RAPORT STIINTIFIC

perioada ianuarie 2021 – decembrie 2021

Proiect PN-III-PCCF-0180

Perceptie Vizuala Semantica si Control Integrat pentru Sisteme Autonome

Rezumatul etapei

In cadrul activitatilor de cercetare desfasurate la Institutul de Matematica al Academiei Romane s-a continuat aprofundarea problematicii reconstructiei 3D profunde a persoanelor prin investigarea unor modele geometrice structurate profunde cu semantica asociata (T1.2). Mai precis, s-a realizat modelarea 3D a auto-contactului asa cum apare in imagini monoculare ale persoanelor, realizare care faciliteaza o analiza fina a scenelor 3D precum si intelegera mai nuantata a comportamentelor. De asemenea, s-a realizat un sistem automat care poate reconstrui miscarea si postura 3D a corpului uman in scenarii diverse, atat in interior cat si afara, cu aplicare particulara pentru practici de antrenament. Un pas semnificativ in intelegerea mai profunda a unei scene 3D l-a constituit realizarea unei reconstructii tridimensionale a unor oameni aflati in interactiune asa cum apar in imagini monoculare. Metoda este cu atat mai importanta cu cat modelele actuale de reconstructie 3D a persoanelor fie le trateaza in izolare, ignorand contextul in care apar, fie reconstruiesc mai multe persoane fara a fi capabile sa recupereze corect interactiunile dintre ei, mai ales cand oamenii sunt aproape unul de celaltalt.

La IMAR au continuat activitatile de investigare a învățarii active vizuale încorporate, în care un agent este setat să exploreze un mediu 3D cu scopul de a dobândi înțelegere vizuală a scenei prin selectarea activă a view-urilor („vederilor”) pentru care să solicite adnotări. Cercetarile s-au desfasurat in cadrul work package-ului A3 destinat controlului semantic optimal. Tot in cadrul aceluiasi work package, mai exact in directia explorarii tehnicilor de control optimal direct si invers (T3.1), s-a obtinut generarea automata de scenarii critice dpdv al sigurantei pentru testarea vehiculelor autonome care evita folosirea metodelor euristice pentru comportamentul persoanelor, recurgand alternativ la propunerea unor modele pentru miscarea persoanelor obtinute prin reformularea problemei ca o problema de invatare a locului in care se pot plasa persoanele astfel incat scenariile rezultate sa fie predispuse pentru coliziunea persoanei cu un vehicul autonom.

In cadrul Universitatii Tehnice din Cluj-Napoca au continuat cercetarile in cadrul work package-urilor A1, A2 si A4, conform planului. Mai exact, echipa condusa de dl profesor Nedevschi a obtinut o serie de rezultate dezvoltand o metoda de invatare nesupervizata a profunzimii in doua etape, care nu necesita ground truth, o metoda de estimare a profunzimii bazata pe informatie semantica, precum si alte metode de estimare a profunzimii descrise in cele ce

urmeaza. In privinta task-urilor legate de recunoasterea si localizarea vizuala (work package A2), UTCN a dezvoltat o serie de modele semantice semi-supervizate cu componente multiple si raspunsuri partiale detaliate de asemenea in raport. In ceea ce priveste optimizarea si integrarea sistemelor, UTCN a realizat progrese in extinderea setului de date geometrice WildUAV, a creat un set de date aeriene adnotate semantic din date reala pentru monitorizarea padurilor si a creat harti 3D imbogatite semantic folosind capabilitati de augmentare a setului de date WildUAV. Au fost adaptate si alte seturi de date pentru solutii de segmentare semantica, detectie si estimarea profunzimii. Folosind date intr-un simulator, s-a dezvoltat o metodologie de estimare a ariilor defrisate. In privinta integrarii navigatiei vizuale si a intelegerii scenei 3D, UTCN a realizat un sistem avansat de perceptie cu acoperire de 360 de grade proiectat pentru a fi operat in timp real la bordul unui vehicul autonom. In plus, s-a adresat problema navigarii autonome a unei drone cu evitarea obstacolelor dintr-o padure, pe baza unei traiectorii de puncte 3D din mediu pe care o primeste sistemul.

Toate aceste rezultate sunt prezentate in continuare pe scurt, cu evidentierea work package-ului (respectiv, task-ului) de care apartin, insotite de referintele la articolele care faciliteaza accesul descrierea lor detaliata.

A1. Reconstructia 3D profunda (IMAR)

Task 1.2: Modele geometrice structurate profunde cu semantica asociata

Într-o lucrare publicata la AAAI-21, autorii Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata (Marinoiu), Alin-Ionut Popa, Vlad Olaru si Cristian Sminchisescu isi propun sa estimeze auto-contactul 3d în imagini monoculare ale persoanelor, task fundamental pentru analiza fina a scenelor 3d sau și pentru înțelegerea comportamentelor. Metodele curente de reconstructie 3d nu se concentreaza pe regiunile corpului uman aflate în auto-contact și adesea recupereaza configuratii care sunt fie departe una de cealaltă, fie se intersectează, când de fapt regiunile implicate doar se ating. Acest lucru conduce la reconstructii incorecte, nerealiste și previne înțelegerea fina a comportamentelor. Articolul abordeaza aceste probleme și propune mai multe contributii: (1) un model pentru predictia auto-contactului, care estimeaza segmentarea și semnatura lui și care profita de (2) suportul spatial din imaginea auto-contactului propus de autori, atat în timpul antrenamentului modelui cât și la testare; (3) s-au colectat doua setul de date de dimensiuni mari pentru a sprijini invatarea și evaluarea metodei prezentare, și anume HumanSC3D, un set de date cu acuratete mare a capturii miscarii 3d cu 1032 de secvente continand 5058 de evenimente de contact și 1,246,487 de posturi 3d groud truth sincronizate cu imagini inregistrare din mai multe perspective (multiple views), precum și FlickrSC3D, un set de date de 3969 de imagini cu 25297 corespondente de suprafete în contact. Mai mult, (4) lucrarea demonstrează felul în care reconstructii 3d mult mai expresive pot fi obtinute folosind semnaturile de autocontract drept constrangeri și (5) prezinta detectia monoculara a

atingeri fetei (auto-contact mana-fata) ca una dintre multiplele aplicații relevante (mai ales în contextul actual epidemic) care devin posibile datorită predictiei auto-contactului. Exemple de rezultate calitative ale metodei noastre sunt prezentate în figura de mai jos:

Fig. Reconstructia posturii și a formei corpului uman folosind adnotarile noastre de autocontact. Stânga: imaginea originala. Centru: Reconstructia fără considerarea autocontractului și a functiei de loss asociate. Dreapta: Reconstructia care folosește adnotarile de auto-contact precum și functia corespunzatoare de loss

Referinta

M. Fieraru, M. Zanfir, E. Oneata, A. Popa, V. Olaru, C. Sminchisescu, “Learning Complex 3D Human Self-Contact”, Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)

O lucrare publicata la CVPR 2021 analizează felul în care un sistem automat poate reconstrui mișcarea și postura 3d a corpului uman pentru a o integra în bune practici de antrenament și pentru a identifica erorile de exercițiu și a furniza din timp feedback util. Sistemul poate fi folosit în interior sau afara. El poate reconstrui postura 3d și mișcarea umana, poate segmenta

în mod sigur repetitiile exercitiilor și poate identifica în timp real deviatiile intre executia standard invatata de la antrenori și executia unui elev. Drept urmare, este posibila oferirea unui feedback cantitativ și localizat pentru executia corecta a exercitiilor, reducerea riscului de accidentare și imbunatatirea continua a performantelor. Pentru a sprijini cercetarea și evaluarea, articolul introduce primul set de date pe scara larga, Fit3D, care conține peste 3 milioane de imagini și configuratiile corespunzatoare ale formei 3d a corpului și miscarii, cu peste 37 de exerciții repetate, acoperind toate grupele majore de muschi, exerciții executate atât de instructori cât și de elevi. Articolul evalueaza calitatea reconstructiei și propune modele de analiza și feedback. Antrenorul statistic utilizeaza un parametru global care captureaza gradul de critica pe care îl impune în analizarea performantei elevului pentru adaptarea la nivelul de antrenament al elevului (incepator, avansat, expert) sau la acuratetea dorita de la metoda de reconstructie 3d. Pentru diferite valori ale parametrului global, sistemul de feedback bazat pe estimari ale posturii 3d atinge o buna acuratete comparativ cu sistemul care folosește informația ground-truth de captura a miscarii. Antrenorul statistic ofera elevului feedback în limbaj natural, cu ancorare vizuala (în imagini) a suportului spatial și temporal al feedback-ului. Exemple calitative pot fi examinate în figura următoare:

Fig. feedback-ul textual și vizual produs pe videoclipuri din lumea reală, surprinse cu o cameră obișnuită a unui smartphone. Pentru fiecare exemplu, sunt prezentate: o imagine cu eroarea identificată a elevului (rândul de sus), reconstrucția 3d a elevului (al doilea rând), imaginea corespunzătoare cu execuția corectă a instructorului (randul trei) și feedback textual (rândul de jos). Cele două exemple din (stânga) arată feedback despre caracteristicile active, în timp ce cele două din (dreapta) arată feedback despre caracteristicile pasive. Se observa generalizarea la diferiți oameni în diferite medii și puncte de vedere ale camerei.

Referinta

Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, Cristian Sminchisescu. AIFit: Automatic 3D Human-Interpretable Feedback Models for Fitness Training. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2021

Reconstrucția tridimensională a mai multor oameni care interacționează data fiind o imagine monoculară este crucială pentru problema generală de înțelegere a scenei, deoarece surprinderea subtilităților interacțiunii este adesea motivul realizării unei fotografii. Metodele actuale de reconstrucție umană 3D fie tratează fiecare persoană în mod independent, ignorând cea mai mare parte a contextului, fie reconstruiesc oamenii împreună, dar nu pot recupera corect interacțiunile atunci când oamenii sunt în imediata apropiere. Într-un articol publicat la NeurIPS 2021, Mihai Fieraru, Mihai Zanfir, Teodor Alexandru Szente, Eduard Gabriel Bazavan, Vlad Olaru și Cristian Sminchisescu au prezentat REMIPS, un model pentru reconstrucția 3D a mai multor persoane care interacționează folosind modelare cu supervizare slabă (weaksupervision). REMIPS poate reconstrui un număr variabil de persoane direct din imagini monoculare. La baza metodologiei prezentate se află o nouă rețea de tip Transfomer care combină tokeni neordonati pentru persoane (unul pentru fiecare om detectat) cu token-uri codificate pozițional din patch-uri cu caracteristicile imaginii. Articolul introduce un nou model unificat pentru auto-coliziuni ca și pentru coliziuni de întrepătrundere bazat pe o aproximare a suprafetei 3D a corpului calculată prin aplicarea unor operatori de decimare. Metoda se bazeaza pe funcții de loss autosupervizate (self-supervized) pentru o mai buna flexibilitate și generalizare în medii reale și încorporeaza functiile de loss pentru auto-contact și interacțiuni direct în procesul de învățare. REMIPS genereaza rezultate cantitative state-of-the-art pe benchmark-uri obișnuite chiar și în cazurile în care nu este utilizată supervizare 3D. În plus, rezultatele vizuale calitative arată că reconstrucțiile sunt plauzibile în ceea ce privește postura și forma și coerente pentru imagini provocatoare, colectate în medii reale, unde oamenii interacționează adesea. Rezultate calitative ale metodei pot fi observate în figura următoare:

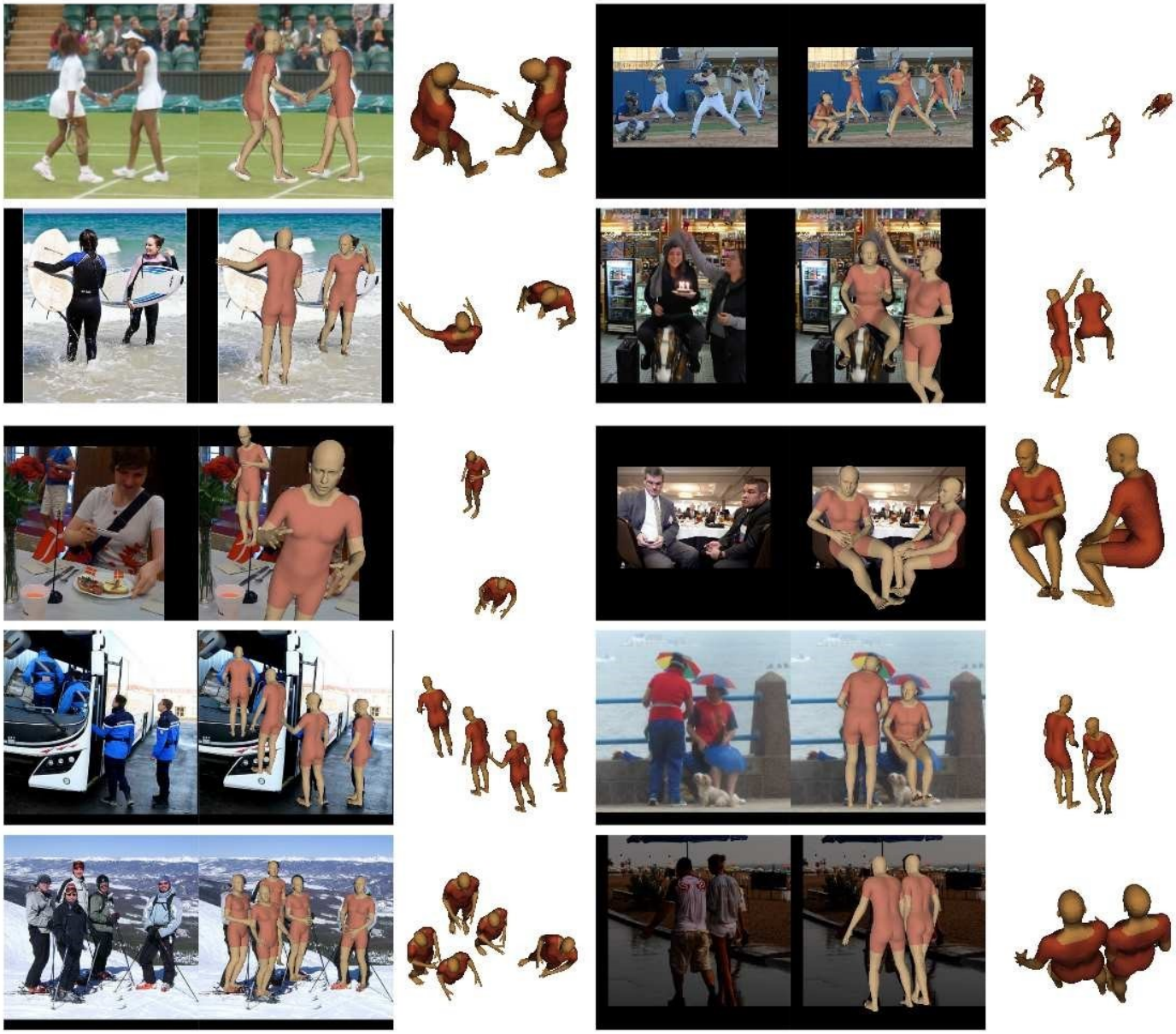

Fig. Predictii 3D ale posturii și formei corpului uman pe setul de date COCO (liniile 1-5) pentru imagini din lumea reala. Este aratata imaginea initiala împreuna cu reconstructiile mesh-urilor suprapuse peste imagine precum și redarea dintr-un alt unghi pentru a ilustra și mai bine consistenta fizica a reconstructiilor REMIPS.

Referinta

Mihai Fieraru, Mihai Zanfir, Teodor Alexandru Szente, Eduard Gabriel Bazavan, Vlad Olaru, Cristian Sminchisescu. REMIPS: Physically Consistent 3D Reconstruction of Multiple Interacting People under Weak Supervision. Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)

A1. Reconstrucție 3D profundă (UTCN)

1.1. Modele profunde structurate geometric cu informație semantică (A1.2)

1.1.1 Folosirea pseudo-adnotărilor pentru învățarea nesupervizată a profunzimii dintr-o singură imagine (A1.2.1)

În articolul "**Pseudo-annotation based unsupervised monocular depth estimation**", submis spre acceptare conferinței CVPR 2022 [11], propunem o nouă metodă de învățare nesupervizată a profunzimii în două etape, care nu necesită ground truth. În prima etapă, rețeaua se antrenează în mod nesupervizat pe imagini de rezoluție mare, cu funcția de pierdere fotometrică. Se folosesc 3 cadre consecutive și se învață mișcarea camerei și profunzimea pentru cadrul din mijloc. Pentru a învăța profunzimea, se folosesc modele de reproiecție 3D între cadrele vecine și cadrul curent. Imaginea curentă este reconstruită folosind cadrele vecine, iar o funcție de pierdere fotometrica este optimizată. Apoi, folosind această rețea se generează pseudo-adnotări. În al doilea pas, rețeaua se antrenează în mod supervizat folosind pseudo-adnotările. În metodele nesupervizate, exista problema ambiguității scalei profunzimii, care se rezolvă de obicei folosind profunzimea ground truth. Pentru a rezolva această problemă, propunem scalarea pseudo-profunzimii printr-o metoda automată. De asemenea, realizăm o filtrare a pseudo-profunzimii bazată pe consistența 3D între cadre consecutive, pentru a elimina din erori. Toate experimentele sunt realizate pe setul de date KITTI.

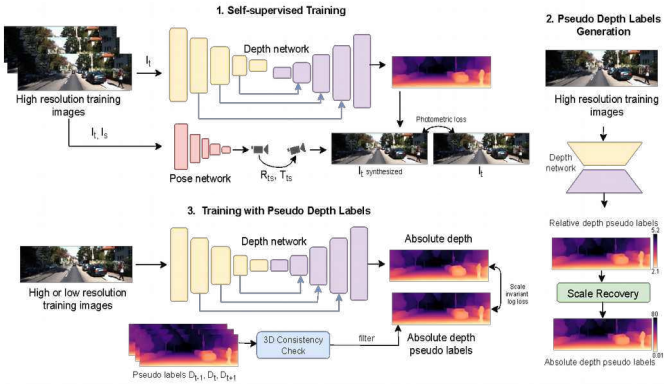

Fig. 1. Metoda de învățare a profunzimii pentru imagini monoculare.

Tabel 1. Comparație cu stadiul actual al cercetării pe setul de date KITTI. Metoda propusă obține cele mai bune rezultate în cazul producerii unor imagini de rezoluție medie, și se afla pe locul 2 după ManyDepth în cazul rezoluțiilor mari. ManyDepth însă folosește două cadre consecutive la inferență, iar metoda propusă folosește doar cadrul curent.

Method	Test frames	Backbone	Sem	Resolution	AbsRel ↓	SqRel ↓	RMS ↓	RMSlog ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
GeoNet [48]	1	ResNet-50	no	192 × 640	0.153	1.328	5.737	0.23	0.802	0.934	0.972
DF-Net [52]	1	ResNet-50	no	192× 640	0.146	1.182	5.215	0.213	0.818	0.943	0.978
Guizilini et al. [19]	1	ResNet-50	yes	192 × 640	0.113	0.831	4.663	0.189	0.878	0.971	0.983
SGDepth [25]	1	ResNet-50	yes	192 × 640	0.112	0.833	4.688	0.190	0.884	0.961	0.981
Monodepth2 (baseline) [15]	1	ResNet-50	no	192 × 640	0.110	0.831	4.642	0.187	0.883	0.962	0.982
Jung et al. [24]	1	ResNet-50	yes	192 × 640	0.102	0.675	4.393	0.178	0.893	0.966	0.984
Ours	1	ResNet-50	no	192 × 640	**0.100**	**0.661**	**4.264**	**0.172**	**0.896**	**0.967**	**0.985**
Monodepth2 (baseline) [15]	1	ResNet-50	no	320 × 1024	0.110	0.831	4.642	0.187	0.883	0.962	0.982
Shu et al. [40]	1	ResNet-50	yes	320 × 1024	0.104	0.729	4.481	0.179	0.893	0.965	**0.984**
ManyDepth [45]	2 (-1, 0)	ResNet-50	no	320 × 1024	**0.091**	**0.694**	**4.245**	**0.171**	**0.911**	**0.968**	**0.983**
Ours	1	ResNet-50	no	320 × 1024	0.098	0.721	4.415	0.180	0.890	0.965	**0.984**

1.1.2 Estimarea monoculară a profunzimii bazată pe informația semantică (A1.2.2)

Am studiat efectul a trei posibile integrări ale semanticii în procesul de estimare a profunzimii: 1) *Antrenarea în comun a ambelor sarcini*. Profunzimea și trăsăturile semantice au mai multe aspecte comune. Ambele folosesc structuri de rețea similare (codor-decodor, extractoare de

caracteristici multi-scale) prin utilizarea unor componente similare (convoluții dilatate, legături reziduale). Aceste caracteristici pot fi combinate într-un singur extractor de caracteristici, care este apoi trecut la un decodor. O pierdere multi-sarcină este apoi folosită în peste rețeaua neuronală, combinând pierderea bazată pe entropia încrucișată la nivel de segmentare cu o pierdere bazată pe MDE (în munca noastră am folosit și o pierdere bazată pe clasificare pentru MDE).

2) *Fuziunea timpurie a caracteristicilor semantice și de profunzime*. În acest caz, am folosit diferite extractoare de caracteristici: unul de segmentare, unul pentru estimarea adâncimii. Deoarece scopul principal al acestei lucrări este obținerea unei adâncimi precise (nu neapărat o hartă semantică mai bună), folosim un extractor de caracteristici bazat pe semantică pre-antrenat, care are doar rolul de a oferi cunoștințe de nivel înalt pentru MDE. Cele două hărți de caracteristici generate sunt apoi combinate în a doua parte a rețelei, cu scopul de a converge către o profunzime mai bună (utilizăm aici doar o singură pierdere MDE).

3) *Fuziunea târzie a caracteristicilor semantice și de profunzime*. În al treilea caz sunt utilizate două rețele separate pentru segmentare și estimarea adâncimii. Rețeaua semantică este antrenată separat, oferind harta semantică finală ca ghid pentru învățarea în profunzime. Cunoștințele de nivel înalt ale obiectului sunt utilizate în mecanismul de învățare, obținând o hartă de profunzime separată pentru fiecare clasă de obiecte. Astfel, în acest caz informația semantică este utilizată doar în funcția de pierdere.

1.1.3 Discretizare adaptabilă pentru estimare robustă a profunzimii (A1.2.3)

Majoritatea algoritmilor MDE regresează direct o hartă de profunzime (fracțională) folosind o pierdere bazată pe MSE (mean-square error), care modifică iterativ valoarea estimată conform valorilor de profunzime din datele de antrenare. Lucrările recente au arătat că se obțin rezultate mai bune atunci când problema MDE este văzută mai degrabă ca o clasificare decât ca o regresie. Scenariilor aeriene le lipsește o structură clară în scenă (de exemplu, scenarii rutiere sau în interiorul clădirilor), în care obiectele sunt ordonate în mod previzibil [1], așa că o discretizare specifică scenariului este mai greu de prezis. Propunem crearea unei estimări inițiale a distribuției punctului de adâncime și apoi folosirea acestei distribuții de adâncime pentru a genera o discretizare adecvată. Astfel, scopul principal este de a deduce o histogramă de profunzime, folosind pierderea KL-Divergence ca mecanism de învățare.

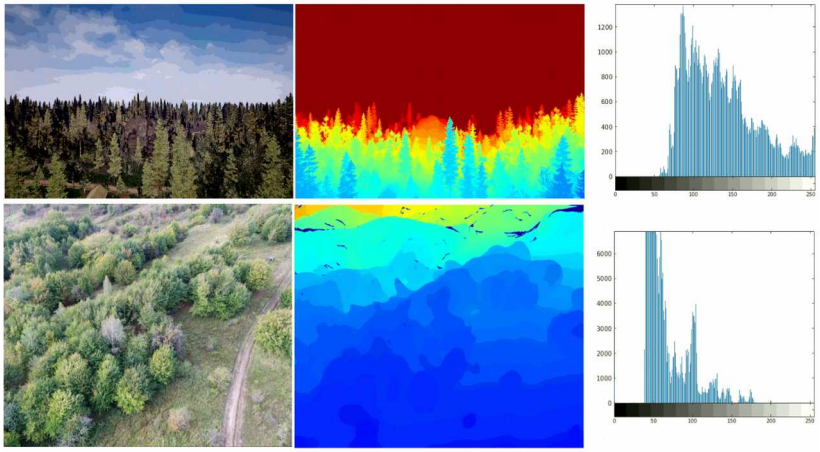

Fig. 2. Histograma estimărilor de profunzime pentru scenarii aeriene;Sus: imagini sintetice, Jos: imagini reale.

1.1.4 Experimente de antrenare a soluțiilor de estimare a profunzimii în context aerian (A1.2.4)

În articolul "**WildUAV: Monocular UAV Dataset for Depth Estimation Tasks**" au fost prezentate rezultatele antrenării mai multor sisteme de estimare a profunzimii bazate pe învățare profundă folosind setul de date WildUAV. Au fost testați patru algoritmi supervizați: un model simplu bazat pe regresie, un model bazat pe clasificare (adâncimea fiind discretizată în mai multe clase), un model de regresie ordinală (set discretizat de valori ordonate pentru adâncime) precum și un model propus special pentru scenarii aeriene (ce combină clasificarea clasică cu cea ordinală), dezvoltat în cadrul acestui proiect și publicat în articolul "**Monocular Depth Estimation With Improved Long-Range Accuracy for UAV Environment Perception**" [1]. Alternativ, au fost testați trei algoritmi nesupervizați (self-supervised), întrucât flexibilitatea acestora, ce derivă din lipsa nevoii pentru imagini de profunzime în timpul antrenării, reprezintă un avantaj important în anumite aplicații.

Performanța sistemelor supervizate a fost comparată între antrenarea pe date aeriene sintetice și antrenarea pe datele reale WildUAV, rezultatul favorizând cel de-al doilea caz. Mai mult, sistemul special dezvoltat pentru scenariul aerian a fost cel mai performant, și în același timp a beneficiat de cea mai importantă îmbunătățire după folosirea noului set de date. Evaluarea calitativă arată că antrenarea pe setul de mapare împreună cu cel video ajută la îmbunătățirea robusteții sistemului. Dificultatea estimării mișcării camerei rămâne un factor limitativ, în special în acest context aerian unde pot apărea mișcări pe o plajă mult mai largă de direcții și magnitudini (comparativ cu scenariul tradițional rutier), necesitând studiu continuat.

A2. Recunoașterea și localizarea vizuală (UTCN)

2.1. Modele semantice semi-supervizate cu componente multiple și răspunsuri parțiale (A2.1)

2.1.1 Crearea unui set de date aeriene sintetice, pentru monitorizarea pădurilor(A2.1.1)

Deoarece colectarea și adnotarea datelor necesită mult timp, ne-am îndreptat atenția către simulatoare, care pot fi utilizate pentru a înregistra seturi care îndeplinesc criterii specifice și oferă acces la API-uri care permit controlul dronei în mediul virtual. Astfel am creat setul de date FI (Forest Inspection), compus atât din înregistrări sintetice, expuse în această secțiune, cât și din date reale. Folosind AirSim, am înregistrat 18 secvențe într-un mediu forestier virtual creat în Unreal Engine 4, care conțin peste 26.000 de imagini color, împreună cu segmentări semantice și informații de profunzime. Sunt furnizate date suplimentare de poziționare și

orientare ale dronei. Setul a fost înregistrat în 2 condiții meteorologice: însorit și înnorat, de la altitudini de: 30, 50 și 80 de metri, cu 3 grade de înclinare.

2.1.2 Învățarea segmentării semantice utilizând o metodă semi-supervizată pe secvențe video aeriene (A2.1.2)

În lucrarea "**Weakly Supervised Semantic Segmentation Learning on UAV Video Sequences**", prezentată la EUSIPCO 2021, am propus o metodologie îmbunătățită bazată pe o soluție de învățare profundă recurentă care preia segmentarea semantică din cadrul anterior, utilizează fluxul optic pentru a o deforma la momentul actual și o alimentează, împreună cu imaginile color anterioare și actuale, către o componentă spațio-temporală GRU care corectează și rafinează predicția segmentării semantice. Am abordat acest lucru dintr-o perspectivă de învățare semi-supervizată și am realizat mai multe experimente care ne permit să reducem semnificativ cantitatea optimă de date necesare pentru procesul de învățare. În plus, am îmbunătățit componenta GRU prin introducerea convoluțiilor separabile în funcție de adâncime, care vizează îmbunătățirea acurateței și a vitezei necesare pentru etichetarea precisă a secvențelor video.

Tabel 2. Rezultatele IoU pe setul de test din Mid-Air, unde GRU(k) reprezintă sistemul complet, antrenat cu fiecare a k-a etichetă, folosind fluxul optic din PWC-Net și segmentarea semantică din ERFNet.

Class	ERFNet(4)	GRU(4)	ERFNet(8)	GRU(8)	ERFNet(16)	GRU(16)	ERFNet(32)	GRU(32)	ERFNet(64)	GRU(64)
Sky	93.58	95.39	92.34	95.75	90.43	94.91	89.55	93.47	87.63	91.56
Trees	86.65	89.88	85.23	89.63	83.56	84.06	83.16	83.39	78.96	82.84
Dirt ground	75.48	81.13	70.73	75.48	68.20	72.23	67.52	73.70	65.83	67.16
Ground vegetation	83.16	85.15	79.61	80.39	77.91	84.38	76.22	82.10	73.54	74.40
Rocky ground	68.35	70.74	67.17	68.91	66.72	68.22	66.23	71.71	60.47	62.87
Boulders	67.56	71.66	65.74	68.84	65.30	69.52	65.19	69.98	58.26	60.82
Water plane	86.36	95.28	86.27	88.25	83.33	86.43	81.81	83.58	81.27	85.17
Road	89.38	93.03	86.22	87.72	83.54	87.52	81.33	82.83	79.61	81.12
Train track	81.48	87.90	73.92	74.46	71.80	75.78	70.59	74.54	68.53	69.82
Road sign	48.14	53.58	36.54	41.09	35.18	38.03	34.37	36.52	34.02	35.74
Man-made objects	57.71	65.35	55.41	58.98	54.73	63.60	53.70	56.87	50.84	51.66
Mean IoU	76.17	80.83	72.65	75.41	70.97	74.97	69.97	73.52	67.18	69.38

În Tabel 2 prezentăm rezultatele obținute, care ne arată că performanța rețelei scade odată cu rata de eșantionare, așa cum era de așteptat. Cu toate acestea, observăm că utilizarea componentei STGRU duce la îmbunătățirea rezultatelor segmentării, indiferent de procentul din setul de date pe care îl folosim. Predicțiile rețelei se îmbunătățesc pentru fiecare clasă, deoarece sistemul exploatează atât informațiile spațiale, cât și cele temporale necesare atunci când lipsesc cantități de date adnotate din setul de antrenare.

2.2. Structuri și metode de învățare activă și adversarială pentru date dinamice (A2.2)

2.2.1 Camere semantice pentru percepție la 360 de grade (A2.2.1)

Am dezvoltat un sistem de percepție la 360 de grade care a fost integrat într-un vehicul prototip, prezentat în articolul „**Semantic Cameras for 360-degree Environment Perception in Automated Urban Parking and Driving**", submis spre publicare [10]. Am implementat camere virtuale semantice bazate pe învățarea profundă, care oferă segmentare semantică, de instanță și panoptică prin procesarea imaginilor de la cinci camere: patru camere fisheye și o cameră cu câmp vizual îngust.

2.2.2 Segmentare panoptica a secvențelor video folosind module Transformer Timp-Spațiu (A2.2.2)

Segmentarea panoptică a secvențelor video realizează simultan segmentarea semantică la nivel de pixel, segmentarea la nivel de instanță și urmarirea instanțelor. În articolul "**Time-Space Transformers for Video Panoptic Segmentation**", acceptat la conferința WACV 2022, propunem rețeaua VPS-Transformer, cu o arhitectura hibridă pornind de la rețeaua Panoptic-DeepLab care realizează segmentarea panoptică a imaginilor, pe care o extindem cu un nou modul video care folosește arhitectura Transformer, care utilizează mecanismul de atenție pentru a modela legăturile spatio-temporale între imaginile dintr-o secvență video. Propunem metode de optimizare a arhitecturii Transformer, pentru a obține procesare eficientă a imaginilor de înaltă rezoluție.

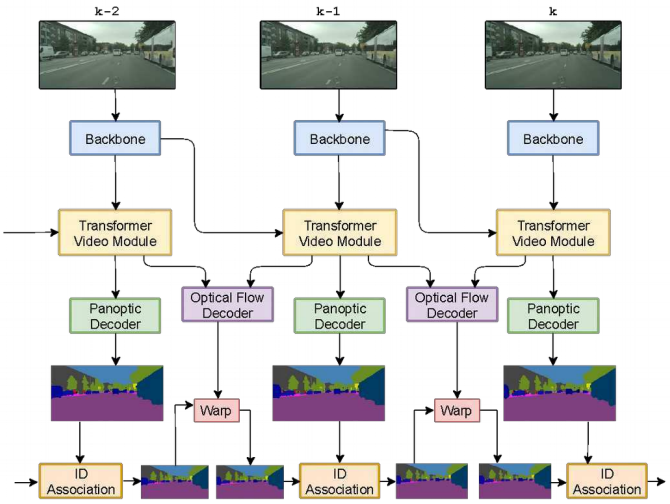

Fig. 3. Arhitectura soluției propuse pentru segmentarea panoptică a secvențelor video.

Tabel 3. Comparație cu stadiul actual al cercetării pentru segmentarea panoptică a secvențelor video. Obținem un VPQ similar cu Vip-DeepLab, dar un timp de inferență de 2x mai mic.

Method	Backbone	PQ	VPQ	Time (ms)
VPSNet	ResNet50	62.7	56.1	770
ViP-DeepLab + MV	WR-41	67.9	59.9	> 400
Baseline - Panoptic DeepLab	ResNet50	63.0	52.0	86
VPS-Transformer (ours)	ResNet50	64.8	57.3	112
Baseline - Panoptic DeepLab	HRNet-W48	66.1	55.1	168
VPS-Transformer (ours)	HRNet-W48	67.6	59.8	185

A3. Control semantic optimal (IMAR)

Task 3.1. Control optimal direct și invers

Intr-un articol publicat în Annual Conference on Robot Learning (CoRL2021, Proceedings of Machine Learning Research), Maria Priisalu, Aleksis Pirinen, Ciprian Paduraru si Cristian Sminchisescu discuta limitarile inerente impuse de dataseturile colectate manual pentru dezvoltarea metodologiilor pentru mașini autonome (self-driving cars) în termeni de variabilitate a cazurilor de test și în particular dificultatea de a genera scenarii provocatoare metodologic și tehnic, de tipul coliziunilor care implica pietoni. O modalitate de a atenua consecințele acestor aspecte este generarea automata de scenarii critice dpdv al sigurantei

pentru testarea vehiculelor autonome. Abordarile existente pentru generarea unor asemenea scenarii folosesc modele euristice pentru comportamentul pietonilor. În articol se propune alternativ un framework care poate folosi modele state-of-the-art pentru mișcarea pietonilor, acest lucru fiind realizat prin reformularea problemei ca o problemă de invatare a locului în care se pot plasa pietonii astfel încât scenariile rezultate să fie predispuse la coliziune pentru un anumit vehicul autonom. Modelul de amplasare inițială a pietonilor poate fi utilizat împreună cu orice model de pieton dirijat de obiective, ceea ce face posibilă testarea unui vehicul autonom cu o gamă largă de comportamente ale pietonilor acest lucru asigură că vehiculul autonom poate evita coliziunile cu orice pieton pe care îl întâlnește. În articol se arata că este posibil să se învețe un model de generare a unui scenariu de căutare a coliziunilor atunci când atât pietonul, cât și vehiculul autonom evită coliziunea. Modelul de amplasament inițial este condiționat de semantica și ocluziile scenei pentru a asigura plauzibilitatea semantică și vizuală, ceea ce crește realismul scenariilor generate. Modelul poate fi folosit pentru a testa orice tip de vehicul autonom dacă se iau în considerare suficiente constrângeri. Exemple de traiectorii antrenate simultan, folosing o inițializare suspecta de a conduce la coliziune, ca și una care nu conduce la coliziune pentru ca vehicului autonom accelereaza, pot fi analizate calitativ în figura următoare:

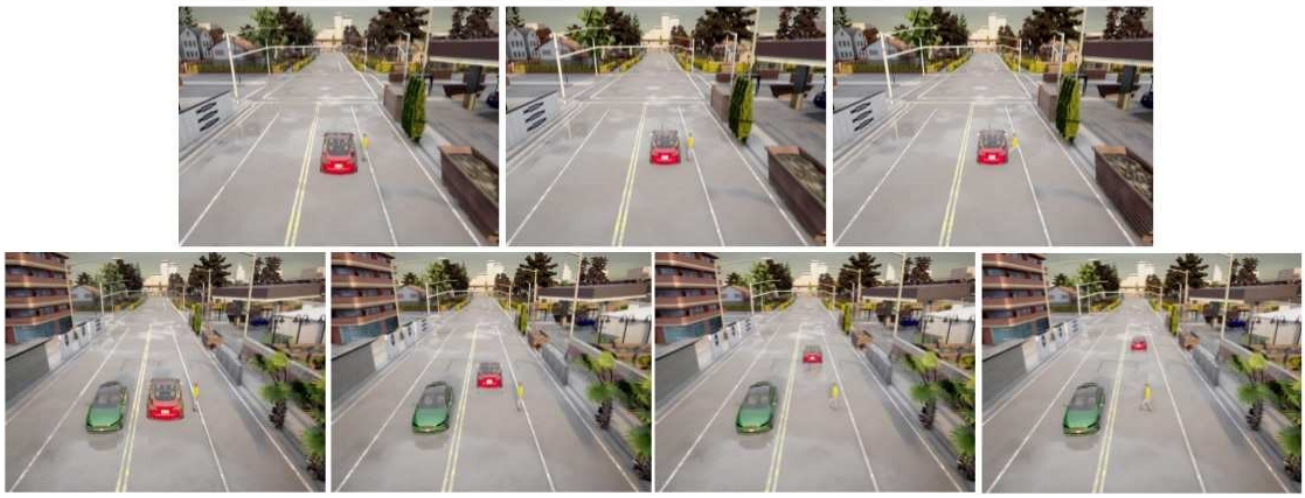

Fig. Exemple de traiectorii simultane. Primul rând: Vehiculul autonom schimba viteza determinand pietonul sa estimeze incorect mișcarea vehiculului autonom și sa intre în acesta. Al doilea rând: Pietonul așteaptă sa treacă vehiculul autonom înainte sa traverseze strada.

Referinta

Maria Priisalu, Aleksis Pirinen, Ciprian Paduraru, Cristian Sminchisescu. Generating Scenarios with Diverse Pedestrian Behaviors for Autonomous Vehicle Testing, In Proceedings of Machine Learning Research (Proceedings of CoRL 2021)

Task 3.2. Reprezentari și metode pentru calcul eficient

Într-un articol publicat la AAAI-2021 se studiaza task-ul de învățare activă vizuală încorporată, în care un agent este setat să exploreze un mediu 3D cu scopul de a dobândi înțelegere vizuală a scenei prin selectarea activă a view-urilor („vederi”) pentru care să solicite adnotări. Deși sunt precise în ceea ce privește anumite repere, pipeline-urile de recunoaștere vizuală profundă de astăzi tind să nu generalizeze bine în anumite scenarii din lumea reală sau pentru viewpoint-uri neobișnuite. Percepția robotică, la rândul său, necesită capacitatea de a rafina capabilitățile de recunoaștere pentru condițiile în care funcționează sistemul mobil, inclusiv medii interioare aglomerate sau iluminare slabă. Acest lucru motivează task-ul propus, în care un agent este plasat într-un mediu nou cu obiectivul de a-și îmbunătăți capacitatea de recunoaștere vizuală. Pentru a studia învățarea activă vizuală încorporată, s-a dezvoltat o colectie de agenți - atât învățați, cât și prespecificați - și cu diferite niveluri de cunoaștere a mediului. Agenții sunt echipați cu o rețea de segmentare semantică și caută să dobândească vederi informative, să se deplaseze și să exploreze pentru a propaga adnotări în vecinătatea acelor vederi, apoi să perfecționeze rețeaua de segmentare subiacentă prin reinstruire online. Metoda antrenabilă folosește deep reinforcement learning cu o funcție de recompensă care echilibrează două obiective concurente: performanța sarcinii, reprezentată ca acuratețea recunoașterii vizuală, care necesită explorarea mediului și cantitatea necesară de date adnotate solicitate în timpul explorării active. Sunt evaluate pe larg modelele propuse utilizând simulatorul fotorealist Matterport3D și se arată că o metodă pe deplin învățată depășește metodele comparabile prespecificate similare, chiar și atunci când se solicită mai puține adnotări. Rezultate calitative sunt prezentate în figurile următoare:

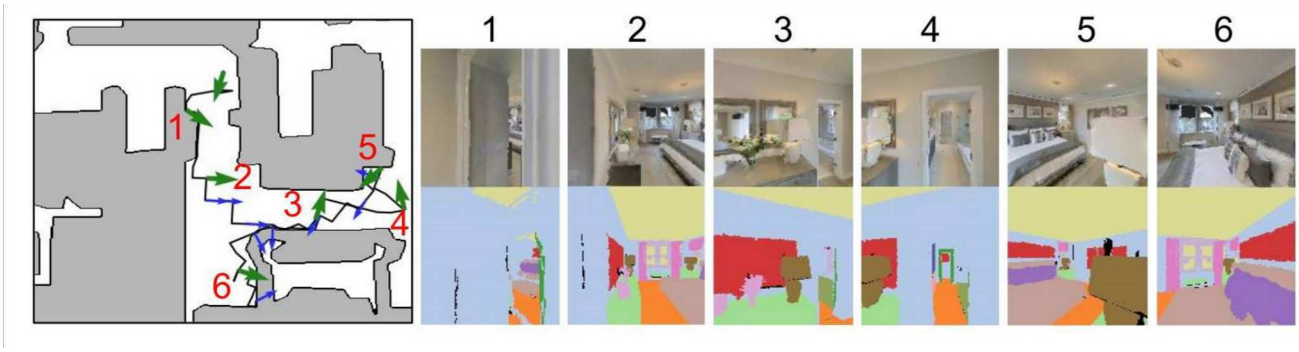

Fig. Primele șase adnotări solicitate de către agentul RL într-o cameră din setul de test. Stânga: Hartă care arată traiectoria agentului și primele șase adnotări solicitate (săgeți verzi). Adnotarea dată inițial nu este indicată cu un număr. Săgețile albastre indică acțiuni de colectare. Dreapta: Pentru fiecare adnotare (numerotată 1 - 6) figurile arată imaginea văzută de agent și ground-truth-ul primit atunci când agentul a solicitat adnotări. După cum se vede, agentul explorează rapid camera și solicită adnotări care conțin diverse clase semantice.

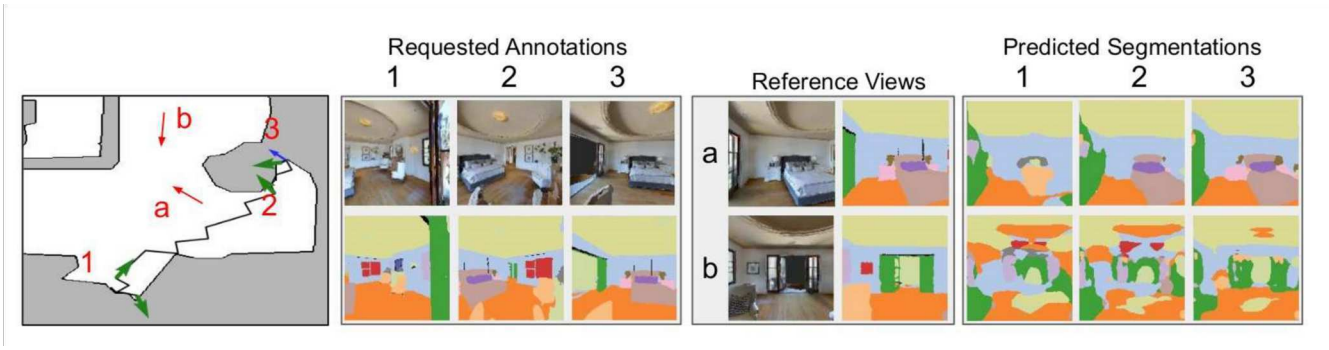

Fig. Exemplu de selecție a viewpoint-ului agentului RL și modul în care percepția acestuia se îmbunătățește în timp. Sunt arătate rezultatele a două vederi (views) de referință după primele trei adnotări ale agentului RL. Stânga: calea de mișcare a agentului este desenată cu negru pe hartă. Adnotările (săgețile verzi) sunt numerotate de la 1 la 3, iar vederile (views) asociate sunt afișate imediat în dreapta hărții (adnotarea dată inițial nu este afișată). Săgețile roșii etichetate a - b indică vederile (views) de referință. Dreapta: vederi (views) de referință și măști groundtruth, urmate de segmentarea prezisă după una, două și trei adnotări. Se observă îmbunătățiri clare de segmentare pe măsură ce agentul solicită mai multe adnotări. Mai exact, se observă cum vederea (view) de referință se îmbunătățește drastic cu adnotarea 2, deoarece patul este vizibil în acea vedere și cu adnotarea 3, unde este văzut sertarul. Este de reținut, de asemenea, cum se îmbunătățește segmentarea pentru vizualizarea de referință b după ce ușa este văzută în adnotarea 3.

Referinta

D. Nilsson , A. Pirinen, E. Gärtner , C. Sminchisescu. “Embodied Visual Active Learning for Semantic Segmentation”, Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)

A4. Optimizarea și integrarea sistemelor (UTCN)

4.1: Localizare simultană și cartografiere semantică(A4.1)

4.1.1 Extinderea setului de date geometrice WildUAV(A4.1.1)

Setul de date denumit WildUAV conține 1500 de imagini de rezoluție înaltă capturate în modul de mapare, pentru care sunt disponibile informații de poziționare și imaginea de profunzime asociată, cărora li se adaugă 11 secvențe video ce totalizează peste 34000 de cadre. Postprocesarea datelor a constat în generarea mesh-urilor 3D pe baza secvențelor de mapare folosind OpenDroneMap, care au fost mai apoi utilizate pentru obținerea imaginilor de profunzime prin proiecția de tip ray-tracing a suprafețelor mesh-ului în imagine, folosind parametrii intrinseci și extrinseci asociați imaginilor din setul de mapare. În același timp, din secvențele video au fost eliminate manual intervalele fără mișcare a camerei, ce perturbă procesul de învățare nesupervizat. Setul de date a fost prezentat în lucrarea “**WildUAV:**

Monocular UAV Dataset for Depth Estimation Tasks” [6] și urmează să fie făcut public după finalizarea procesului de anonimizare a datelor.

4.1.2 Crearea unui set de date aeriene adnotate semantic folosind date reale, pentru monitorizarea pădurilor(A4.1.2)

Datele reale din setul FI (Forest Inspection) introdus în secțiunea dedicată sarcinii 2.1. sunt preluate din setul WildUAV [6]. Imaginile din 4 seturi de mapare au fost adnotate manual, ceea ce a dus la segmentări semantice dense și precise. Am obținut peste 3.200 de imagini reale de segmentare semantică care cuprind 4 secvențe de cartografiere și un video.

4.1.3 Dezvoltarea capabilităților de augmentare a setului de date WildUAV(A4.1.3)

Generarea de noi cadre prin survolarea virtuală a hărților 3D obținute în procesul de mapare

Întrucât achiziția și procesarea datelor reale sunt procese complexe ce necesită resurse considerabile de timp, iar condițiile din teren pot influența semnificativ calitatea și similaritatea datelor capturate deasupra aceleași arii, a fost dezvoltat un proces alternativ de generare a noi imagini pentru antrenarea algoritmilor de percepție. În acest scop, noi traiectorii de zbor, la diferite altitudini și orientări ale camerei, sunt simulate deasupra hărții 3D obținute în procesul anterior de mapare, rezultând noi imagini color și de profunzime, împreună cu date de poziționare. Folosind parametrii intrinseci și extrinseci cunoscuți ai camerei, putem modela o camera virtuală identică cu cea reală, și, pe baza mesh-ului disponibil, să generăm imaginile sintetice. Imaginea de adâncime este generată aplicând direct algoritmul Depth buffer. Pentru a genera imaginea color, ca pas intermediar, se reține inițial doar punctele corespunzătoare vertex-urilor din mesh, care au asociată informația de culoare. Al doilea pas este de interpolare a culorii la nivelul tuturor triunghiurilor din mesh, vizibile conform algoritmului Depth buffer. Soluția pentru a genera imagini color cu grad mare de realism se bazează pe capacitățile crescute de texturare disponibile în biblioteca OpenGL.

Crearea hărții 3D îmbogățite semantic, reconstrucția și adnotarea semantică automată de imagini virtuale

Pe baza unui set redus de imagini segmentate semantic (adnotare manuală) și utilizând mecanismul din modulul prezentat anterior, se pot genera imagini din poziții noi ale camerei virtuale pentru care este disponibilă și imaginea de segmentare semantică, obținută în mod automat. Primul pas, care se aplică o singură dată asupra unui mesh, este cel de propagare a informației semantice din setul restrâns de imagini etichetate înspre vertex-uri. Pentru fiecare imagine etichetată, vertex-urile sunt proiectate folosind parametrii originali ai camerei, iar pentru fiecare vertex, se stabilește care este clasa semantică corectă analizând setul de clase memorate. Al doilea pas constă în generarea de noi imagini etichetate semantic, prin extinderea mecanismului de generare a imaginilor din poziții simulate. Astfel, în paralel cu

generarea imaginii de adâncime, se va genera și imaginea etichetată semantic, pe baza claselor disponibile în vertex-urile din mesh.

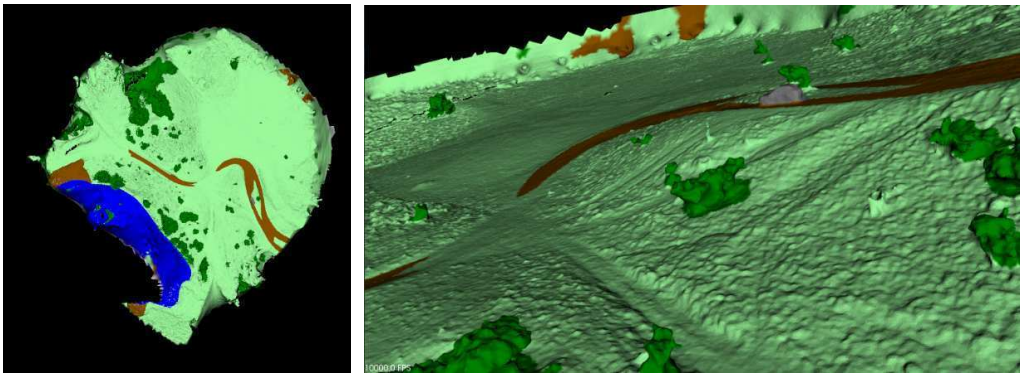

Fig. 4. Mesh-ul îmbogățit cu informație semantică (vedere de sus și perspectivă).

4.1.4 Adaptarea altor seturi de date pentru soluții de segmentare semantică, detecție și estimare a profunzimii(A4.1.4)

Pentru crearea unui set adițional de date de antrenare care să permită dezvoltarea atât a algoritmilor de estimare a profunzimii cât și cei de segmentare semantică, s-a dezvoltat o extensie a setului de date UAVid. Întrucât nu este disponibilă informația de profunzime, a fost folosită aplicația COLMAP ce oferă capabilități de procesare folosind algoritmi de Structure-from-Motion (SfM) și Multi-View-Stereo (MVS) pentru generarea hărților de profunzime. Deoarece imaginile înregistrate în China conțin un număr foarte mare de obiecte în mișcare ce ridică mari dificultăți pentru astfel de abordări, au fost procesate doar secvențele din Germania.

Rezultatele obținute folosind COLMAP pentru setul UAVid au motivat efectuarea unor experimente adiționale folosind o cameră stabilizată pe două axe (DJI Osmo Pocket) transportată manual în zone de pădure deasă la o înălțime de cca. 3m, simulând astfel mișcarea unei drone (eliminând riscul ridicat de coliziune a dronei asociat acestui context). Metoda de reconstrucție a fost validată și în acest scenariu, făcând posibile viitoare extensii ale seturilor de date folosite la dezvoltarea metodelor bazate pe învățare profundă care să acopere și astfel de scene. Procesarea atât a datelor UAVid cât și a celor capturate în vegetația densă a evidențiat dificultățile sistemelor de reconstrucție bazate pe algoritmi SfM, precum: sensibilitatea la inițializare, lipsa de robustețe în cazul rotirii camerei la viteză mărită și erorile generate de obiecte dinamice sau de zonele cu textură redusă, limitări ce motivează încă o dată nevoia pentru soluții monoculare.

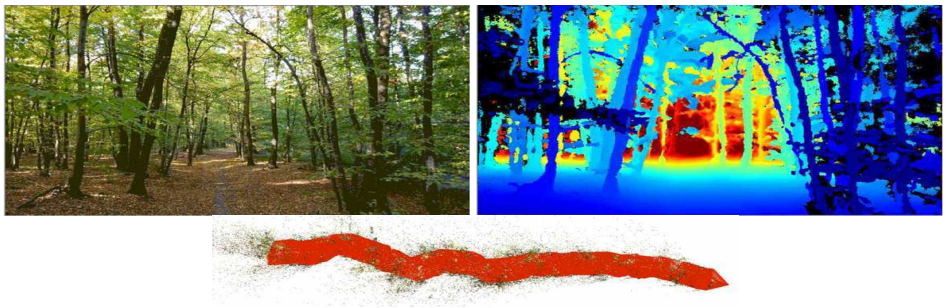

Fig. 5. Rezultate experiment în zonă de pădure: Imaginea originală, cea de profunzime și reconstrucția traseului camerei.

4.1.5 Estimarea ariilor defrișate (A4.1.5)

Pornind de la mediul virtual din simulator, am dezvoltat o metodologie pentru a descrie gradul de împădurire al zonei de interes. Primul pas constă în cartografierea pădurii, care se realizează prin înregistrarea imaginilor color și semantice dintr-o dronă care survolează zona. Al doilea pas este reprezentat de construirea unui nor de puncte 3D cu informație color și de segmentare semantică, pentru care am utilizat librăria Open3D. Prin fuziunea temporală a înregistrărilor se obține o hartă pe baza căreia putem extrage procentajul de pomi sănătoși, dar avem și informații despre gradul pomilor defrișați și al ariei despădurite. Ultimul pas constă în aplicarea unei operații morfologice de dilatare prin care am obținut o hartă mai densă de informație semantică, pe baza căreia am extras contururile zonelor despădurite, rezultate care se observă în Fig. 6.

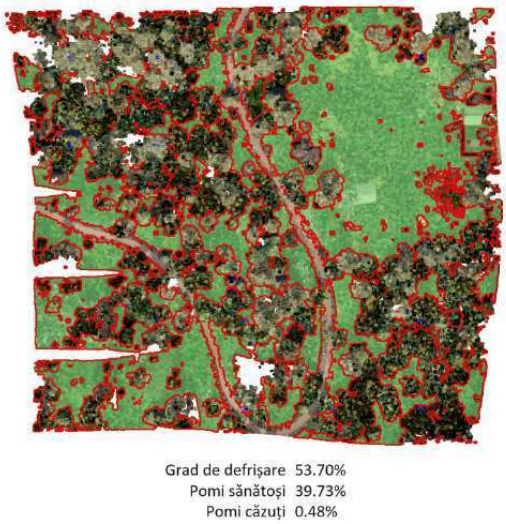

Fig. 6. Extragerea regiunilor fără pomi și informații referitoare la gradul defrișărilor raportat la zona analizată.

4.2: Integrarea navigației vizuale și a înțelegerii scenei (A4.2)

4.2.1 Percepție avansată a mediului bazat pe fuziunea informațiilor geometrice și semantice (A4.2.1)

În articolul "**Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion**" [10], submis jurnalului IEEE Transactions on Intelligent Transportation Systems, prezentăm un sistem avansat de percepție cu acoperire 360° ce a fost proiectat și dezvoltat cu succes pentru a fi operat în timp real la bordul unui autovehicul autonom. Soluția de percepție se bazează pe fuziunea la nivel jos între norii de puncte 3D măsurați de mai mulți senzori LiDAR și informația semantică a scenei obținută prin procesarea imaginilor furnizate de un sistem de camere multiple. Segmentările de tip semantic, instanță și panoptic sunt generate pe baza unor algoritmi de învățare profundă, eficienți și optimizați

pentru resursele hardware folosite. În același timp, norii 3D sunt segmentați folosind un algoritm tradițional bazat pe voxeli, ce asigură un timp de rulare scăzut. Pe baza datelor geometric-semantice obținute în urma fuziunii au fost dezvoltați algoritmi eficienți ce asigură detecția și clasificarea obiectelor, precum și localizarea vehiculului.

4.2.2 Navigare autonomă a unei drone cu evitarea obstacolelor dintr-o pădure (A4.2.2)

În această secțiune ne propunem să rezolvăm problema navigării autonome a dronei, cu evitarea obstacolelor din scenă. Primind o traiectorie reprezentată de puncte 3D din mediu, drona trebuie să navigheze prin modificarea unghiului de orientare, fără a se ciocni de obiectele prezente în scenă, în timp ce menține o distanță constantă față de sol. Navigarea pe bază de profunzime utilizează fâșia din mijlocul zonei de interes, care este împărțită în 11 cadrane egale. Dacă în cadranul din mijloc, poziționat pe direcția de mers în față, există obiecte mai apropiate decât un prag stabilit de noi, atunci vom lua decizia să ne îndreptăm înspre cadranul în care am calculat cea mai mare distanță din cele 10 posibile. În cazul în care nu există obstacole, unghiul de deplasare al dronei este diferența între vectorul ei de orientare înainte și vectorul de la poziția dronei până la punctul obiectiv. Altitudinea față de orientarea în jos este ajustată astfel încât să se mențină o distanță egală față de sol sau clădire. În cadrul acestei aplicații vom introduce algoritmii dezvoltați de noi pentru estimarea profunzimii, cum ar fi MDE sau SSDE, iar pentru segmentare semantică vom utiliza segmentarea panoptică sau cea semi-supervizată din videoclipuri.

Publicații UTCN

Lucrări publicate:

[1] V.C. Miclea, S. Nedevschi, “Monocular Depth Estimation With Improved Long-Range Accuracy for UAV Environment Perception”, *IEEE Transactions on Geoscience and Remote Sensing,* Early Access, MAR 2021, DOI: 10.1109/TGRS.2021.3060513. (Q1, FI 5.6)
[2] R. Brehar, M.P. Muresan, M. Tiberiu, C. Vancea, N. Mihai, S. Nedevschi, “Pedestrian Street-Cross Action Recognition in Monocular Far Infrared Sequences”, IEEE ACCESS, Vol. 9, pp. 74302-74324, JUN 2021, DOI:10.1109/ACCESS.2021.3080822. (Q2, FI 3.367)
[3] M.P. Muresan, S. Nedevschi, R. Danescu, “Robust Data Association using Fusion of Data-Driven and Engineered Features for Real Time Pedestrian Tracking in Thermal Images”, *SENSORS,* Vol. 21 Issue 23, AN 8005, NOV 2021, DOI: 10.3390/s21238005. (Q1, FI 3.567)

[4] A. Petrovai, S. Nedevschi, “Time-Space Transformers for Video Panoptic Segmentation”, accepted at 2022 IEEE Winter Conference on Applications of Computer Vision (WACV 2022), 4-8 January 2022, Waikoloa, Hawai, USA.
[5] B.C.Z. Blaga, S. Nedevschi, “Weakly Supervised Semantic Segmentation Learning on UAV Video Sequences”, EUSIPCO 2021, Dublin, Ireland.
[6] H. Florea, V.C. Miclea, S. Nedevschi, “WildUAV: Monocular UAV Dataset for Depth Estimation Tasks”, in Proceedings of 17th 2021 IEEE International Conference Intelligent Computer Communication and Processing (ICCP 2021).
[7] S. Deac, C. Vancea, S. Nedevschi, “MVGNet: 3D object detection using multi-volume grid representation in urban traffic scenarios”, in Proceedings of 17th 2021 IEEE International Conference Intelligent Computer Communication and Processing (ICCP 2021), 27-20 October, 2021.

[8] B. Maxim, S. Nedevschi, "OccTransformers: Learning occupancy using attention", in Proceedings of 17th 2021 IEEE International Conference Intelligent Computer Communication and Processing (ICCP 2021).
[9] B. Maxim, S. Nedevschi, "A survey on the current state of the art on deep learning 3D reconstruction", in Proceedings of 17th 2021 IEEE International Conference Intelligent Computer Communication and Processing (ICCP 2021).
[10] M.P. Muresan, R. Marchis, S. Nedevschi, R. Danescu "Stereo and Mono Depth Estimation Fusion for an Improved and Fault Tolerant 3D Reconstruction", in Proceedings of 17th 2021 IEEE International Conference Intelligent Computer Communication and Processing (ICCP 2021).

Lucrări submise pentru publicare:
[11] Petrovai, S.Nedevschi, "Pseudo-annotation based unsupervised monocular depth estimation", submis la 2022 Conference on Computer Vision and Pattern Recognition (CVPR).
[12] Petrovai, S. Nedevschi, "Semantic Cameras for 360-degree Environment Perception in Automated Urban Parking and Driving", submis la IEEE Transactions on Intelligent Transportation Systems.
[13] H. Florea, A. Petrovai, I. Giosan, F. Oniga, R. Varga, S. Nedevschi, "Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion", submis la IEEE Transactions on Intelligent Transportation Systems.

Lucrări prezentate la workshopul „Semantic and Geometric Visual Perception" organizat la 2021 IEEE ICCP de Universitatea Tehnica din Cluj-Napoca in cadrul proiectului PN-III-P4-ID-PCCF-2016-0180 Integrated Semantic Visual Perception and Control for Autonomous Systems - SEPCA.

[14] Bianca Cerasela Zelia Blaga: "Large aerial dataset creation using real and synthetic data for forest monitoring", Semantic and Geometric Visual Perception" workshop organizat la 2021 IEEE ICCP.
[15] Vlad Miclea: "Supervised Monocular Depth Estimation for Aerial Images", Semantic and Geometric Visual Perception" workshop organizat la 2021 IEEE ICCP.
[16] Horatiu Florea: "Overview of Self-Supervised Monocular Depth Estimation", Semantic and Geometric Visual Perception" workshop organizat la 2021 IEEE ICCP.

[17] Selma Evelyn Catalina Deac: "Overview of 3D Object Detection from Point Clouds", Semantic and Geometric Visual Perception" workshop organizat la 2021 IEEE ICCP.

Concluzii: Consideram obiectivele etapei ca fiind indeplinite, activitatile proiectului desfasurandu-se conform calendarului stabilit si intr-un ritm semnificativ, marcat de publicatiile UTCN (vezi lista de mai sus) precum si cele ale IMAR : 2 articole publicate la AAAI-21, 1 articol publicat la CVPR 2021, 1 articol publicat la 5th Annual Conference on Robot Learning (CoRL2021, Proceedings of Machine Learning Research), 1 articol publicat în ACM Transactions on Computer-Human Interaction (august 2021), și 1 articol publicat la NeurIPS (NIPS) 2021, care se adauga publicatiilor IMAR din anii trecuti (4 publicatii la IEEE CVPR 2018, 1 publicatie la NIPS 2018, dintre care un Best Paper Award Honorable Mention -- Deep Learning for Graph Matching, A. Zanfir si C. Sminchisescu --, 1 publicație la AAAI-20, 1 articol publicat la IEEE CVPR 2020, 1 articol publicat la ACCV 2020).