de drum.

Tabela 6: Studiu privind grosimea drumurilor folosind modelul cu dilatare maximă 32.

		IoU Antrenare	IoU Validare
Aceeași lățime	Subțire (≈ 4m)	0,6282	0,6123
	Gros (2x subțire)	0,6918	0,6751
Lățime variabilă	Subțire (original)	0,6432	0,6483
	Gros (2x subțire)	0,7254	0,6889

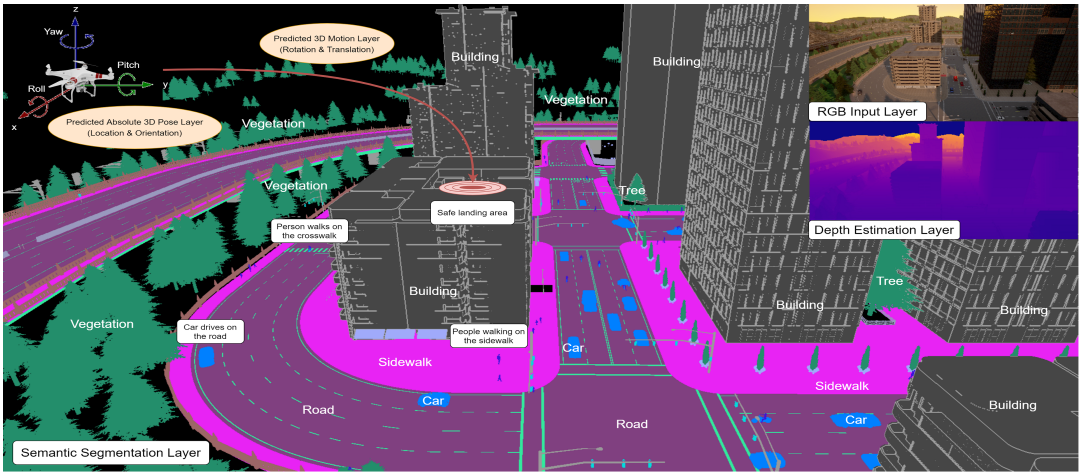

Figura 4: NGC poate combina diferite interpretări ale scenei dinamice, cum ar fi structura 3D, poziția, mișcarea, segmentarea semantică a obiectelor și activitățile din diferite regiuni ale spațiului și timpului, într-un graf neuronal unificat, în care multiple căi care ajung la un nod dat devin profesor prin acorduri consensuale pentru orice rețea de muchii singulară care ajunge la același nod. Antrenat în acest mod auto-supervizat, NGC poate atinge o învățare nesupervizată robustă în fața datelor neetichetate. Scena din figură este luată dintr-un mediu virtual folosit pentru a colecta date pentru experimentele noastre.

Ambele metode arată îmbunătățiri față de bazele de referință. Lucrările viitoare includ îmbunătățirea hărților existente, reprezentări multi-nivel combinând vederi de la sol și aeriene, și abordarea variațiilor lățimii drumurilor.

4 Spre o înțelegere completă a lumii cu o dronă

Ne propunem să înțelegem mai bine lumea din multiple reprezentări. Mai întâi, prezentăm SafeUAV, o soluție de aterizare sigură care folosește adâncimea și normalele suprafeței pentru a genera regiuni de aterizare sigure. Apoi prezentăm Neural Graph Consensus, un algoritm auto-supervizat pentru îmbunătățirea înțelegerii scenei bazat pe un set divers de reprezentări, așa cum este arătat în Figura 4.

4.1 Aterizare sigură pentru UAV-uri

Propunem SafeUAV-Net, un sistem încorporabil bazat pe rețele convoluționale profunde pentru estimarea adâncimii și a zonei de aterizare sigură folosind doar intrare RGB. Producem un set de date sintetice și antrenăm pe acesta, arătând performanțe convingătoare pe imagini reale de dronă.

4.1.1 SafeUAV-Net pentru estimarea adâncimii și orientării planului

Scopul nostru este de a prezice adâncimea și de a clasifica orientarea planului în trei clase: orizontal, vertical și altele. Sarcinile noastre sunt legate de segmentarea semantică, deoarece prezic o valoare categorică pentru fiecare pixel. Folosim o variantă a modelului U-Net propus de [40] pentru segmentarea

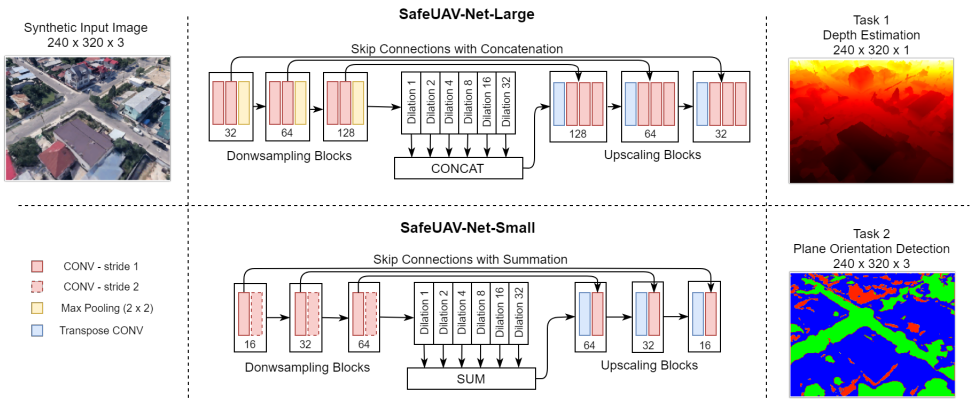

Figura 5: SafeUAV-Net-urile noastre propuse atât pentru procesarea la bord, cât și în afara bordului, antrenate pentru estimarea adâncimii și predicția orientării planului.

Tabela 7: Rezultatele predicției HVO pentru SafeUAV-Net-Large și SafeUAV-Net-Small antrenate pe setul complet de date.

Model	Intrare	Acuratețe	Precizie	Recall	mIoU
U-net [38]	RGB	0.729	0.560	0.505	0.356
DeepLabv3+ [42]	RGB	0.840	0.753	0.739	0.597
Small	RGB	0.823	0.728	0.693	0.551 heightLarge
RGB	**0.846**	**0.761**	**0.748**	**0.607**	

imaginilor aeriene. Am dezvoltat două variante: SafeUAV-Net-Large rulează la 35 FPS pe Nvidia Jetson TX2, în timp ce SafeUAV-Net-Small rulează la 130 FPS. Arhitecturile detaliate sunt descrise în Figura 5.

4.1.2 Set de date

Ne construim setul de date virtual folosind reconstrucțiile 3D Google Earth [41]. Setul de date constă din 11.907 eșantioane într-o divizare de 80

4.1.3 Experimente

Raportăm rezultate calitative și cantitative pentru estimarea adâncimii și segmentarea HVO pe toate cele patru regiuni din setul nostru de date. Tabelele 7 și 8 arată rezultatele. Experimentele noastre pe cazuri de test sintetice nevăzute arată că sistemul nostru este numeric precis, fiind în același timp

Tabela 8: Rezultate pentru estimarea adâncimii pentru SafeUAV-Net-Large și SafeUAV-Net-Small antrenate pe setul complet de date. Erorile sunt exprimate în metri.

Model	Intrare	RMSE	Metri
U-net [38]	RGB	0.041	9.63
DeepLabv3+ [42]	RGB	0.034	8.49
Small	RGB	0.031	7.22
Large	RGB	**0.026**	**6.09**

rapid pe un GPU embedded. Credem că utilizarea abordării noastre pe drone comerciale ar putea îmbunătăți siguranța zborului în zone urbane sau suburbane la viteze mari și ar completa senzorii de la bord.

4.2 Învățare Semi-Supervizată pentru Înțelegerea Multi-Sarcină a Scenei folosind Consensul Grafului Neural

Propunem Neural Graph Consensus (NGC), un model nou pentru învățarea semi-supervizată a mai multor interpretări ale scenei. NGC conectează mai multe rețele profunde într-un graf neural mare, unde fiecare nod reprezintă o interpretare diferită a scenei (de exemplu, adâncime, segmentare semantică, poziție). Muchiile dintre noduri sunt rețele profunde care transformă o reprezentare în alta. Modelul NGC este ilustrat în Figura 6.

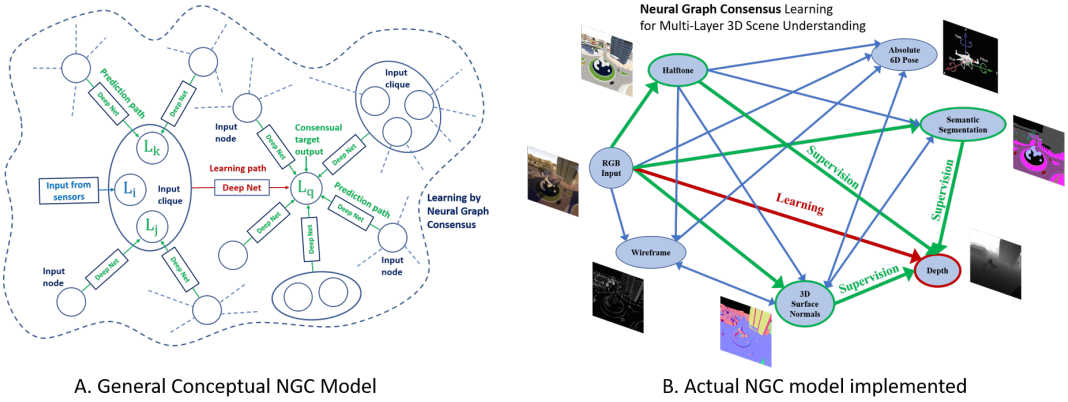

Figura 6: Arhitectura modelului Neural Graph Consensus (NGC)

4.3 Modelul Neural Graph Consensus

Fiecare nod i din graf are un strat asociat L_i, codificând o viziune sau interpretare specifică a lumii spațio-temporale. Straturile de la noduri pot fi prezise din alte straturi prin rețele profunde care formează muchii în graf. În timpul învățării nesupervizate, fiecare rețea devine un student al grafului NGC și este antrenată prin consensul mutual din căile contextuale care ajung la același nod de ieșire. NGC devine un sistem auto-supervizat în care acordul este profesorul suprem pentru datele neetichetate.

4.3.1 Analiză experimentală

Captăm un set mare de date folosind un simulator CARLA personalizat [43], unde o dronă zboară deasupra unui oraș prezicând adâncimea scenei, normalele suprafeței 3D, poziția absolută 6D, scheletul scenei și segmentarea semantică dintr-o singură imagine. Am dezvoltat un cadru general NGC pe baza PyTorch [44]. Pentru EdgeNets am folosit arhitecturile Map2Map și Map2Vector, fiecare cu aproximativ 1,1M parametri antrenabili. Modelul NGC total are 27 de rețele de muchii, totalizând aproximativ 30M parametri. Tabelul 9 arată rezultatele pentru ansamblul nostru propus NGC și EdgeNets distilate pe 6 reprezentări pe parcursul a 2 iterații de învățare nesupervizată. Am comparat NGC cu metodele de învățare multi-sarcină de ultimă generație NDDR [45] și MTL-NAS [46], și metoda de învățare semi-supervizată CCT [47]. Tabelele 10 și 11 arată rezultatele. Figura 7 vizualizează îmbunătățirile obținute de NGC față de modelele de bază cu sarcină unică pentru diverse sarcini de înțelegere a scenei.

4.4 Concluzii

Am prezentat SafeUAV-Net pentru estimarea adâncimii și a zonei de aterizare sigură și modelul Neural Graph Consensus pentru învățare semi-supervizată multi-sarcină. Ambele abordări arată rezultate

Reprezentare	Metric de Evaluare	Iterația 0	Iterația 1		Iterația 2	
		EdgeNet	NGC	EdgeNet Distilat	NGC	EdgeNet Distilat
Adâncime	L1 (metri)	4.9844	3.4867	4.2802	**3.2994**	**3.9508**
Normale de Suprafață (C)	L1 (grade)	8.4862	7.7914	8.2891	**7.4503**	**7.6773**
Normale de Suprafață (W)	L1 (grade)	11.8859	8.8248	10.7500	**8.5282**	**8.6714**
Segmentare Semantică	mIOU	0.4840	0.4978	0.4980	**0.5258**	**0.5159**
Schelet	Acuratețe	0.9617	0.9655	0.9654	**0.9661**	**0.9655**
Poziție	L2 (metri)	25.7597	15.5383	20.0204	**12.0764**	**15.5599**
Orientare	L1 (grade)	3.8439	2.5001	3.3961	**2.2088**	**3.0005**

Tabela 9: Rezultate pentru ansamblul nostru propus NGC și EdgeNets distilate pe 6 reprezentări, pe parcursul a 2 iterații de învățare nesupervizată.

Sarcină	Metrică	EdgeNet(iter 0)	NGC	NDDR*(fără preantrenare)	NDDR*(preantrenare)	NDDR	MTL-NAS
Segmentare	mIOU	0.484	**0.498**	0.141	0.343	0.315	0.368
Semantică	Acuratețe	90.017	**91.816**	48.7	84.2	86.9	87.8
Normale (C)	Eroare (grade)	8.4862	7.7914	9.820	7.727	6.801	**6.533**

Tabela 10: Rezultate de învățare multi-sarcină. Toate metodele au fost antrenate pe aceleași date supervizate (setul nostru de antrenament) și testate pe setul de evaluare.

promițătoare pe date sintetice și reale. Lucrările viitoare ar putea explora utilizarea continuității spațiale și temporale în secvențe video pentru predicții mai robuste și îmbunătățirea siguranței prin geolocalizare vizuală.

5 Către construirea eficientă a structurii 3D

Propunem o metodă eficientă pentru învățarea nesupervizată a estimării adâncimii metrice dintr-o singură imagine în contextul videoclipurilor necondiționate capturate de UAV-uri. Combinăm precizia unei soluții analitice bazate pe odometrie cu puterea învățării profunde. Abordarea noastră, numită UFODepth, depășește metodele de ultimă generație pe un set de date UAV pe care îl extindem semnificativ.

5.1 UFODepth: Învățare nesupervizată cu optimizarea odometriei bazată pe flux pentru estimarea adâncimii metrice

5.1.1 Introducere

Învățarea nesupervizată a estimării adâncimii metrice din video și odometrie poate avea un impact puternic asupra navigației autonome. Pentru UAV-uri, aterizarea în siguranță este o sarcină crucială care ar trebui realizată cu precizie ridicată folosind doar senzorii disponibili la bord. Abordarea noastră combină două direcții complementare:

UFODepth are o soluție analitică pentru estimarea adâncimii din fluxul optic și vitezele măsurate ale camerei - corectate cu o nouă abordare de optimizare. Adâncimea analitică este apoi utilizată atât ca intrare, cât și ca termen de cost suplimentar pentru antrenarea rețelei finale UFODepth pentru estimarea adâncimii.

Metrică	EdgeNet (iter 0)	NGC (iter 2)	EdgeNet (iter 2)	CCT (supervizat)	CCT (semi-supervizat)	heightmIOU
0.484	**0.526**	0.516	0.353	0.353		
Acuratețe	0.9001	0.9245	**0.9283**	0.8463	0.8503	

Tabela 11: Comparații de segmentare semantică pe setul nostru de evaluare cu CCT[47] semi-supervizat.

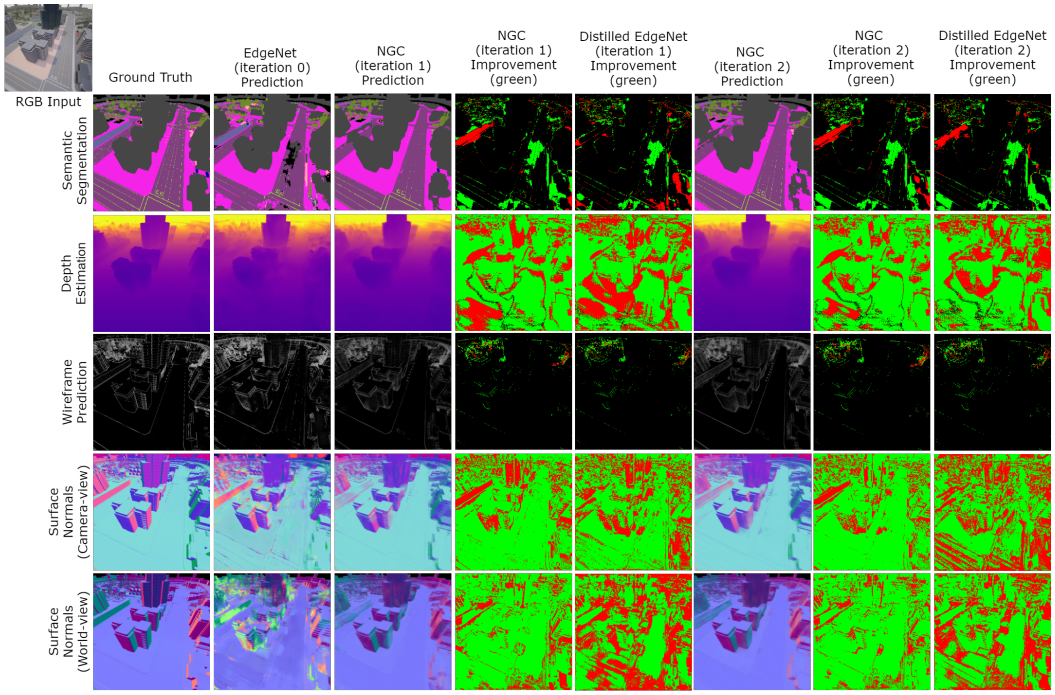

Figura 7: Comparație vizuală a rezultatelor NGC cu metodele de bază pentru diverse sarcini de înțelegere a scenei

5.1.2 Context științific

Metodele SLAM în timp real recente au ca rezultat o ieșire cu rezoluție scăzută (256x192 pixeli) [48]. Alte abordări care realizează reconstrucția 3D în timp real (de exemplu, NeuralRecon [49]) se bazează pe reprezentări SDF care nu produc direct o hartă de adâncime. Metodele recente de învățare profundă, cum ar fi estimarea consistentă a adâncimii [50] sau abordările bazate pe NeRF [51, 52] produc adâncime metrică doar dacă încep de la rezultate intermediare SfM scalate și necesită ajustare fină pe fiecare scenă. Mai multe metode se bazează pe învățare nesupervizată [53–55]. Cele două lucrări cele mai similare pentru adâncimea nesupervizată care pretind operare în timp real sunt [56] și [57]. Alte abordări recente care oferă o ieșire de adâncime de înaltă calitate necesită în general pre-antrenare intensivă pe seturi de date foarte mari [58, 59].

5.1.3 Optimizarea odometriei bazată pe flux

Ne propunem să calculăm adâncimea metrică robustă din fluxul optic și odometria camerei. Mai întâi derivăm o soluție analitică, pe care o folosim apoi pentru a corecta măsurătorile odometrice inițiale. Fluxul optic de la imagine la imagine cauzat de mișcarea camerei poate fi scris ca o sumă a două componente: fluxul linear $\mathbf{F}_\nu$ cauzat de translații și fluxul rotațional $\mathbf{F}_\omega$ produs de rotație. Derivata temporală a unui punct 3D din scenă $P = (X, Y, Z)$ în sistemul de coordonate al camerei este legată de mișcarea camerei prin vitezele liniare (ν) și unghiulare (ω):

$$\dot{\mathbf{P}} = -\omega \times \mathbf{P} - \nu. \tag{1}$$

Putem defini fluxul optic ca o funcție de mișcarea instantanee a camerei și adâncime:

$$\left(\dot{\bar{u}}\nu \;\; \dot{\bar{v}}\nu\right) = \frac{1}{Z}\begin{pmatrix} -f & 0 & \bar{u} & 0 \\ 0 & -f & \bar{v} & \end{pmatrix}\begin{pmatrix} \nu_x \\ \nu_y \\ \nu_z \end{pmatrix}. \tag{2}$$

$$\left(\dot{\bar{u}}\omega \;\; \dot{\bar{v}}\omega\right) = \begin{pmatrix} \frac{\bar{u}\bar{v}}{f} & -\frac{f^2+\bar{u}^2}{f} & \bar{v} & \frac{f^2+\bar{v}^2}{f} & -\frac{\bar{u}\bar{v}}{f} & -\bar{u} \end{pmatrix}\begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix}. \tag{3}$$

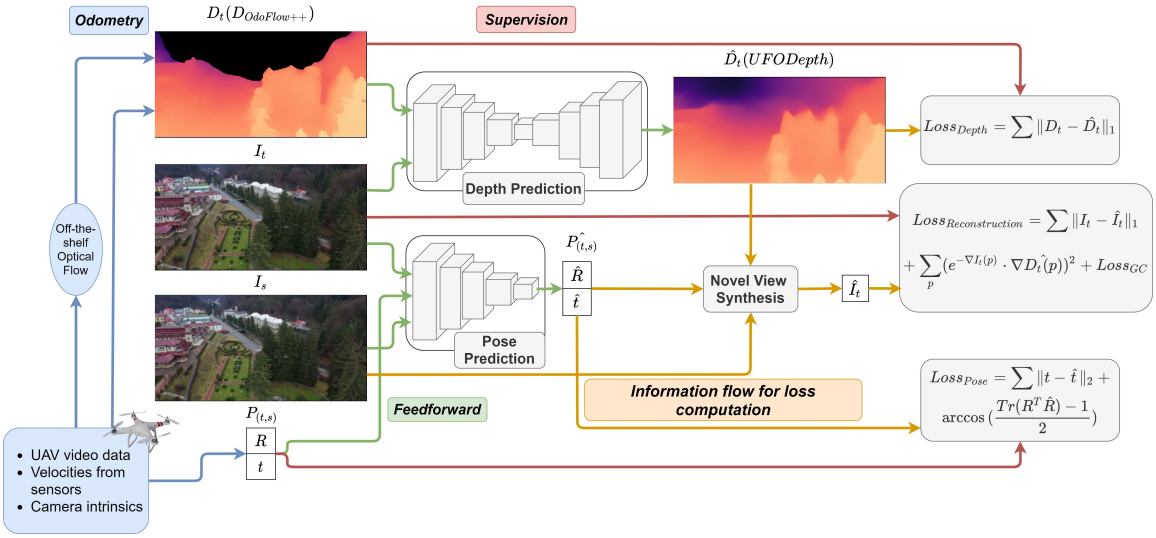

Figura 8: Prezentare generală a abordării noastre (UFODepth). Combinăm trei tipuri de pierderi cu o formulare matematică îmbunătățită pentru adâncimea din fluxul optic.

Tabela 12: Erori medii absolute și relative față de adâncimea de referință D_{SfM}.

Metodă	Slanic		Chilia		Olanesti		Herculane		Oveselu		General		
	Metric	Rel.	Metric	Rel.	Metric	Rel.	Metric	Rel.	Metric	Rel.	Metric	Rel.	Acoperire
$D_{Triangulation}$[60]	25,39	18,9	14,79	11,2	51,82	49,3	23,19	13,4	21,79	19,9	21,29	15,6	74,0 %
$D_{OdoFlow}$ [56]	18,33	13,7	**12,90**	11,0	13,5	**11,4**	17,36	9,6	19,50	17,7	16,32	12,7	74,0 $D_{OdoFlow++}$
16,06	**12,4**	13,92	**10,4**	**13,4**	11,7	**14,46**	**8,6**	**18,52**	**17,3**	**15,27**	**12,0**	74,0 %	

5.1.4 Antrenare nesupervizată

Procedura noastră de antrenare este inspirată de [55]. Ne bazăm pe pierderea de consistență geometrică, denumită $\mathcal{L}GC$, și introducem două pierderi suplimentare, $\mathcal{L}Depth$ și $\mathcal{L}_{Pose}$. Obiectivul nostru este minimizarea:

$$\mathcal{L} = \mathcal{L}Reconstruction + \mathcal{L}GC + \mathcal{L}Depth + \mathcal{L}Pose. \tag{4}$$

O prezentare generală a abordării noastre este prezentată în Figura 8.

5.1.5 Analiză experimentală

Testăm și extindem un set de date recent publicat pentru cercetarea viziunii UAV. **Setul de date Slanic și Herculane:** Acest set de date include un total de 20 de minute de secvențe video 4K din două scene urbane din Europa de Est [56]. Include informații odometrice la o frecvență de 10 Hz. **Set de date odometrie extins:** Introducem trei scene noi (Oveselu, Olanesti, Chilia) cu peisaje variate. Setul de date are un total de 33 de minute de zbor real cu drona, la 30 FPS.

5.1.6 Rezultate

Tabelul 12 arată performanța metodei noastre de estimare analitică a adâncimii $D_{OdoFlow++}$ comparativ cu lucrările anterioare. Tabelul 13 compară abordarea noastră completă UFODepth cu metodele de ultimă generație. Rezultatele noastre arată că UFODepth depășește metodele anterioare în medie pe toate scenele. Componenta analitică $D_{OdoFlow++}$ oferă estimări precise ale adâncimii acolo unde este valabilă. Abordarea completă UFODepth se generalizează bine la scene noi, menținând în același timp o viteză de inferență competitivă.

Tabela 13: Erori medii absolute și relative față de adâncimea de referință D_{SfM}. Metodele care nu oferă estimări metrice sunt scalate către $D_{OdoFlow++}$ pentru o comparație corectă.

Metodă	Slanic		Chilia		Olanesti		Herculane		Oveselu		General	
	Metric	Relativ	Metric	Relativ	Metric	Relativ	Metric	Relativ	Metric	Relativ	Metric	Relativ
D_{Unsup}[55]	25,00	15,4	44,52	23,4	25,85	18,4	34,40	16,0	31,10	22,3	32,17	19,1
$D_{Ensemble}$[56]	24,83	14,9	37,93	18,4	22,46	**15,2**	34,28	16,6	33,15	22,1	30,53	17,44
Tiny-16 [56]	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 [59]	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 [58]	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 [57]	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**

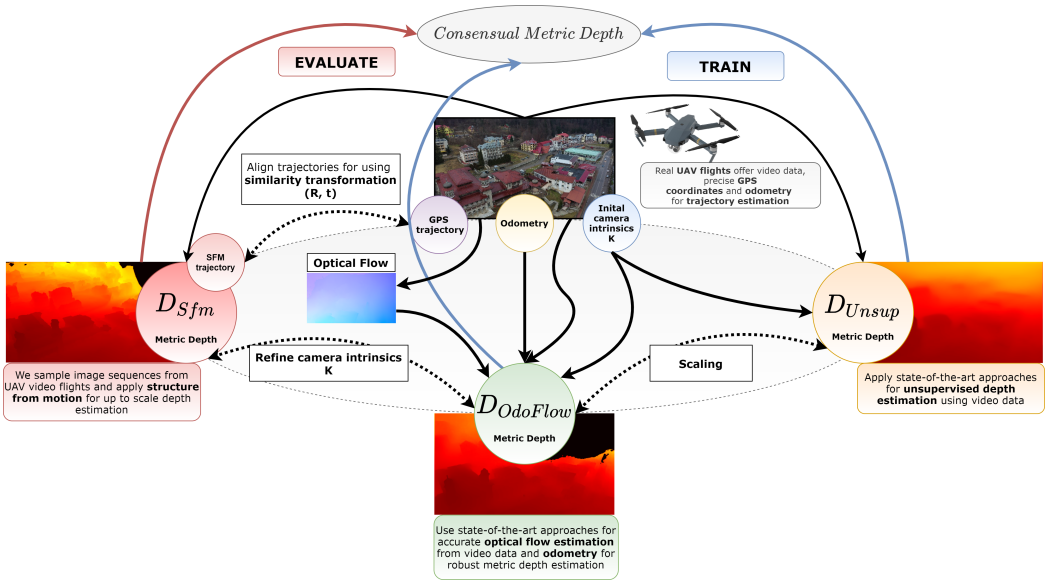

Figura 9: Prezentare generală a abordării noastre de distilare a adâncimii, combinând mai multe căi complementare pentru estimarea precisă a adâncimii metrice.

5.2 Distilarea adâncimii: estimarea nesupervizată a adâncimii metrice pentru UAV-uri

Explorăm, de asemenea, o abordare de distilare a adâncimii care combină metode analitice și bazate pe date. O prezentare generală este prezentată în Figura 9.

Folosim trei tipuri de hărți de adâncime:

D_{SfM}: Adâncimea din structure-from-motion, utilizată ca referință pentru evaluare. $D_{OdoFlow}$: Adâncimea analitică din fluxul optic și odometrie. D_{Unsup}: Adâncimea dintr-o rețea profundă nesupervizată, scalată pentru a fi metrică folosind $D_{OdoFlow}$.

Formăm un profesor ansamblu combinând $D_{OdoFlow}$ și D_{Unsup} și folosim acesta pentru a antrena o rețea student ușoară pentru estimarea adâncimii în timp real.

5.2.1 Rezultate

Tabelele 14 și 15 prezintă rezultatele pe întreaga zonă a imaginii și pe zona "bună" unde $D_{OdoFlow}$ este valid. Rețelele student distilate (Tiny-16 și Large-16) depășesc profesorii lor pe setul de test Slanic. Deși performanța scade pe setul de date Herculane nevăzut, studenții rămân competitivi. Tabelul 16 prezintă viteza de inferență pe GPU-uri desktop și embedded.

	Slanic		Herculane	
	Metric	Relativ	Metric	Relativ
D_{Unsup}	27,28 m	17,10 %	44,39 m	20,29 %
$D_{OdoFlow}$	26,05 m	16,34 %	**39,67 m**	**17,53 %**
$D_{Ensemble}$	25,63 m	15,88 %	41,18 m	18,29 %
$Tiny - 16$	**21,58 m**	**14,58 %**	46,77 m	24,09 %
$Large - 16$	21,84 m	14,65%	48,00 m	23,97 %

Tabela 14: Erori medii absolute și relative pe întreaga hartă validă față de adâncimea de referință D_{SfM}.

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

Tabela 15: Erori absolute și relative în zona bună, unde atât predicțiile D_{SfM}, cât și $D_{OdoFlow}$ sunt valide.

Metodă	Parametri	Desktop[FPS]	Embedded[FPS]
D_{Unsup}[55]	14.842.236	$166,703 \pm 3,27$	$11,631 \pm 0,55$
$OpticalFlow(doar)$[61]	15.263.888	$83,404 \pm 2,65$	$43,699 \pm 3,13$
$D_{OdoFlow}$	n/a	$26,807 \pm 6,17$	$10,127 \pm 0,73$
$Tiny - 16$[62]	1.119.862	$54,922 \pm 1,33$	$10,357 \pm 0,22$
$Large - 16$[62]	2.005.239	$51,969 \pm 1,32$	$9,045 \pm 0,13$

Tabela 16: Cadre pe secundă pe GPU-uri desktop și embedded. Algoritmul de adâncime din flux rulează pe CPU. Desktop-ul are un GPU RTX 2080 și CPU Ryzen 7 3700.

5.3 Concluzii

Propunem două abordări complementare pentru estimarea nesupervizată a adâncimii metrice din videoclipuri UAV:

UFODepth combină o metodă analitică de estimare a adâncimii cu învățare profundă nesupervizată, obținând rezultate de ultimă generație pe setul nostru de date UAV extins. O abordare de distilare a adâncimii care folosește un ansamblu de metode analitice și nesupervizate ca profesor pentru a antrena o rețea student ușoară.

Ambele metode demonstrează o bună generalizare la scene noi și o performanță aproape în timp real pe platforme embedded, făcându-le potrivite pentru implementare la bord pe UAV-uri. Lucrări viitoare ar putea explora încorporarea unor constrângeri geometrice suplimentare și optimizarea pentru performanță în timp real pe dispozitive embedded.

6 O abordare ciclică neural-analitică auto-supervizată pentru sinteza de vederi noi și reconstrucția 3D

Generarea de noi vederi din înregistrări video este crucială pentru a permite navigația autonomă a dronelor (UAV). Progresele recente în renderizarea neurală au facilitat dezvoltarea rapidă a metodelor capabile să redea noi traiectorii. Cu toate acestea, aceste metode eșuează adesea în a generaliza bine în regiuni îndepărtate de datele de antrenament fără o traiectorie de zbor optimizată, ducând la reconstrucții sub-optimale. Propunem o soluție ciclică neural-analitică auto-supervizată care combină rezultatele de înaltă calitate ale renderizării neurale cu informații geometrice precise din metode analitice. Soluția noastră îmbunătățește atât reconstrucțiile RGB, cât și cele de tip mesh pentru sinteza de vederi noi, în special în zone sub-eșantionate și regiuni complet distincte de setul de date de antrenament.

6.1 Introducere

Abordările tradiționale pentru sinteza de vederi noi se concentrează predominant pe seturi de date sintetice sau centrate pe obiecte, caracterizate prin zgomot minim în datele de intrare. Testarea are loc de obicei pe aceleași imagini folosite pentru antrenament sau la intervale regulate într-o secvență [63]. Deși tehnicile actuale de renderizare neurală pot produce reconstrucții de înaltă calitate pe datele de antrenament – cu valori PSNR adesea depășind 30 [64], traiectoriile de cameră care se îndepărtează semnificativ de datele de antrenament duc frecvent la reconstrucții inferioare. O strategie prevalentă pentru atenuarea acestor limitări implică segmentarea seturilor de date în regiuni mai mici și consistente [65, 66]. Cu toate acestea, această abordare necesită resurse computaționale substanțiale, deoarece trebuie antrenat un model separat pentru fiecare secvență [67, 68]. Introducem o abordare ciclică neural-analitică care utilizează punctele forte ale metodelor de structure-from-motion și de renderizare neurală pentru a sintetiza imagini RGB de înaltă fidelitate în poziții noi, departe de setul de antrenament. Soluția noastră folosește o strategie de reconstrucție în două faze. Inițial, o reconstrucție 3D analitică se aliniază cu rezultatele ramurii de renderizare neurală pentru a rafina consistența geometrică. Ulterior, adaptăm un transformer ușor pentru restaurarea imaginilor pentru a obține o reconstrucție de înaltă rezoluție a vederilor 2D noi. Principalele noastre contribuții sunt:

1. Combinăm metode analitice și neurale de reconstrucție 3D printr-o nouă abordare ciclică auto-supervizată bazată pe transformeri și demonstrăm îmbunătățiri atât pentru sinteza de vederi noi, cât și pentru reconstrucția 3D prin învățare iterativă.

2. Cadrul nostru demonstrează capacități de generalizare, generând imagini de calitate din locații de test fără date de antrenament suplimentare.

3. Arătăm îmbunătățiri comparativ cu metodele state-of-the-art pentru cazuri dificile de scene din lumea reală capturate de drone, acoperind zone largi cu zgomot semnificativ în poziție și adâncime.

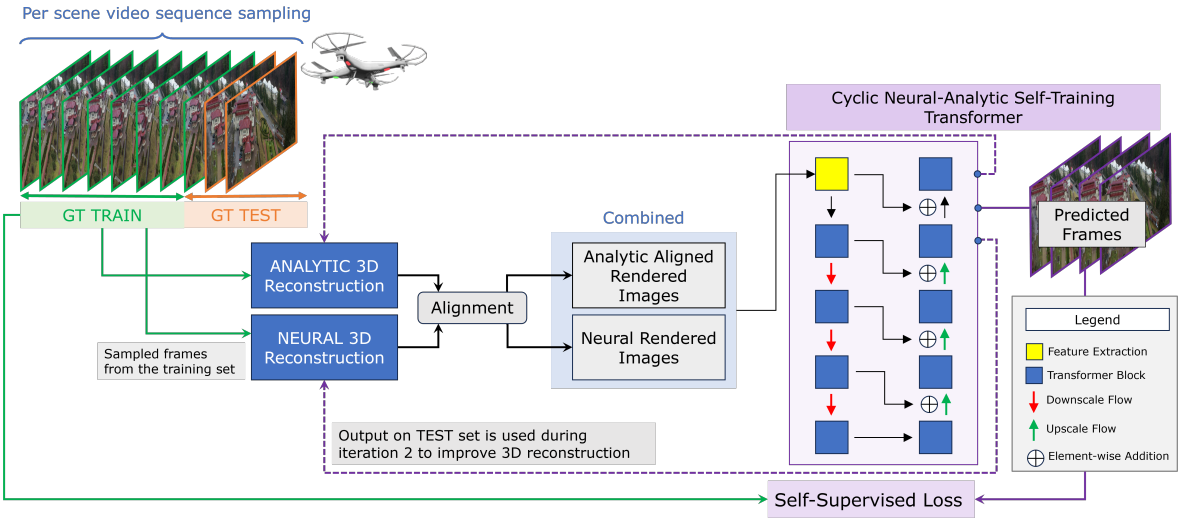

Figura 10: Privire de ansamblu asupra soluției noastre ciclice neural-analitice auto-supervizate pentru sinteza de vederi 2D noi. Ne bazăm atât pe metode tradiționale, cât și moderne de reconstrucție 3D pe care le combinăm printr-un model auto-supervizat de tip U-net bazat pe transformeri pentru o reconstrucție îmbunătățită a imaginilor. Folosim o procedură de învățare iterativă în care rezultatele primei iterații de învățare devin intrări pentru următoarea pentru a rafina în continuare rezultatele în termeni de RGB și mesh, fără imagini noi suplimentare. Lucrăm în domeniul video al dronelor și folosim ultimele 20% din secvența de imagini pentru testare pentru a simula un scenariu de reconstrucție mai realist. **Cadrele RGB originale din setul de TEST sunt folosite exclusiv în scopuri de evaluare.**

6.2 Metodă

Soluția noastră constă din module analitice și neurale complementare (Figura 10). Ambele sunt capabile să redea imagini din poziții noi. Iterația inițială folosește setul de antrenament pentru ambele module. Noi imagini sunt generate din pozițiile de test și incluse în datele de antrenament. Repetăm ciclul cu imaginile nou generate.

6.2.1 Reconstrucție 3D Analitică

Folosim un pipeline standard de reconstrucție 3D fotogrammetrică [14, 69, 70]. Pipeline-ul procesează imagini neordonate pentru a reconstrui scene 3D, prin extragerea de caracteristici, potrivire, structure-from-motion și stereo multi-vedere. Norul de puncte dens rezultat este transformat într-un mesh texturat.

6.2.2 Reconstrucție 3D Neurală

Folosim metoda Gaussian Splatting [64], care reprezintă scena folosind gaussiene 3D caracterizate prin poziție, matrice de covarianță anizotropică și opacitate. Această reprezentare captează eficient detaliile scenei fără eșantionare densă. Procesul de renderizare utilizează rasterizare bazată pe dale, proiectând gaussienele 3D pe un plan 2D pentru o amestecare eficientă.

6.2.3 Sinteza de Vederi Noi Neural-Analitică Auto-Supervizată

Combinăm reconstrucția 3D analitică și neurală anterioară adaptând o arhitectură ușoară bazată pe transformeri [71] pentru o reconstrucție îmbunătățită a imaginilor. Modelul seamănă cu o arhitectură de tip U-net cu blocuri de transformeri folosite în locul straturilor convoluționale simple. Antrenăm modelul în mod auto-supervizat folosind cadre RGB din secvențe video ca supervizare atunci când calculăm pierderea de eroare pătratică medie pixel cu pixel pe predicții.

Tabela 17: Comparație cu metodele state-of-the-art pentru sinteza de vederi noi. Rezultatele de reconstrucție PSNR pe setul de date Aerial [72] exclusiv pentru setul de test. Cele mai bune numere pentru fiecare set individual sunt **îngroșate**.

MetodăScenă	Slanic	Olanesti	Chilia	Herculane	Media
SFM [70]	18.43	17.63	18.75	18.27	18.27
DVGO [77]	15.34	14.34	16.43	15.43	15.38
Plenoxels [78]	17.22	15.92	18.59	16.07	15.38
Instant-NGP [13]	16.54	19.25	21.01	19.12	18.98
Neuralangelo [76]	16.79	16.20	16.96	15.82	17.29
GeoView [79]	20.28	18.64	19.04	18.17	19.03
Gaussian Splatting [80]	19.37	19.34	21.85	20.85	20.35
CNA (Al nostru)	**19.85**	**19.81**	**22.21**	**21.39**	**20.82**

Tabela 18: Comparație pe setul de test cu Instant-NGP, Gaussian Splatting și metoda de reconstrucție analitică de bază pe trei scene suplimentare. Am selectat scena *5b69cc0cb44b61786eb959bf* din BlendedMVS, scena Rubble din MIL-19 și scena Family din Tanks and Temples. Structura de împărțire este aceeași ca în experimentele noastre anterioare (80% antrenare, 20% testare). Cele mai bune numere pentru fiecare set individual sunt **îngroșate**.

MetodăScenă	BlendedMVS	Rubble	Tanks and temples	Media
Instant-NGP [13]	14.54	11.54	14.61	13.56
Analitic [70]	13.95	13.65	15.73	14.44
Gauss.Splatt. [80]	21.52	15.53	19.63	18.89
CNA (Al nostru)	**21.67**	**16.04**	**19.81**	**19.17**

6.2.4 Rafinare Ciclică 3D

Metoda noastră include o rafinare ciclică care îmbunătățește rezultatele fără supervizare suplimentară. Prin reintroducerea imaginilor redate din pozițiile de test înapoi în algoritm, observăm câștiguri de performanță în reconstrucția imaginilor.

6.3 Analiză Experimentală

Aplicăm cadrul nostru pe o serie de scene în aer liber, cu scopul de a îmbunătăți performanța pe setul de test fără etichete sau reantrenare suplimentară. Folosim următoarele seturi de date:

- **Setul de date Aerial** [72]: Conține date telemetrice și peisaje diverse capturate de drone.
- **BlendedMVS** [73]: Un set de date de reconstrucție a scenelor în aer liber cu imagini redate.
- **Rubble** [74]: Un set de date capturat cu drona cu imagini detaliate.
- **Tanks and Temples** [75]: Folosim scena Family pentru evaluarea centrată pe obiect.

Împărțim fiecare scenă în 80% date de antrenament și 20% date de test. Rezultatele sunt evaluate pe setul de test pentru a asigura generalizarea.

6.4 Rezultate și Discuție

Comparăm metoda noastră cu abordări state-of-the-art, incluzând Instant-NGP [13], Neuralangelo [76], Direct Voxel Grid Optimization [77], Plenoxels [78] și Gaussian Splatting [64]. Tabelul 6.4 prezintă rezultatele pe setul de date Aerial. Abordarea noastră Ciclică Neural-Analitică (CNA) oferă rezultate îmbunătățite fără etichete suplimentare, cu îmbunătățiri consistente pe toate scenele. Rezultatele pe alte seturi de date sunt prezentate în Tabelul 6.4. Observăm îmbunătățiri pe diferite tipuri de scene, inclusiv cele centrate pe obiect, capturate cu drona și reconstruite sintetic. Studiul nostru de ablație (Tabelul 6.4) arată o îmbunătățire consistentă de-a lungul iterațiilor, demonstrând eficacitatea

Tabela 19: Studii de ablație pentru fiecare dintre componentele metodei noastre propuse, CNA pe setul de date Aerial [72]. Rezultatele de reconstrucție PSNR sunt raportate exclusiv pe setul de test. (1) și (2) denotă prima și a doua iterație a CNA. Cele mai bune numere pentru fiecare set individual sunt **îngroșate**.

MetodăScenă	Slanic		Olanesti		Chilia		Herculane		Media	
Iterație	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Analitic (doar)	18.43	18.66	17.63	16.99	18.75	19.12	18.27	18.83	18.27	18.40
Analitic (aliniat)	19.11	19.33	17.73	17.05	18.95	19.81	18.42	19.01	18.55	18.80
Neural (doar)	19.49	19.73	19.48	**19.85**	21.28	22.19	21.27	21.38	20.38	20.79
CNA (Al nostru)	19.65	**19.85**	19.81	19.81	21.90	**22.21**	21.32	**21.39**	20.67	**20.82**

abordării noastre ciclice. Evaluăm, de asemenea, calitatea reconstrucției 3D comparând mesh-urile generate doar din datele de antrenament cu cele îmbunătățite de imaginile redate de CNA. Rezultatele arată o îmbunătățire consistentă în acuratețea reconstrucției 3D (Tabelul 6.4).

Tabela 20: Erori de reconstrucție 3D pe parcursul iterațiilor pe setul de date Aerial [72] exclusiv pentru setul de test. Cele mai bune numere pentru fiecare set individual sunt **îngroșate**.

MetodăScenă	Slanic		Olanesti		Chilia		Herculane	
	Medie	Mediană	Medie	Mediană	Medie	Mediană	Medie	Mediană
Iterația 1	0.6102	0.5263	0.5053	0.5013	0.5324	0.4932	0.5289	0.4802
Iterația 2	**0.4832**	**0.4392**	**0.3806**	**0.3646**	**0.3553**	**0.2934**	**0.4125**	**0.4225**

6.5 Concluzii

Am prezentat o soluție ciclică neural-analitică auto-supervizată care îmbină renderizarea neurală cu metode analitice bazate pe mesh. Soluția noastră îmbunătățește atât reconstrucțiile RGB, cât și cele de mesh 3D pentru poziții de vedere noi, semnificativ diferite de setul de antrenament. Experimentele au demonstrat eficacitatea atât a modulelor neurale, cât și a celor analitice, precum și utilitatea ciclării. Designul nostru permite înlocuirea modulelor, acoperind un teren nou în acest domeniu și având potențialul de a împinge limitele mai departe către sinteza de vederi noi înrădăcinată în lumea fizică 3D. În timp ce majoritatea metodelor sunt testate pe scene sintetice și pe aceleași imagini folosite pentru antrenament, noi am evaluat și am arătat îmbunătățiri față de state-of-the-art din puncte de vedere drastic diferite de cele văzute în antrenament. Mai mult, am arătat că metoda noastră suferă de o degradare mică între datele văzute (antrenament) și nevăzute (testare), spre deosebire de competiția recentă care prezintă o degradare semnificativă, indicând un overfitting puternic.

6.6 Lucrări Viitoare

Ne propunem să îmbunătățim eficiența pipeline-ului de antrenare pentru a beneficia de cele mai recente inovații în reconstrucția 3D - antrenare mai rapidă, acuratețe mai bună, adâncime, normale de suprafață și mesh ca reprezentări de ieșire. Ideal, am îmbunătăți iterativ un mesh texturat care a utilizat toate informațiile din imaginea RGB. Preprocesarea pentru rolling shutter și blur de mișcare ar putea duce la o ieșire PSNR mai mare, așa cum s-a arătat în 3GS Deblur [81]. Compararea mesh-urilor în timp este un domeniu important de interes (de exemplu, detectarea schimbărilor de mediu precum creșterea vegetației sau degradarea clădirilor istorice) și ne propunem să adaptăm conceptul prezentat aici către o ieșire de tip mesh în loc de RGB. Am folosi experiența acumulată din renderizarea datelor sintetice și reconstrucția 3D pentru a atribui clase de pixeli și a construi o ieșire mai robustă. Figura 11 ilustrează îmbunătățirea consistentă a PSNR pe cadrele de test pentru metoda noastră CNA comparativ cu liniile de bază. Acest lucru demonstrează robustețea abordării noastre în generarea de vederi noi de înaltă calitate. În concluzie, abordarea noastră ciclică neural-analitică

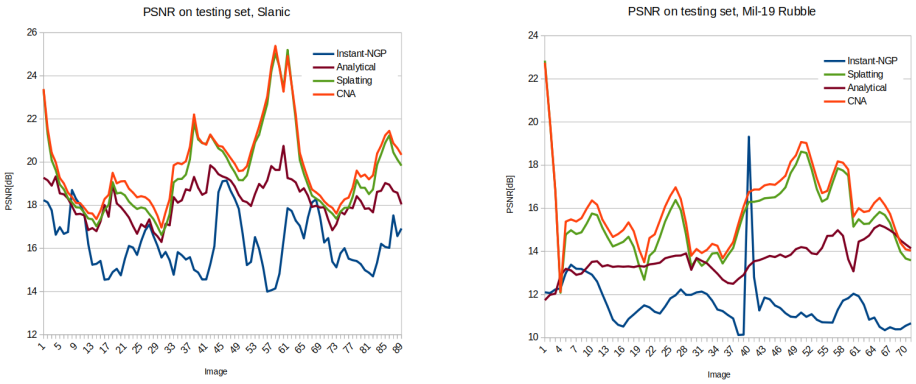

Figura 11: Eroarea PSNR pe Slanic pentru fiecare cadru de test. CNA îmbunătățește constant față de celelalte linii de bază, în ciuda faptului că nu primește imagini RGB suplimentare. Câștigurile de performanță nu sunt doar în medie, ci pe marea majoritate a cadrelor.

oferă o direcție promițătoare pentru îmbunătățirea sintezei de vederi noi și a reconstrucției 3D, în special pentru scene dificile din lumea reală capturate de drone. Prin combinarea punctelor forte ale metodelor analitice și neurale într-un cadru iterativ, obținem o mai bună generalizare la puncte de vedere nevăzute și o acuratețe îmbunătățită a reconstrucției 3D. Lucrările viitoare se vor concentra pe îmbunătățirea în continuare a eficienței, încorporarea înțelegerii suplimentare a scenei și extinderea abordării la comparații temporale de mesh pentru aplicații de monitorizare a mediului.

7 Dincolo de Scene Virtuale Statice: Un Chat Non-Verbal în Timp Real pentru Interacțiunea Om-IA

Prezentăm o abordare nouă a interacțiunii Om-IA prin intermediul unui chat non-verbal în timp real, valorificând expresiile faciale și mișcările corpului pentru a îmbunătăți angajamentul. Sistemul nostru își propune să capteze și să răspundă la emoțiile utilizatorului folosind tehnici de computer vision, operând în timp real cu resurse computaționale minime. Oferim trei abordări complementare bazate pe tehnici de recuperare, statistice și de învățare profundă, integrând o componentă artistică pentru a transmite emoții. Experimentele noastre compară modele de difuzie pentru traducerea emoțiilor în 2D și introduc un avatar 3D, Maia, cu mișcări faciale și corporale pentru o experiență mai naturală.

7.1 Introducere

Progresele recente în IA, în special Modelele Lingvistice Mari (LLM), au arătat performanțe impresionante în sarcini bazate pe text. Cu toate acestea, ele nu au capacitatea de a se angaja în interacțiuni complexe, multi-modale care caracterizează comunicarea umană. Sistemele actuale bazate pe LLM sunt limitate în capacitatea lor de a adapta răspunsurile la nevoile individuale și ratează indicii non-verbale cruciale precum gesturi, limbajul corpului și expresiile faciale. Aceste elemente transmit informații emoționale și intenționale complexe dincolo de textul scris. Lucrarea noastră abordează aceste limitări concentrându-se pe aspectele non-verbale ale comunicării, cu scopul de a reduce decalajul dintre capacitățile IA și natura nuanțată a interacțiunii umane. Propunem un sistem care interpretează și răspunde la o gamă largă de emoții folosind puncte cheie faciale și mișcări ale corpului, potențial conducând la o comunicare Om-IA mai intuitivă și cuprinzătoare.

7.2 Metodologie

7.2.1 Setul de Date pentru Expresii Non-Verbale (NED)

Din cauza lipsei de seturi de date pentru transmiterea expresiilor non-verbale, am colectat propriul nostru set de date, pe care îl vom face disponibil public. Setul de date conține 30 de videoclipuri

Set	Oameni (Medie)	GPT-4o
3 emoții	**91,66**	55,56
5 emoții	**50,69**	47,22

Tabela 21: Evaluarea la nivel uman și de mașină pentru 3 și 5 emoții generate cu DreamBooth și ControlNet. Acuratețea peste toate clasele în procente.

Metodă	Oameni (Medie)	GPT-4o
NN	55,56	27,78
PCA	**66,67**	22,22
Recuperare	44,44	38,89
Toate	55,56	29,63

Tabela 22: Evaluarea la nivel uman și de mașină pentru fiecare metodă propusă de generare a expresiei faciale 3D. Acuratețea peste toate clasele în procente.

pentru fiecare tip de emoție, variind de la 3 la 6 secunde per videoclip, capturate la 30 FPS. Pentru experimentele noastre, ne-am concentrat pe emoții pozitive: fericit, râzând și surprins.

7.2.2 Introducerea lui Maia

Am creat pe Maia, un personaj animat convertit în 3D folosind VRoidStudio [82] și picturi în ulei artistice originale. Pentru animație, folosim VSeeFace [83]. Textura folosită pentru crearea avatarului nostru este derivată dintr-o pictură care reprezintă emoția "fericit".

7.2.3 Proceduri de Evaluare

Folosim două mijloace de evaluare: o evaluare automată folosind GPT-4 [84] și o evaluare la nivel uman. Pentru evaluarea automată, GPT-4 analizează cadre individuale din videoclipurile de test, atribuind etichete emoționale bazate pe expresiile faciale detectate și indicii contextuale. Pentru evaluarea umană, am recrutat indivizi pentru a evalua același set de test folosit în evaluarea automată.

7.2.4 Generare 2D cu difuzie

Abordarea noastră 2D folosește OpenCV pentru procesarea input-ului video, OpenPose [85] pentru estimarea pozei, și un pipeline de estimare a adâncimii [86] pentru a genera hărți de adâncime. Folosim un model Stable Diffusion XL [87] cu condiționare dual ă ControlNet [88] pentru generarea de imagini.

7.2.5 Generarea expresiei faciale 3D

Propunem trei metode pentru răspunsul facial 3D:

Reconstrucția Spațiului de Expresie: Folosește PCA pentru reducerea dimensionalității pentru a crea un spațiu încorporat pentru emoții. Distilarea Reacției: Un cadru de învățare nesupervizată folosind o paradigmă Profesor-Student cu o rețea neuronală bazată pe MLP. Recuperarea Emoției Similare: Recuperează secvența de puncte cheie cea mai similară dintr-un set de date predefinit.

7.2.6 Generarea mișcării corpului 3D

Ne-am extins experimentele pentru a include mișcări ale corpului, concentrându-ne pe trei emoții: "fericit să te vadă", "entuziast" și "râzând". Am colectat în jur de 100 de videoclipuri de 5 secunde fiecare pentru aceste emoții și am evaluat cât de bine a transmis Maia emoția dorită prin mișcarea corpului.

7.3 Rezultate și Discuție

Pentru generarea 2D cu difuzie, evaluatorii umani au obținut o acuratețe medie de 91,66% în scenariul cu trei emoții, depășind semnificativ acuratețea de 55,56% a GPT-4o (Tabelul 21). Pentru generarea expresiei faciale 3D, metoda PCA a performant cel mai bine când a fost evaluată de evaluatori umani,

Set	Oameni (Medie)	GPT-4o zero-shot	GPT-4o few-shot 3 emoții
93,44	57,67	73,00	

Tabela 23: Evaluarea la nivel uman și automată pentru experimentele de mișcare a corpului Maia pe setul nostru de test.

Figura 12: Procesul de creare a avatarului nostru 3D Maia. Am folosit un avatar feminin implicit VRoidStudio peste care am aplicat textura facială din pictură pe fața avatarului, potrivind manual fiecare corespondență și apoi personalizând-o cu păr și îmbrăcăminte bazate pe personalitatea ei. Am folosit o pictură diferită ca fundal pentru Maia.

în timp ce abordarea de Recuperare a obținut cel mai mare scor cu GPT-4o (Tabelul 22). Pentru generarea mișcării corpului 3D, evaluatorii umani au obținut o acuratețe medie de 93,44%, în timp ce GPT-4o a obținut 57,67% în scenariul zero-shot și 73,00% în scenariul few-shot (Tabelul 23).

7.4 Considerații Etice

Prioritizăm confidențialitatea și protecția datelor în abordarea noastră. Metoda noastră se bazează pe puncte cheie faciale și corporale mai degrabă decât pe date video brute, servind ca o caracteristică de păstrare a confidențialității. Extragem aceste puncte cheie din datele faciale și din insight-urile emoționale, anonimizând informațiile și eliminând input-ul video original. Această abordare ne permite să construim un lac de date cuprinzător de emoții fără a risca confidențialitatea utilizatorului.

7.5 Impact și Lucrări Viitoare

Sistemul nostru are potențiale aplicații în educație, terapie pentru persoanele cu provocări de comunicare și expoziții interactive în muzee. Ar putea fi deosebit de benefic pentru copiii cu nevoi speciale care necesită angajament și interacțiune constantă. Lucrările viitoare se vor concentra pe generarea de avatare animate bazate pe puncte cheie faciale și corporale combinate în timp real. De asemenea, plănuim să explorăm proceduri de învățare online și continuă pentru a îmbunătăți interacțiunea Om-IA bazată pe emoții, implicând un sistem dinamic care își actualizează continuu cunoștințele și comportamentul bazat pe interacțiuni în timp real.

7.6 Concluzie

Sistemul nostru de chat non-verbal în timp real reprezintă un avans semnificativ în interacțiunea Om-IA. Prin reducerea decalajului dintre comunicarea IA verbală și non-verbală, deschidem noi căi pentru crearea unor sisteme IA mai empatice, angajante și naturale, apropiindu-ne de obiectivul unor interfețe om-mașină cu adevărat intuitive prin artă.

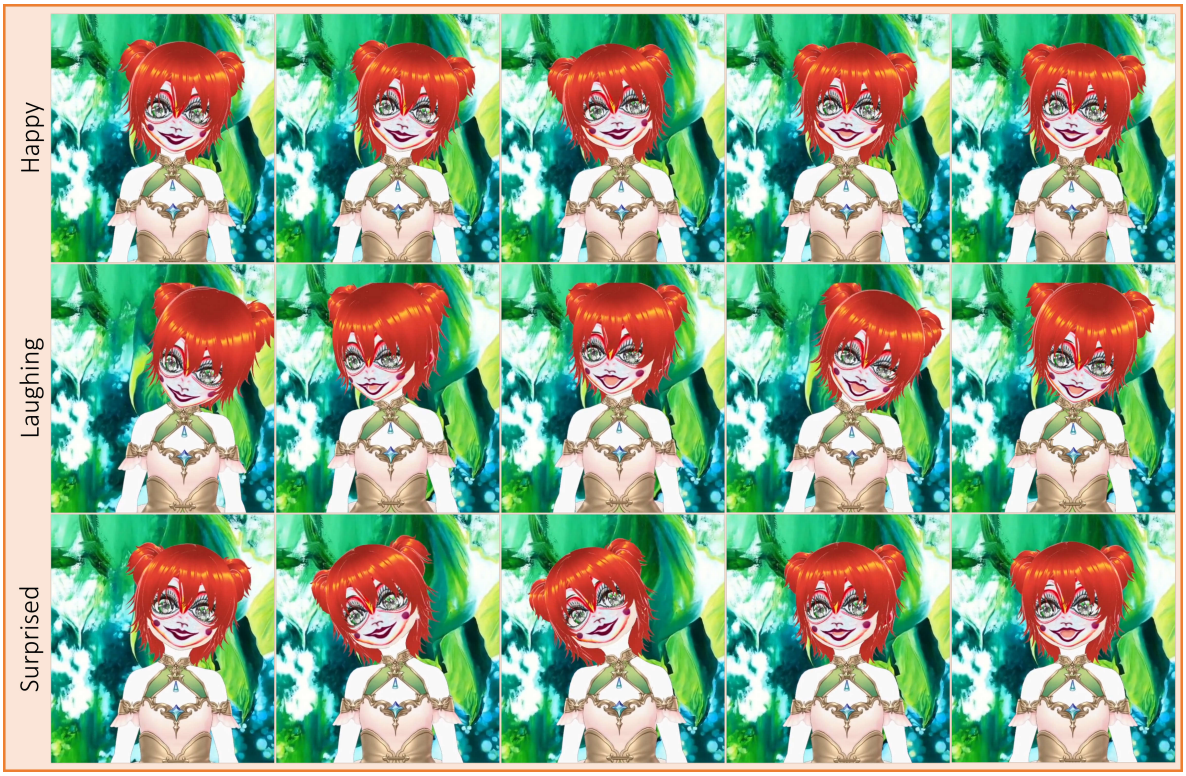

Figura 13: Maia în acțiune. Prezentăm rezultate vizuale ale metodei noastre PCA aplicată avatarului nostru 3D. În prezent, ne-am concentrat pe 3 emoții pozitive, dar poate fi extins la multe altele.

8 Concluzie

Această teză a explorat mai multe aspecte cheie ale construirii unor sisteme eficiente și robuste pentru înțelegerea scenelor aeriene, reconstrucția și sinteza de noi perspective utilizând vehicule aeriene fără pilot (UAV). Lucrarea acoperă mai multe domenii interconectate, inclusiv localizarea, segmentarea semantică, estimarea adâncimii, reconstrucția 3D și interacțiunea non-verbală om-AI. De-a lungul diferitelor capitole, am adus mai multe contribuții care avansează stadiul actual al tehnicii în aceste domenii.

8.1 Idei Cheie

Ideile de mai jos rezumă învățămintele cheie din teză și indică, de asemenea, direcții pentru cercetări viitoare în domeniul viziunii computaționale aeriene și roboticii:

1. **Sinergia Abordărilor Geometrice și Bazate pe Învățare** Combinarea tehnicilor geometrice clasice cu abordările moderne de învățare profundă a dus adesea la soluții robuste și eficiente. Această sinergie valorifică interpretabilitatea metodelor geometrice și flexibilitatea abordărilor bazate pe învățare. De exemplu, pipeline-ul nostru de sinteză a noilor perspective utilizează sculptura voxelilor multi-scală alături de un modul de randare neurală.

2. **Importanța Învățării Nesupervizate și Auto-Supervizate** Având în vedere provocările obținerii de date etichetate la scară largă pentru sarcini de viziune aeriană, tehnicile de învățare nesupervizată și auto-supervizată s-au dovedit valoroase. De la estimarea adâncimii la sinteza noilor perspective, am demonstrat cum să valorificăm coerența geometrică și temporală inerentă în date pentru a antrena modele fără adnotări manuale.

3. **Conectarea Simulării cu Realitatea** Am explorat tehnici pentru reducerea decalajului dintre datele de antrenament sintetice și implementarea în lumea reală. Lucrarea noastră privind estimarea zonelor sigure de aterizare și predicția adâncimii a arătat că seturile de date sintetice

atent proiectate, combinate cu tehnici de adaptare a domeniului, pot duce la modele care se generalizează la scenarii din lumea reală.

4. **Eficiență și Operare în Timp Real** O temă constantă a fost concentrarea pe dezvoltarea de metode care sunt eficiente din punct de vedere computațional și adecvate pentru operarea în timp real pe platforme embedded. Metodele rezultate, cum ar fi SafeUAV-Net și pipeline-ul nostru eficient de sinteză a noilor perspective, demonstrează că este posibil să se obțină performanțe competitive în limitele computaționale ale platformelor UAV.

5. **Importanța Reprezentărilor Multi-Scală** În diverse sarcini, de la detectarea drumurilor la reconstrucția 3D, am constatat că reprezentările multi-scală au fost utile în gestionarea gamei largi de scale prezente în imagistica aeriană. Abordarea noastră de sculptură voxel multi-scală pentru sinteza noilor perspective este un exemplu de aplicare a acestui principiu.

8.2 Concluzii

Această teză a explorat mai multe aspecte ale construirii de sisteme pentru înțelegerea scenelor aeriene, reconstrucție și sinteza noilor perspective utilizând UAV-uri. Lucrarea acoperă mai multe domenii interconectate, inclusiv localizarea, segmentarea semantică, estimarea adâncimii, reconstrucția 3D și interacțiunea non-verbală om-AI. **Capitolul 2: Localizarea din drumuri și intersecții în imagini aeriene** Am prezentat o metodă pentru geolocalizarea automată a imaginilor aeriene fără informații GPS, prin potrivirea drumurilor și intersecțiilor detectate cu date cartografice disponibile public. Abordarea noastră a obținut rezultate de localizare chiar și atunci când a fost antrenată pe imagini dintr-un oraș și testată pe altul [**?**]. Contribuțiile cheie ale acestui capitol includ:

- O metodă pentru geo-localizare automată în imagini aeriene fără informații GPS
- O rețea neurală convoluțională profundă cu stream dual local-global pentru detectarea drumurilor și intersecțiilor
- O procedură de aliniere geometrică pentru îmbunătățirea localizării și detectării drumurilor

Capitolul 3: Detectarea drumurilor și clădirilor în imagini aeriene Am introdus o rețea generativă adversarială cu salt dual (DH-GAN) pentru producerea segmentării pixelice a drumurilor și intersecțiilor simultan la două niveluri de interpretare [**?**]. Contribuțiile cheie ale acestui capitol includ:

- O arhitectură GAN cu salt dual pentru detectarea simultană a drumurilor și intersecțiilor
- O abordare de optimizare bazată pe netezire (SBO) pentru extragerea structurilor grafice ale drumurilor din segmentarea pixelică
- O evaluare pe un set de date de drumuri europene, demonstrând îmbunătățiri față de metodele anterioare

Capitolul 4: Către o înțelegere completă a lumii cu o dronă Am abordat două probleme pentru operarea UAV: estimarea zonelor sigure de aterizare și înțelegerea comprehensivă a scenei. Pentru estimarea zonelor sigure de aterizare, am propus SafeUAV-Net, o rețea neurală convoluțională pentru estimarea atât a adâncimii, cât și a zonei sigure de aterizare folosind input RGB [**?**]. Contribuțiile cheie includ:

- O arhitectură CNN potrivită pentru implementare embedded
- Un set de date sintetice pentru antrenarea modelelor de estimare a zonelor sigure de aterizare
- Rezultate atât pe date sintetice, cât și pe imagini reale capturate de drone

Pentru înțelegerea comprehensivă a scenei, am introdus modelul Neural Graph Consensus (NGC), o abordare pentru învățarea semi-supervizată multi-task [**?**]. Aspectele cheie ale acestei lucrări includ:

- O arhitectură bazată pe grafuri pentru învățarea multi-task
- O abordare pentru învățarea semi-supervizată prin consens

- Rezultate pe multiple sarcini de înțelegere a scenei, inclusiv estimarea adâncimii, segmentarea semantică și estimarea poziției

Capitolul 5: Către construirea eficientă a structurii 3D Am introdus două contribuții în domeniul estimării adâncimii din imagistică monoculară: În primul rând, am propus UFODepth, o metodă pentru învățarea nesupervizată a estimării adâncimii metrice dintr-o singură imagine în contextul videoclipurilor neconstrânse capturate de UAV-uri [**?**]. Inovațiile cheie includ:

- O procedură de optimizare bazată pe flux pentru corectarea măsurătorilor de odometrie cu zgomot
- O metodă analitică de estimare a adâncimii bazată pe fluxul optic și vitezele corectate ale camerei
- O arhitectură de învățare nesupervizată care încorporează atât constrângeri analitice, cât și bazate pe date

În al doilea rând, am explorat o abordare de distilare a adâncimii care combină multiple căi complementare pentru estimarea adâncimii metrice [**?**]. Aspectele cheie includ:

- O abordare de ansamblu care combină estimarea adâncimii analitică și bazată pe învățare
- O procedură de distilare pentru a antrena o rețea studentă compactă
- Generalizare la scene noi și performanță pe hardware embedded

Capitolul 6: Îmbinarea reconstrucției 3D și sintezei noilor perspective Am abordat problema sintezei noilor perspective pentru scene aeriene la scară largă. Am adus două contribuții în acest domeniu: În primul rând, am introdus o abordare auto-supervizată pentru sinteza noilor perspective 2D a scenelor la scară largă capturate de UAV-uri [**?**]. Inovațiile cheie includ:

- O metodă geometrică multi-scală bazată pe sculptura voxelilor
- Un modul de randare neurală auto-supervizat pentru rafinarea reconstrucției geometrice
- Performanță pe date din lumea reală cu informații de poziție zgomotoase

În al doilea rând, am extins această lucrare la o abordare ciclică neuro-analitică care rafinează iterativ atât sinteza noilor perspective 2D, cât și calitatea reconstrucției 3D [**?**]. Aspectele cheie ale acestei lucrări includ:

- O procedură de rafinare ciclică care alternează între reconstrucția geometrică și neurală
- Performanță îmbunătățită atât pentru sinteza noilor perspective, cât și pentru sarcinile de reconstrucție 3D
- Gestionarea scenelor la scară largă și a datelor zgomotoase din lumea reală

Capitolul 7: Dincolo de Scene Virtuale Statice: Un Chat Non-Verbal în Timp Real pentru Interacțiunea Om-AI Am explorat o direcție în interacțiunea om-AI propunând un sistem de chat non-verbal în timp real bazat pe expresii faciale. Am introdus Maia, un avatar animat capabil să răspundă la emoțiile umane în timp real folosind indicii vizuale [**?**]. Contribuțiile cheie ale acestei lucrări includ:

- Un set de date de expresii non-verbale pentru antrenarea modelelor de recunoaștere și generare a emoțiilor
- Abordări pentru generarea răspunsurilor expresive, inclusiv generarea de imagini 2D cu modele de difuzie și animație facială și corporală 3D
- Un cadru de evaluare care compară evaluarea umană și metricile automate

Această lucrare deschide posibilități pentru interacțiuni AI, cu potențiale aplicații în educație, terapie și expoziții interactive.

8.3 Lucrări viitoare

Deși această teză a adus contribuții în diverse aspecte ale înțelegerii scenelor aeriene și reconstrucției, există numeroase direcții pentru cercetări viitoare care ar putea construi pe baza acestei lucrări și o pot extinde. Prezentăm mai jos câteva direcții promițătoare pentru investigații viitoare în diferitele domenii explorate în teză. **Geolocalizare și Cartografiere**

- Fuziune multi-modală: Încorporarea senzorilor suplimentari (IMU-uri, GPS de cost redus) pentru îmbunătățirea acurateței.
- Coerență temporală: Utilizarea secvențelor video pentru o mai bună stabilitate în zonele dificile.
- Actualizarea adaptivă a hărților: Dezvoltarea algoritmilor online pentru rafinarea continuă a hărților.
- Îmbogățire semantică: Extinderea detectării pentru a include clădiri, vegetație și corpuri de apă.

Segmentare Semantică și Extragerea Drumurilor

- Învățare multi-task: Integrarea detectării clădirilor și clasificării utilizării terenurilor.
- Învățare slab supervizată: Utilizarea datelor cartografice zgomotoase pentru antrenare la scară mai largă.
- Coerență temporală: Încorporarea constrângerilor pentru secvențe video.
- Atribute de drum fine: Detectarea benzilor, tipurilor de drumuri și condițiilor de trafic.

Estimarea Zonelor Sigure de Aterizare

- Fuziune multi-senzor: Integrarea LiDAR și camerelor termice pentru robustețe.
- Detectarea obstacolelor dinamice: Urmărirea vehiculelor și pietonilor în mișcare.
- Înțelegere semantică: Diferențierea între tipurile de suprafețe pentru aterizare.
- Învățare prin întărire: Dezvoltarea politicilor end-to-end în medii simulate.

Estimarea Adâncimii

- Coerență multi-vedere: Impunerea coerenței între mai multe cadre.
- Optimizare comună: Optimizarea adâncimii, poziției camerei și structurii 3D împreună.
- Eșantionare adaptivă: Concentrarea resurselor pe regiunile informative.
- Transfer de învățare: Dezvoltarea tehnicilor pentru adaptarea rapidă la medii noi.

Sinteza Noilor Perspective

- Scene la scară largă: Dezvoltarea abordărilor pentru reconstrucția la scara orașului.
- Scene dinamice: Gestionarea obiectelor în mișcare și a iluminării schimbătoare.
- Sinteză semantică: Încorporarea informațiilor semantice pentru îmbunătățirea realismului.
- Randare în timp real: Avansarea către sinteza în timp real pe platforme embedded.

Interacțiunea Om-AI

- Interacțiune multimodală: Încorporarea gesturilor, poziției și semnalelor fiziologice.
- Personalizare: Adaptarea la stilurile individuale ale utilizatorilor în timp.
- Răspunsuri contextuale: Luarea în considerare a istoricului conversației și factorilor de mediu.
- Considerații etice: Abordarea preocupărilor legate de confidențialitate și potențialul abuz.

Teme Transversale

- Estimarea incertitudinii: Dezvoltarea tehnicilor pentru propagarea robustă a incertitudinii.
- Învățare continuă: Investigarea metodelor pentru adaptarea continuă fără uitare.
- Explicabil CopyRetryClaude's response was limited as it hit the maximum length allowed at this time.Iplease continueEditContinuarea traducerii în română:
- Explicabilitate: Crearea unor sisteme complexe mai interpretabile pentru utilizatori.
- Benchmarking: Crearea unor seturi de date de evaluare mai cuprinzătoare și realiste.
- Eficiență energetică: Dezvoltarea algoritmilor optimizați pentru bugete de putere limitate.

Abordarea acestor provocări va necesita colaborare interdisciplinară între domeniile viziunii computaționale, roboticii și învățării automate. Progresele în aceste domenii au potențialul de a permite aplicații transformative ale UAV-urilor în monitorizarea urbană, răspunsul la dezastre și multe altele.

Bibliografie

[1] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 5007–5015. IEEE, 2015.

[2] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[3] Helmut Mayer, Stefan Hinz, Uwe Bacher, and Emmanuel Baltsavias. A test of automatic road extraction approaches. *International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences*, 36(3):209–214, 2006.

[4] Yucong Lin and Srikanth Saripalli. Road detection and tracking from aerial desert imagery. *Journal of Intelligent & Robotic Systems*, 65(1-4):345–359, 2012.

[5] Ivan Laptev, Helmut Mayer, Tony Lindeberg, Wolfgang Eckstein, Carsten Steger, and Albert Baumgartner. Automatic extraction of roads from aerial images based on scale space and snakes. *Machine Vision and Applications*, 12(1):23–31, 2000.

[6] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *Computer Vision–ECCV 2010*, pages 210–223. Springer, 2010.

[7] 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.

[8] Alina Elena 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.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[10] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[12] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[13] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022.

[14] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Kostas Karpouzis and Georgios N Yannakakis. *Emotion in games*. Springer, 2016.

[16] Angelo Cafaro, Hannes Högni Vilhjálmsson, and Timothy Bickmore. First impressions in human–agent virtual encounters. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(4): 1–40, 2016.

[17] Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620 (7976):982–987, 2023.

[18] Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602(7896): 223–228, 2022.

[19] Dan Klang. Automatic detection of changes in road data bases using satellite imagery. *International Archives of Photogrammetry and Remote Sensing*, 32:293–298, 1998.

[20] Armin Gruen and Haihong Li. Road extraction from aerial and satellite images by dynamic programming. *ISPRS Journal of Photogrammetry and Remote Sensing*, 50(4):11–20, 1995.

[21] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[22] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007.

[23] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 255–268. Springer, 2010.

[24] Fernando Caballero, Luis Merino, Joaquín Ferruz, and Aníbal Ollero. Unmanned aerial vehicle localization based on monocular vision and online mosaicking. *Journal of Intelligent and Robotic Systems*, 55(4-5):323–343, 2009.

[25] Xiaoming Li. A software scheme for uav's safe landing area discovery. *AASRI Procedia*, 4:230–235, 2013.

[26] Sumair Aziz, Rao Muhammad Faheem, Mudassar Bashir, Adnan Khalid, and Amanullah Yasin. Unmanned aerial vehicle emergency landing site identification system using machine vision. *Journal of Image and Graphics*, 4(1):36–41, 2016.

[27] Timo Hinzmann, Thomas Stastny, Cesar Cadena Lerma, Roland Siegwart, and Igor Gilitschenski. Free lsd: Prior-free visual landing site detection for autonomous planes. *IEEE Robotics and Automation Letters*, 2018.

[28] Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1424–1445, 2000.

[29] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.

[30] Gellert Mattyus, Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Enhancing road maps by parsing aerial images around the world. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[31] Jonghyuk Kim and Salah Sukkarieh. Real-time implementation of airborne inertial-slam. *Robotics and Autonomous Systems*, 55(1):62–71, 2007.

[32] Fernando Caballero, Luis Merino, Joaquin Ferruz, and Aníbal Ollero. Vision-based odometry and slam for medium and high altitude flying uavs. *Journal of Intelligent and Robotic Systems*, 54 (1-3):137–161, 2009.

[33] Alina Marcu and Marius Leordeanu. Dual local-global contextual pathways for recognition in aerial imagery. *arXiv preprint arXiv:1605.05462*, 2016.

[34] Dragos Costea, Alina Marcu, Emil-Ioan Slusanschi, and Marius Leordeanu. Creating roadmaps in aerial images with generative adversarial networks and smoothing-based optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2100–2109, 2017.

[35] Per Enge Frank van Diggelen. The world's first gps mooc and worldwide laboratory using smartphones. *Proceedings of the 28th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2015)*, pages 361 – 369, 2015.

[36] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[37] Alina Marcu and Marius Leordeanu. Object contra context: Dual local-global semantic segmentation in aerial images. *AAAI Workshops*, 2017. URL https://www.aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15177/14658.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[39] 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.

[40] 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.

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

[42] 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.

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

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[45] Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan L Yuille. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3205–3214, 2019.

[46] Yuan Gao, Haoping Bai, Zequn Jie, Jiayi Ma, Kui Jia, and Wei Liu. Mtl-nas: Task-agnostic neural architecture search towards general-purpose multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11543–11552, 2020.

[47] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[48] Jan Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J Davison. Deepfactors: Real-time probabilistic dense monocular slam. *IEEE Robotics and Automation Letters*, 5(2):721–728, 2020.

[49] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021.

[50] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (TOG)*, 39(4):71–1, 2020.

[51] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. *arXiv preprint arXiv:2107.02791*, 2021.

[52] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *ICCV*, 2021.

[53] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3838, 2019.

[54] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019.

[55] 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.

[56] 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 (CVPR) Workshops*, pages 3215–3223, June 2021.

[57] 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.

[58] 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.

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

[60] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[61] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Rife: Real-time intermediate flow estimation for video frame interpolation. *arXiv preprint arXiv:2011.06294*, 2020.

[62] 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.

[63] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[64] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

[65] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *CVPR*, 2023.

[66] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.

[67] Ruilong Li, Sanja Fidler, Angjoo Kanazawa, and Francis Williams. Nerf-xl: Scaling nerfs with multiple gpus, 2024.

[68] Teppei Suzuki. Fed3DGS: Scalable 3D Gaussian Splatting with Federated Learning. *arXiv preprint arXiv:2403.11460*, 2024.

[69] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[70] Carsten Griwodz, Simone Gasparini, Lilian Calvet, Pierre Gurdjos, Fabien Castan, Benoit Maujean, Gregoire De Lillo, and Yann Lanthony. Alicevision meshroom: An open-source 3d reconstruction pipeline. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pages 241–247, 2021.

[71] Haiyu Zhao, Yuanbiao Gou, Boyun Li, Dezhong Peng, Jiancheng Lv, and Xi Peng. Comprehensive and delicate: An efficient transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14122–14132, 2023.

[72] 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.

[73] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks, 2020.

[74] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12922–12931, June 2022.

[75] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[76] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023.

[77] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022.

[78] Fridovich-Keil and Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022.

[79] Alexandra Budisteanu, Dragos Costea, Alina Marcu, and Marius Leordeanu. Self-supervised novel 2d view synthesis of large-scale scenes with efficient multi-scale voxel carving, 2023.

[80] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

[81] Otto Seiskari, Jerry Ylilammi, Valtteri Kaatrasalo, Pekka Rantalankila, Matias Turkulainen, Juho Kannala, and Arno Solin. Gaussian splatting on the move: Blur and rolling shutter compensation for natural camera motion, 2024.

[82] Nozomi Isozaki, Shigeyoshi Ishima, Yusuke Yamada, Yutaka Obuchi, Rika Sato, and Norio Shimizu. Vroid studio: a tool for making anime-like 3d characters using your imagination. In *SIGGRAPH Asia 2021 Real-Time Live!*, pages 1–1. ACM, 2021.

[83] Vseeface. https://www.vseeface.icu/, 2023. Accessed: 2023-08-03.

[84] OpenAI. Gpt-4. https://openai.com/gpt-4, 2024. Accessed: 2024-08-25.

[85] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[86] L. Young et al. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2303.16817*, 2023.

[87] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[88] L. Zhang, Maneesh Agrawala, Frédo Durand, and X. Zhou. Controlnet: Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.