Izboljšava panoptične detekcije ovir na vodni domeni

Lojze Žust, Matej Kristan

Fakulteta za računalništvo in informatiko, Unverza v Ljubljani
E-pošta: lojze.zust@fri.uni-lj.si

Abstract

Robust maritime obstacle detection is crucial for autonomous surface vehicles (ASVs). Recently, the domain is undergoing a shift toward panoptic segmentation [16, 21] as a successor to the widely used semantic segmentation approaches, enabling the separation of obstacle instances. However, current panoptic segmentation methods still lack the desired accuracy required for practical applications. This work aims to address several common issues by proposing two extensions for transformer-decoder-based methods, namely to improve the separation of object instances and disentangle the processing of background and foreground classes. We combine these contributions into a new panoptic method AnchorFormer, which achieves state-of-the-art results on the largest maritime panoptic benchmark [21].[1]

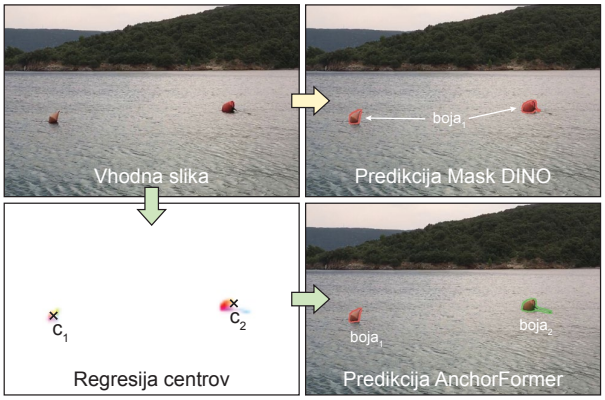
1 Uvod

Računalniški vid za avtonomna plovila je hitro razvijajoče se področje [8] s številnimi aplikacijami, od manjših (npr. avtomatiziran nadzor marin) do daljnosežnih (npr. avtomatiziran transport ljudi in tovora). Za varno in uspešno navigacijo avtonomnih plovil, je ključnega pomena pravočasno in natančno zaznavanje ovir na vodni gladini.

Pomemben del raziskav se usmerja v zaznavanje na podlagi vizualne informacije iz kamer, ki so priljubljene zaradi cenovne ugodnosti, gostote informacij in številnih uspehov na drugih področjih računalniškega vida (npr. avtonomna vozila). Za namen zaznave ovir se je še posebej obetaven izkazal segmentacijski pristop [11, 1, 20], ki vsebino slike razdeli na več semantičnih kategorij (npr. voda, nebo, ovire). Tak pristop lahko na splošen način naslovi različne tipe in oblike ovir in je praktičen z vidika navigacije, saj direktno segmentira plovno površino.

Kljub temu, tak pristop ni idealen, saj ni zmožen ločevanja med posameznimi instancami ovir in ozadjem, ter posledično onemogoča sledenje in napovedovanje trajektorij dinamičnih ovir (npr. čolnov). Zaradi tega razloga so se v zadnjem času pojavile učne in evalvacijske zbirke [16, 21] s panoptičnimi anotacijami, ki zahtevajo napovedi na nivoju objektov. Rezultati na teh zbirkah kažejo, da trenutne splošne panoptične metode še niso dovolj sposobne za potrebe detekcije ovir.

[1]Delno financirano iz ARRS programa P2-0214 in projekta J2-2506

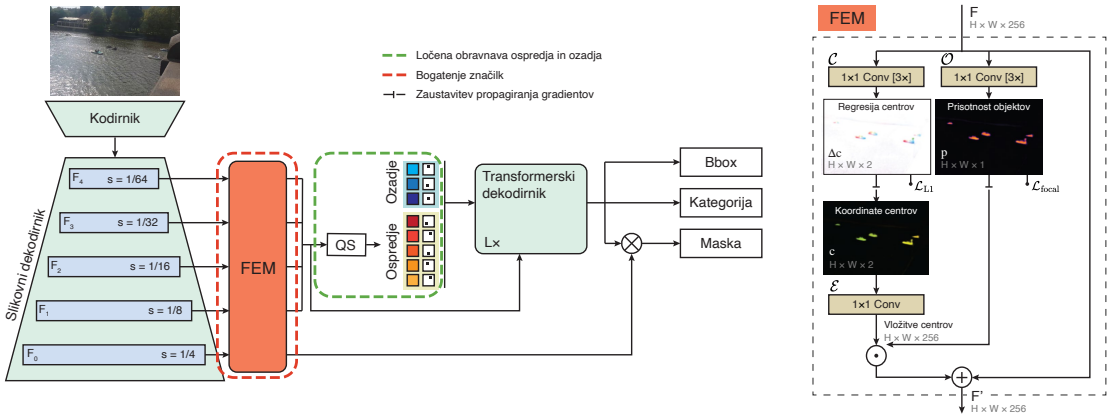


Slika 1: Obstoječe metode imajo težave z združevanjem več objektov v eno masko (zgoraj desno). Predlagana metoda AnchorFormer značilke obogati z informacijo iz vmesne predikcije centrov objektov (spodaj levo), kar mreži olajša ločevanje med objekti (spodaj desno).

Najboljši predstavniki trenutnih panoptičnih metod [6, 13] uporabljajo zelo podobno arhitekturo dekodiranja poizvedb (angl. queries). Vsaka poizvedba predstavlja eno instanco objekta in se na koncu preslika v napoved razreda in maske objekta. V tem delu analiziramo dva pomembna vira napak metod, ki temeljijo na tem pristopu in predlagamo arhitekturne spremembe, ki ju naslovijo.

Prva pomankljivost trenutnih pristopov je, da razrede ozadja in ospredja obravnavajo na enak način, kljub ključnim razlikam med njimi – razredi ospredja so običajno relativno kompaktni objekti, značilne oblike in velikosti (npr. čolni, boje, plavalci), medtem ko so razredi ozadja splošne kategorije brez oblike (npr. voda, nebo, obala). Metode, ki izhajajo iz segmentacijskega sveta [6] uporabljajo statične učljive poizvedbe, ki kodirajo predhodno informacijo o videzu in lokaciji objektov. Tak pristop odpove v primeru majhnih objektov in v scenah z velikim številom objektov.

Nasprotno pa pristopi, ki izhajajo iz področja detekcije [13], dinamično pripravijo predloge poizvedb na podlagi izbora najbolj izrazitih značilk in njihovih lokacij iz značilk slike. Tak pristop bistveno bolje naslovi ločevanje in detekcijo majhnih objektov, vendar temelji na prisotnosti izrazitih ključnih točk objektov, kar v primeru razredov ozadja pogosto ne drži. V tem delu predlagamo



Slika 2: Arhitektura metode. Slikovni dekodirnik na podlagi značilk kodrinika pripravi večnivojske značilke. Slednje nato modul FEM obogati z informacijo pripadnosti objektom na podlagi vmesne napovedi lokacije centrov objektov. Poizvedbe v transformerskem dekodirniku ločimo na dva dela: poizvedbe za razrede ozadja, ki so učljive in statične, medtem ko pridobimo poizvedbe za dinamične ovire s postopkom izbire poizvedb (QS) iz značilk. Vsaka poizvedba se dekodira v napoved kategorije, maske in omejevalnega okvirja.

združitev pozitivnih lastnosti obeh pristopov in razdelimo poizvedbe na dva dela: (i) statične, učljive poizvedbe za naslavljanje amorfnih razredov ozadja in (ii) dinamične predloge poizvedb za zaznavo instanc objektov ospredja.

Druga pomanjkljivost metod je predikcija maske na podlagi podobnosti med poizvedbo in visoko-resolucijskimi značilkami slike. Pristop predpostavlja, da so visoko-resolucijske značilke dovolj diskriminativne za ločevanje med instancami objektov. V praksi pa se izkaže, da metode še vedno pogosto nepravilno združujejo več podobnih objektov v eno samo masko, saj značilke v prvi vrsti še vedno vsebujejo le semantične informacije o videzu piksla. V tem delu predlagamo mehanizem za obogatitev značilk za boljše ločevanje med posameznimi instancami objektov. Specifično, predlagamo modul FEM, ki za vsako lokacijo v značilkah napove center dotičnega objekta (Slika 1). To informacijo, ki vsakemu pikslu implicitno priredi pripadnost objektu, zakodiramo v značilke in posledično olajšamo ločevanje med objekti v preostanku mreže.

V tem delu torej predlagamo dva prispevka, ki se lahko aplicirata na obstoječe transformerske metode: (i) ločena obravnava razredov ospredja (dinamično) in ozadja (statično) izboljša zanesljivost napovedi razredov ozadja, in (ii) modul za bogatenje značilk z informacijo o pripadnosti objektom izboljša ločevanje med instancami objektov ospredja. Razširitve združimo v novo metodo na osnovi Mask DINO [13] in jo poimenujemo *AnchorFormer*. Metodo evalviramo na največji panoptični zbirki v vodni domeni LaRS, kjer dosežemo nove najboljše rezultate.

2 Sorodna dela

V zadnjih letih so se na področju detekcije ovir v vodni domeni zelo uveljavile metode semantične segmentacije [11, 2, 1, 18, 20, 8], saj so sposobne na splošen način nasloviti ovire najrazličnejših oblik in velikosti. Vendar tak tip napovedi ni idealen za praktično uporabo, saj predikcijo izvaja na nivoju pikslov in ne objektov, zato so bili razviti tudi posebni protokoli za evalvacijo praktične uporabnosti napovedanih segmentacij [3].

Zaradi teh pomankljivosti, je v zadnjem času vedno več raziskav usmerjenih v aplikacijo in razvoj metod panoptične segmentacije [16, 15, 21], ki ohranijo prednosti semantične segmentacije, hkrati pa omogočajo predikcijo na nivoju posameznih objektov. Nedavno ustvarjena panoptična zbirka vodnih scen LaRS [21] je omogočila evalvacijo in analizo obstoječih panoptičnih metod, ter razkrila, da le-te še bistveno zaostajajo za metodami semantične segmentacije iz vidika zanesljivosti.

Med njimi so se za posebej perspektivne izkazale metode s transformersko arhitekturo, ki dekodirajo razrede in maske objektov na podlagi poizvedb. Metoda Mask2-Former [6] uporablja učljive poizvedbe za vse razrede, masko pa napove na podlagi podobnosti med dekodirano poizvedbo in visokoresulcijskimi značilkami. V primerjavi s konvolucijskimi metodami je natančnejša v segmentaciji, ima pa več težav z ločevanjem objektov, tudi če so v sliki daleč narazen. Mask DINO [19] sledi podobnemu pristopu, ter ga razširi z nedavnimi dognanji na področju detekcije objektov [12, 14] (npr. generiranje predlogov poizvedb), ki izboljšajo detekcijo in lokalizacijo objektov, nekoliko pa zanemarijo segmentacijo razredov ozadja.

3 Metoda

Metodo AnchorFormer smo zasnovali na podlagi meta-arhitekture transformerksega dekodirnika za segmentacijo [6, 13]. Specifično, izhajamo iz arhitekture Mask DINO, ki jo razširimo z dvema ključnima komponentama: (i) ločenim dekodiranjem razredov ozadja in ospredja (Poglavje 3.2) in (ii) bogatenjem značilk z informacijo o objektih (Poglavje 3.1). Delovanje mreže na visokem nivoju je prikazano na Sliki 2. V nadaljevanju opišemo komponenti, ki predstavljata novost v metodi AnchorFormer, za podrobnosti preostalih delov arhitektur pa bralca naslavljamo na [13].

3.1 Bogatenje značilk z informacijo o objektih

Pri sorodnih arhitekturah za segmentacijo, napoved končne maske dobimo kot podobnost $m_j = q'_j F_0$ med dekodirano poizvedbo q'_j in visoko-resolucijskimi značilkami F_0, kjer je $j = 1 \dots N$ in N število poizvedb. Tak pristop se ključno zanaša na kvaliteto značilk – ne glede na vrednosti dekodiranih poizvedb, instanc objektov ni zmožen ločiti, če so le-ti v prostoru značilk preveč podobni. Ta problem naslovimo z novim modulom za bogatenje značilk (angl. feature enrichment module, FEM).

Cilj našega pristopa je priprava reprezentacije, ki se po vrednosti minimalno razlikuje znotraj pisklov istega objekta in čim bolj razlikuje med piksli različnih objektov. Za dosego tega cilja se zgledujemo po polno-konvolucijskih metodah za detekcijo objektov [17]. Specifično, za vsak piksel f_i v vhodnih značilkah F_k na nivoju k napovemo dve vrednosti: verjetnost prisotnosti objekta $o_i = \mathcal{O}(f_i)$ (t.j. objektnost) in vektor premika $\Delta\mathbf{c_i} = \mathcal{C}(f_i)$ do lokacije centra objekta, kateremu piksel pripada.

Na podlagi premika lahko izračunamo absolutno lokacijo centra trenutnega objekta $\mathbf{c_i} = \mathbf{r_i} + \Delta\mathbf{c_i}$, kjer je $\mathbf{r_i}$ krajevni vektor piksla. Lokacija centra $\mathbf{c_i}$ je dober kandidat za kodiranje informacije o pripadnosti objektu – (i) vsi piksli objekta imajo enako vrednost $\mathbf{c_i}$, in (ii) lokacija centra unikatno opiše posamezen objekt ter posledično omogoča ločevanje med njimi.

Lokacijo centra vložimo v prostor značilk z uporabo preproste mreže $\mathcal{E}$. Na koncu vložene značilke pomnožimo z napovedano verjetnostjo objekta in tako poskrbimo, da se vložitve upoštevajo le na lokacijah objektov in ne na ozadju, kjer le-te niso smiselne. Originalne značilke nato obogatimo tako, da jim prištejemo pridobljene vložitve $f'_i = f_i + o_i\mathcal{E}(\mathbf{c_i})$. Mreže $\mathcal{O}$, $\mathcal{C}$ in $\mathcal{E}$ uporabljajo iste uteži na vseh nivojih F_k, vsak nivo pa ima dodaten učljiv skalirni faktor s katerim se pomnožijo napovedi regresije centrov, da naslovimo razlike v velikosti pikslov na posameznem nivoju.

3.1.1 Učenje

Mreži $\mathcal{C}$ in $\mathcal{O}$ učimo z uporabo *focal loss* za učenje $\mathcal{O}$ in *L1 loss* za $\mathcal{C}$. Pri računanju napake na različnih nivojih F_k pri učenju upoštevamo le objekte, ki po x smeri razpenjajo vsaj 4 piksle pri tej resoluciji značilk. Izjema je najvišji nivo F_0, kjer upoštevamo vse objekte. $\mathcal{E}$ se uči prek standardnih cenilnih funkcij izhodov mreže [13].

3.2 Ločena obravnava ozadja in ospredja

V metodi Mask DINO poizvedbe nastanejo na podlagi postopka izbire poizvedb (angl. query selection, QS), ki iz večnivojskih značilk izbere značilke in lokacije pikslov z največjo napovedano prisotnostjo objekta. Tak pristop je neustrezen za razrede ozadja, ki običajno pokrivajo velik del slike in jih ne moremo učinkovito predstaviti s točkovnimi značilkami. Zato uporabimo dva nabora značilk, N_{th} poizvedb za razrede ospredja, ki uporabljajo QS in dodaten nabor N_{st} *učljivih* poizvedb za razrede ozadja, ki so statične in neodvisne od vhodne slike. Značilke nato vstopijo v skupen transformerski dekodirnik kot običajno.

3.2.1 Učenje

Med učenjem se za povezovanje poizvedb z ustreznimi segmenti iz anotacij uporablja madžarska metoda (angl. Hungarian matching) [4]. Pri metodi AnchorFormer povezovanje prilagodimo tako, da se izvaja ločeno med anotacijami in poizvedbami ozadja, ter med anotacijami in poizvedbami objektov ospredja. Mask DINO med učenjem uporablja tudi posebne pomožne poizvedbe za odstranjevanje šuma, ki pomagajo pri lokalizaciji objektov. Le-te prilagodimo, da se izvedejo le na objektih ospredja, kjer je to smiselno.

Tabela 1: Ablacijska študija AnchorFormer na LaRS. Na osnovno metodo dodajamo ločevanje poizvedb ozadja in ospredja (§ 3.2) in bogatenje značilk z informacijo o centrih objektov (§ 3.1). V oklepajih so prikazane vrednosti na podmnožici težkih primerov.

metoda	PQ	RQ	SQ
Mask DINO	50.1 (43.3)	59.0 (50.8)	74.5 (74.0)
+ ospredje/ozadje	50.5 (44.8)	59.4 (52.2)	74.4 (73.9)
+ FEM	**50.6** (**45.5**)	**59.5** (**52.9**)	**82.2** (**74.2**)

4 Eksperimenti

Za kodirnik mreže uporabimo ResNet-50 [7]. Mreža uporablja $N_{\text{th}} = 250$ dinamičnih poizvedb za razrede ospredja in $N_{\text{st}} = 50$ učljivih poizvedb za razrede ozadja. Metodo evalviramo na zbirki LaRS [21], največji, najbolj raznoliki in najzahtevnejši vodni panoptični zbirki, ki vsebuje več kot 4000 raznolikih vodnih scen. Pri vseh eksperimentih za učenje uporabljamo velikost svežnja 8 na računskem sistemu s 4 × NVIDIA A100 GPU-ji. Ostale podrobnosti implementacije povzamemo po [13].

Pri evalvaciji uporabljamo standardno metriko panoptične kvalitete (PQ) [10], ter kvaliteti prepoznavanja (RQ) in segmentacije (SQ) iz katerih je sestavljena. V nekaterih eksperimentih rezultate poročamo ločeno na razredih ospredja (Th) in ozadja (St). Dodatno s Th_a označujemo vrednosti pri poenostavljenem problemu, kjer vse razrede ospredja združimo v en razred.

V nadaljevanju analiziramo komponente razvite metode AnchorFormer (§ 4.1), nato pa jo primerjamo s trenutno najboljšimi panoptičnimi mrežami na vodni domeni (§ 4.2).

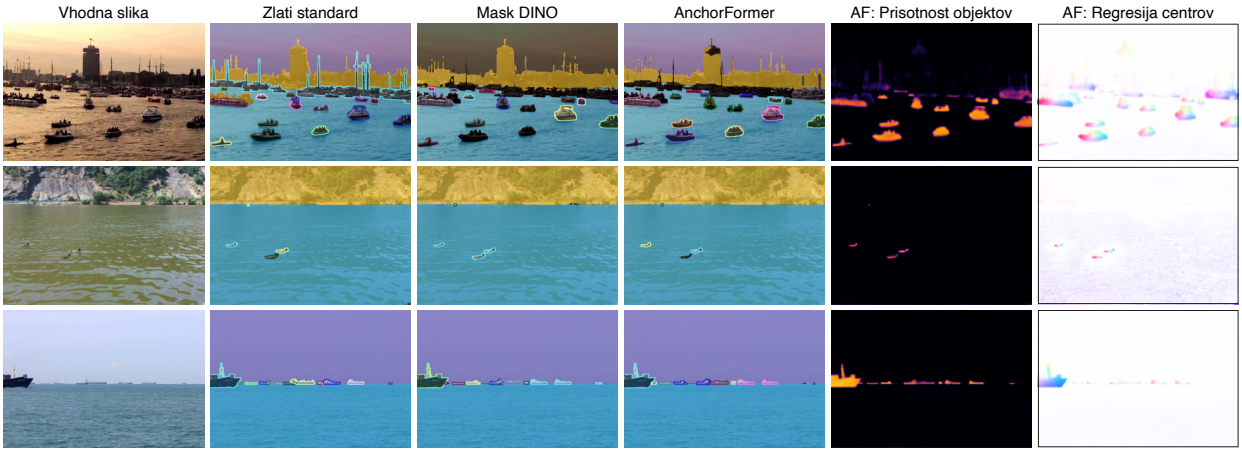
4.1 Ablacijska študija

Ablacijsko študijo izvedemo na testni množici LaRS in na podmnožici težkih primerov, ki smo jih avtomatsko izbrali na podlagi zaznanih napak združevanja objektov trenutno najboljših metod. Rezultati ablacijske študije so podani v Tabeli 1. Za osnovo vzamemo mrežo brez obeh predlaganih komponent, t.j. Mask DINO [13], ki ima $N = 300$ poizvedb.

Najprej vklopimo ločeno obravnavanje ozadja in ospredja v transformerskem dekodirniku (Poglavje 3.2). Leto prinese skok +0.4 % v skupni meri panoptične kvalitete in +1.5% na podmnožici težkih primerov. Ko do-

Tabela 2: Panoptična kvaliteta (PQ), kvaliteta prepoznavanja (RQ) and kvaliteta segmentacije (SQ) na testni množici. Poleg skupne ocene (Vse), prikažemo tudi ocene le na razredih ozadja (St) in dinamičnih ovir (Th), kjer Th_a pomeni, da ne zahtevamo ločevanja med razredi dinamičnih ovir. Z zlato, srebrno in bronasto barvo so označene najboljše tri vrednosti pozamezne metrike.

		PQ (%)				RQ (%)				SQ (%)			
mreža	kodirnik	All	Th	Th_a	St	All	Th	Th_a	St	All	Th	Th_a	St
Panoptic Deeplab [5]	ResNet-50	34.7	13.4	33.0	91.4	40.3	19.3	46.3	96.2	69.5	60.0	71.3	94.9
Panoptic FPN [9]	ResNet-50	40.1	21.7	35.5	89.3	46.9	28.6	45.9	95.8	73.5	66.1	77.3	93.1
Mask2Former [6]	ResNet-50	37.6	17.0	27.9	92.4	43.7	23.6	37.6	97.3	71.3	62.4	74.2	95.0
Mask2Former [6]	Swin-B	41.7	21.8	33.6	**94.7**	48.5	29.7	44.6	**98.5**	78.2	71.5	75.3	96.2
Mask DINO (1 stage) [13]	ResNet-50	41.4	22.3	30.0	92.5	47.1	28.9	37.9	95.8	75.2	67.2	**79.2**	**96.5**
Mask DINO (2 stage) [13]	ResNet-50	50.1	34.2	47.7	92.5	59.0	45.0	61.4	96.1	74.5	66.3	77.8	96.2
AnchorFormer (our)	ResNet-50	**50.6**	**34.5**	**49.1**	93.5	**59.5**	**45.3**	**63.1**	97.2	**82.2**	**77.0**	77.7	96.1



Slika 3: Kvalitativna primerjava delovanja metod Mask DINO [13] in AnchorFormer. Z različnim barvami so prikazane instance objektov. Za metodo AnchorFormer poleg napovedi prikažemo tudi vmesne napovedi prisotnosti objektov in regresije centrov (za najvišji nivo značilk), ki se uporabljajo pri bogatenju značilk. V regresiji centrov barva kodira smer, saturacija pa dolžino vektorja.

damo še bogatenje značilk z modulom FEM, opazimo še dodatno izboljšavo +0.1% in +0.7% na podmnožici težkih primerov. Največjo razliko bogatenje značilk prinese v segmentacijski kvaliteti (SQ), kjer opazimo +7.8% izboljšavo.

4.2 Primerjava z najboljšimi metodami

AnchorFormer primerjamo s trenutno najboljšimi metodami za panoptično segmentacijo. Vključili smo dve klasični konvolucijski metodi [5, 9], ter dve metodi ki sledita isti meta-arhitekturi dekodiranja poizvedb [6, 13] kot AnchorFormer. Rezultati so prikazani v Tabeli 2. AnchorFormer prekaša vse obstoječe metode v vseh treh osnovnih merah. Bistveno boljši (+8,9% PQ) je od Mask2Former, ki uporablja le statične poizvedbe in boljši (+0.5% PQ) od dvo-stopenjskega Mask DINO, ki uporablja le dinamične poizvedbe. AnchorFormer najbolj izstopa po segmentacijski kvaliteti (SQ), kjer je je za 4% boljši od druge najboljše metode Mask2Former.

Vizualna primerjava AnchorFormer z najbolj sorodno mrežo Mask DINO (Slika 3) razkrije, da metoda AnchorFormer zaradi statičnih poizvedb ne zgreši razredov ozadja (1. vrstica), pravilno napove več objektov (1. vrstica) in bolje ločuje med njimi (1. in 2. vrstica). Dodatno se pokaže, da so napovedi lokacije centrov (zadnji stolpec) objektov zelo robustne in pravilno ločujejo med objekti. Včasih je ločevanje med objekti boljše kot v končnih napovedih (2. vrstica: manjkajoča napoved race, 3. vrstica: združene ladje v napovedih), kar nakazuje, da trenutna metoda še ne zna povsem izkoristiti te informacije.

5 Zaključek

V tem delu smo predstavili metodo AnchorFormer, ki naslavlja dve pomanjkljivosti metode Mask DINO: obravnavo razredov ozadja izboljšamo z ločenimi statičnimi poizvedbami, ločevanje med objekti pa izboljšamo z modulom za bogatenje značilk z informacijo o centrih objektov. Metoda dosega nove najboljše rezultate na panoptični vodni zbirki LaRS. V nadaljnjem delu želimo še izboljšati upoštevanje napovedanih lokacij centrov objektov za ločevanje instanc v preostanku mreže.

Literatura

[1] Borja Bovcon and Matej Kristan. WaSR–A Water Segmentation and Refinement Maritime Obstacle Detection

Network. *IEEE Transactions on Cybernetics*, pages 1–14, July 2021.

[2] Borja Bovcon, Rok Mandeljc, Janez Perš, and Matej Kristan. Stereo obstacle detection for unmanned surface vehicles by IMU-assisted semantic segmentation. *Robotics and Autonomous Systems*, 104, 2018.

[3] Borja Bovcon, Jon Muhovič, Duško Vranac, Dean Mozetič, Janez Perš, and Matej Kristan. MODS – A USV-oriented object detection and obstacle segmentation benchmark. *IEEE Transactions on Intelligent Transportation Systems*, May 2021.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12346 LNCS:213–229, May 2020.

[5] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12485, June 2020.

[6] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[8] Benjamin Kiefer, Matej Kristan, Janez Perš, Lojze Žust, Fabio Poiesi, Fabio Andrade, Alexandre Bernardino, Matthew Dawkins, Jenni Raitoharju, Yitong Quan, Adem Atmaca, Timon Höfer, Qiming Zhang, Yufei Xu, Jing Zhang, Dacheng Tao, Lars Sommer, Raphael Spraul, Hangyue Zhao, Hongpu Zhang, Yanyun Zhao, Jan Lukas Augustin, Eui-ik Jeon, Impyeong Lee, Luca Zedda, Andrea Loddo, Cecilia Di Ruberto, Sagar Verma, Siddharth Gupta, Shishir Muralidhara, Niharika Hegde, Daitao Xing, Nikolaos Evangeliou, Anthony Tzes, Vojtěch Bartl, Jakub Špaňhel, Adam Herout, Neelanjan Bhowmik, Toby P. Breckon, Shivanand Kundargi, Tejas Anvekar, Ramesh Ashok Tabib, Uma Mudenagudi, Arpita Vats, Yang Song, Delong Liu, Yonglin Li, Shuman Li, Chenhao Tan, Long Lan, Vladimir Somers, Christophe De Vleeschouwer, Alexandre Alahi, Hsiang-Wei Huang, Cheng-Yen Yang, Jenq-Neng Hwang, Pyong-Kun Kim, Kwangju Kim, Kyoungoh Lee, Shuai Jiang, Haiwen Li, Zheng Ziqiang, Tuan-Anh Vu, Hai Nguyen-Truong, Sai-Kit Yeung, Zhuang Jia, Sophia Yang, Chih-Chung Hsu, Xiu-Yu Hou, Yu-An Jhang, Simon Yang, and Mau-Tsuen Yang. 1st Workshop on Maritime Computer Vision (MaCVi) 2023: Challenge Results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 265–302, 2023.

[9] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6399–6408, April 2019.

[10] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 9396–9405. IEEE Computer Society, June 2019.

[11] Matej Kristan, Vildana Sulić Kenk, Stanislav Kovačič, and Janez Perš. Fast Image-Based Obstacle Detection from Unmanned Surface Vehicles. *IEEE Transactions on Cybernetics*, 46(3), 2016.

[12] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022.

[13] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation, December 2022.

[14] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *International Conference on Learning Representations*, October 2021.

[15] Shailesh Nirgudkar, Michael DeFilippo, Michael Sacarny, Michael Benjamin, and Paul Robinette. MassMIND: Massachusetts Maritime INfrared Dataset, September 2022.

[16] Dalei Qiao, Guangzhong Liu, Wei Li, Taizhi Lyu, and Juan Zhang. Automated Full Scene Parsing for Marine ASVs Using Monocular Vision. *Journal of Intelligent & Robotic Systems*, 104(2):1–20, 2022.

[17] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-Octob, 2019.

[18] L Yao, D Kanoulas, Z Ji, and Y Liu. ShorelineNet: An Efficient Deep Learning Approach for Shoreline Semantic Segmentation for Unmanned Surface Vehicles. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[19] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection, July 2022.

[20] Lojze Žust and Matej Kristan. Temporal Context for Robust Maritime Obstacle Detection. In *2022 IEEE/RJS International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[21] Lojze Žust, Janez Perš, and Matej Kristan. LaRS: A Diverse Panoptic Maritime Obstacle Detection Dataset and Benchmark. In *ICCV 2023*, 2023.