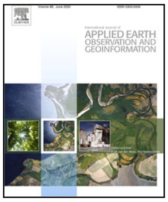
Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Multispectral panoptic segmentation: Exploring the beach setting with worldview-3 imagery

Osmar Luiz Ferreira de Carvalho [a], Osmar Abílio de Carvalho Júnior [b,*], Anesmar Olino de Albuquerque [b], Nickolas Castro Santana [b], Díbio Leandro Borges [a], Argelica Saiaka Luiz [b], Roberto Arnaldo Trancoso Gomes [b], Renato Fontes Guimarães [b]

[a] University of Brasilia, Department of Computer Science, Campus Universitario Darcy Ribeiro, Brasilia, 70910-900, Federal District, Brazil
[b] University of Brasilia, Department of Geography, Campus Universitario Darcy Ribeiro, Brasilia, 70910-900, oederal District, Brazil

ARTICLE INFO

Keywords:
Deep learning
Dataset
Worldview-3
Segmentation
High-resolution image
Remote sensing

ABSTRACT

Panoptic segmentation is a recent and powerful task that tackles individual object recognition (“things”) and multiple backgrounds (“stuff”) simultaneously. Remote sensing studies with panoptic segmentation are still restricted and recent, with great application perspectives. In this sense, we propose the first multispectral panoptic segmentation study, considering the “thing” and “stuff” classes in the beach scenario and evaluating different sets of spectral bands. Our methodology included developing a dataset with 3800 (3200 for training, 300 for validation, and 300 for testing) with 128 × 128 spatial dimensions and eight spectral bands considering fourteen classes (6 “thing” and 8 “stuff” classes). We used WorldView-3 images from Praia do Futuro, Fortaleza, and pan-sharpening to improve spatial resolution. Five different spectral band configurations were considered: (1) all eight bands, (2) $RGB + NIR1 + NIR2$, (3) $RGB + NIR1$, (4) $RGB + NIR2$, and (5) only RGB. The model training used the Panoptic-FPN architecture with the same hyperparameter settings considering three backbones (ResNeXt-101, ResNet-101, and ResNet-50). The best result considered the ResNeXt-101 with all spectral bands. However, the results from the first four configurations were very similar, and the RGB alone was the only configuration with significantly lower results. We also evaluated 15 semantic segmentation models for a benchmark comparison for the Beach Dataset. We show in visual results that even though the semantic models may be precise, they fail at identifying unique targets, especially in crowded locations such as the beach. The panoptic segmentation allowed a necessary detailing and counting of tourist infrastructures and mapping of other background features, establishing an essential tool for inspecting beach areas.

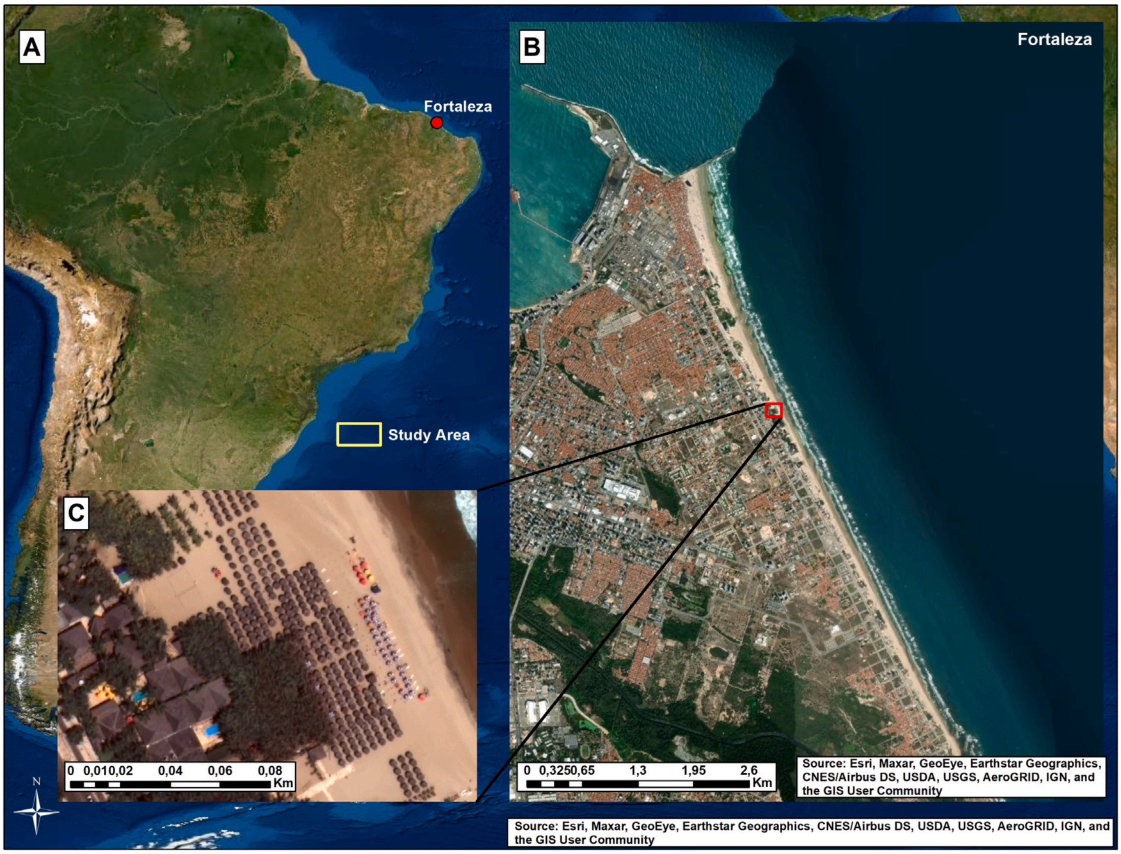
1. Introduction

Deep learning image segmentation techniques are subdivided into the semantic, instance, and panoptic. Semantic segmentation performs a pixel-wise classification, where all elements of the same class acquire the same label, not allowing the recognition of individual elements within the same class (Guo et al., 2018). This approach is like traditional land cover classification, which does not distinguish individual objects of the same class. Instance segmentation is restricted to objects and allows individuals to distinguish all elements of the same class (Hafiz and Bhat, 2020). This approach enables rapid quantification of objects and the extraction of statistical data about their size and shape. Panoptic segmentation unifies semantic and instance segmentation. In this regard, panoptic segmentation is the most advanced segmentation task. The original implementation (Kirillov et al., 2019) uses two branches: (1) instance segmentation (for identifying “things”) and (2) semantic segmentation (for identifying “stuff”). The computer vision community defines objects as “things” and amorphous background elements as “stuff” (Caesar et al., 2018). Panoptic segmentation is promising in orbital or aerial remote sensing studies due to integrating the best qualities of instance and semantic segmentation.

Many studies have been performed on panoptic segmentation for ground-level RGB images (Cheng et al., 2020; Xiong et al., 2019), medical images (Cha et al., 2021; Yu et al., 2020; Zhang et al., 2018), and videos (Kim et al., 2020; Qiao et al., 2021). Panoptic segmentation was first explored in orbital or aerial remote sensing data (top-view

* Corresponding author.
E-mail addresses: osmarcarvalho@ieee.org (O.L.F. de Carvalho), osmarjr@unb.br (O.A. de Carvalho Júnior), anesmar@ieee.org (A.O. de Albuquerque), nickolas.santana@unb.br (N.C. Santana), dibio@unb.br (D.L. Borges), argelica.saiaka@aluno.unb.br (A.S. Luiz), robertogomes@unb.br (R.A.T. Gomes), renatofg@unb.br (R.F. Guimarães).

https://doi.org/10.1016/j.jag.2022.102910
Received 19 April 2022; Received in revised form 4 July 2022; Accepted 5 July 2022
Available online 21 July 2022



Source: Esri, Maxar, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

Source: Esri, Maxar, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

Fig. 1. Study Area, in which (A) shows the region in Brazil, (B) shows a more detailed zoom of the beach area considered in this study, and (C) shows a larger zoom to show what kind of elements is visible with the WorldView-3 images.

images) only recently, presenting few studies (Garnot and Landrieu, 2021; de Carvalho et al., 2022b; De Carvalho et al., 2022; Khoshboresh-Masouleh and Shah-Hosseini, 2021; Hua et al., 2021). Among those studies, Garnot and Landrieu (2021) used remote sensing peculiarities in terms of spectral bands. However, the authors considered a dataset containing only "thing" classes. Khoshboresh-Masouleh and Shah-Hosseini (2021) evaluated the change detection in very high-resolution Google Earth images but only considered the building class (things). de Carvalho et al. (2022b) developed a dataset with "thing" and "stuff" classes, but they used an aerial image only containing the RGB channels, being very similar to traditional ground-level images. Finally, Hua et al. (2021) used datasets previously designed for instance segmentation but adapted for the panoptic segmentation task, considering only RGB images.

Therefore, the multispectral imaging dataset has not yet been explored in panoptic segmentation. Satellite or aircraft-based images often present many different characteristics, such as large spatial dimensions, varying number of channels, image format, and georeferencing (Carvalho et al., 2021). Furthermore, in the field of remote sensing technologies for monitoring the Earth's surface, there is a wide variety of images (multispectral, hyperspectral, Synthetic Aperture Radar (SAR), and thermal) coming from different platforms (satellites, Unmanned Aerial Vehicles (UAV), and aerial images). The particularities of orbital and aerial images differ from datasets produced by the computer vision community such as Common Objects in Context (COCO) (Lin et al., 2014), Mapillary Vistas (Neuhold et al., 2017), Cityscapes (Cordts et al., 2016), which contains Red, Green, and Blue (RGB) images at a ground level. Thus, the development of new tasks, such for instance segmentation (He et al., 2020), panoptic segmentation (Kirillov et al., 2019) and eventual novel methods (Bolya et al., 2020; Mohan and Valada, 2021; Gao et al., 2021) are all designed in the first moment for those traditional ground-level RGB images. Therefore the orbital and aerial image peculiarities require specific software and methodologies to extract the most out of it since even preliminary stages such as generating image tiles with a specific size and annotation format may be challenging (Li et al., 2021; de Carvalho et al., 2022b). Besides, most software that is openly available today, such as Facebook's Detectron2 (Wu et al., 2019) is designed with specifications for RGB images with three channels in conventional formats such as Joint Photographics Experts Group (JPEG) and Portable Network Graphics (PNG). Adapting those configurations may not be straightforward, making the usage of some new methods much harder in the satellite or aircraft-based remote sensing.

The continuous monitoring and inspection of tourist activity along the beaches is essential for achieving effective public and environmental policies. In this context, panoptic segmentation from the remote sensing images can facilitate the inspection process. However, few beach studies used remote sensing data with deep learning. Besides, the beach scene is mainly composed of small objects, which are represented by few pixels even with high-resolution images, being a significant challenge. Deep learning models generally perform poorly on small targets due to their noisy representation and confusion with other targets (de Carvalho et al., 2021; Tong et al., 2020).

This study aims to introduce panoptic segmentation with multispectral remote sensing data, providing theory, application, and methods contributions, as follows:

1. We aim to verify the importance of band selection within the panoptic segmentation task and compare the conceptual results of panoptic, instance, and semantic segmentation.
2. A viable and state-of-the-art application for beach inspection, being the first study to explore the panoptic segmentation in

the beach setting, providing a novel dataset comprising thirteen classes and benchmark results for panoptic, and semantic segmentation.

3. The panoptic segmentation task was initially developed for RGB images, and the present research carried out changes in the original code to allow the joint processing of a varying number of spectral bands. Besides, this study adjusted the ResNeXt-101 backbone for Panoptic-FPN.

2. Material and methods

2.1. Study area

The study area is located in the Praia do Futuro region, Fortaleza, Brazil, with intense tourist and economical activity. Fig. 1A shows the study area highlighted in yellow borders, and Fig. 1B shows a zoomed area containing tourist umbrellas, beach umbrellas, suns, straw sun, trees, buildings, and swimming pools.

2.2. Image acquisition and annotations

We used WorldView-3 images provided by the European Space Agency with a total area of 400 km^2. The high-resolution WV-3 images contain eight (1.24 m) spectral bands and a panchromatic band (0.3 m). We applied the Gram–Schmidt pan-sharping method to obtain a high-resolution color image, conjugating the spatial information from the panchromatic band and spectral information from the multispectral bands.

Geographic Information System (GIS) specialists performed manual annotations considering fourteen distinct features, all listed in the Table 1. Six of these classes were “things”, and eight were “stuff” categories. The most numerous class was the straw umbrella, with nearly 4000 distinct polygons and nearly no pixels with no classes. Fig. 2 shows the annotation pattern, containing three examples of each interest class demarcated by a colored polygon.

The images were cropped into smaller image tiles with their corresponding annotations in the COCO format, which is the standard format for the Detectron2’s Panoptic-FPN model. These annotations require JSON files with specifications for each image, containing the information regarding the “thing” classes (such as bounding boxes) and “stuff” classes. The conversion of the GIS data to the panoptic format used the software developed by de Carvalho et al. (2022b), considering a GIS attribute table where each polygon has two columns with the class value and the polygon value (or unique IDs), in case of the thing category. The software requires point shapefiles to generate the smaller samples, in which each point is the centroid of the frame. We chose a sample size of 128 × 128 pixels. We used the image from 2017 to generate all training samples totaling 3,200. Using the image from 2018, we chose the validation and test samples, 300 of each. Choosing points manually is frequently better than random since we can select high-priority areas. Since the validation and test samples are in the same image, we assured that there was no overlap between them.

2.3. Deep learning experiments

The experiments were subdivided into panoptic and semantic segmentation. The panoptic segmentation approach considered an analysis of different spectral band compositions using three models. Then, we evaluated our dataset using the semantic segmentation task considering 15 models for the best spectral band composition. Note that there is no need to perform an isolated instance segmentation approach since we can retrieve instance-only results from the panoptic models. All experiments were conducted on a computer with an NVIDIA RTX 3090 graphics card with 24 GB RAM and an i9 processor.

Table 1
Categories (in which SP, SBU and TU stand for swimming pool, straw beach umbrella, and tourist umbrella), labels, type (thing or stuff), number of polygons and number of pixels in the Panoptic Beach Dataset.

Category	Label	Type	Polygons	Pixels
Background	0	–	–	–
Ocean	1	Stuff	–	47,799,957
W. Sand	2	Stuff	–	3,874,445
Sand	3	Stuff	–	7,944,837
Road	4	Stuff	–	2,113,790
Vegetation	5	Stuff	–	1,814,849
Grass	6	Stuff	–	1,814,967
Sidewalk	7	Stuff	–	1,360,113
Vehicle	8	Thing	531	52,196
SP	9	Thing	55	33,118
Construction	10	Thing	457	1,097,025
SBU	11	Thing	3,653	247,723
TU	12	Thing	805	45,129
Crosswalk	13	Thing	46	34,138

2.3.1. Panoptic segmentation

This research aims to compare different configurations of spectral bands using the Panoptic-FPN (Kirillov et al., 2019) model, the pioneer model in panoptic segmentation studies. The primary motivation for using FPN to predict semantic segmentation is to establish a simple, single-network baseline, which allows executing the semantic and instance segmentation steps in a chained way and considering a joint task of panoptic segmentation. This model is present in the Detectron2 software that allows implementation and contains detailed documentation for future improvements and replication. The software uses the Pytorch library, which is also widely used, making the code easier to understand. The Panoptic-FPN model comprises two branches: (1) instance segmentation and (2) semantic segmentation. Both branches use a common structure, which is the FPN. The instance segmentation branch uses a Mask-RCNN model and aims to identify the “things” elements (He et al., 2020). The semantic segmentation branch uses upsampling in the feature maps and targets the “stuff” classes. The two branches are combined using a simple heuristic method for combining the instance level “thing” predictions and the background “stuff” elements. The Detectron2 (Wu et al., 2019) software is the most appropriate to do experiments in this task because the documentation is very robust. Besides, previous studies proposed modifications and adaptations for well functioning in remote sensing datasets (Carvalho et al., 2021; de Carvalho et al., 2022b), which has not yet been done to other models. We evaluated three backbones using the Panoptic-FPN model, namely the ResNeXt-101, ResNet-101, and ResNet-50.

We had to leverage the Detectron2 software for working with TIFF multispectral images, allowing it to use a varying number of input bands. Our experiments considered five tests from the pan-sharpening images, considering: (1) all eight spectral bands, (2) $RGB + NIR1 + NIR2$, (3) $RGB + NIR1$, (4) $RGB + NIR2$, and (5) only RGB. All trained models use the same specifications, apart from the input spectral dimensions. The z-score normalization for each channel allowed a faster convergence in the training phase.

Regarding the model hyperparameters, we used: (a) stochastic gradient descent (SGD) optimizer; (b) 0.0005 learning rate; (c) 150,000 iterations; (d) anchor boxes with sizes 8, 16, 32, 64, 128; (e) three aspect ratios (0.5, 1, 2); (f) one image per batch. We evaluated the validation set for every 5,000 iterations, in which the final model considered the best Panoptic Quality results. Besides, we considered the following augmentation steps: (a) random vertical flip (probability chance of 50%), (b) random horizontal flip (probability chance of 50%), and (c) rescaling the image dimensions to 800 × 800 pixels. The augmentation processes in the training set resulted in 9600 different image combinations in the training phase.

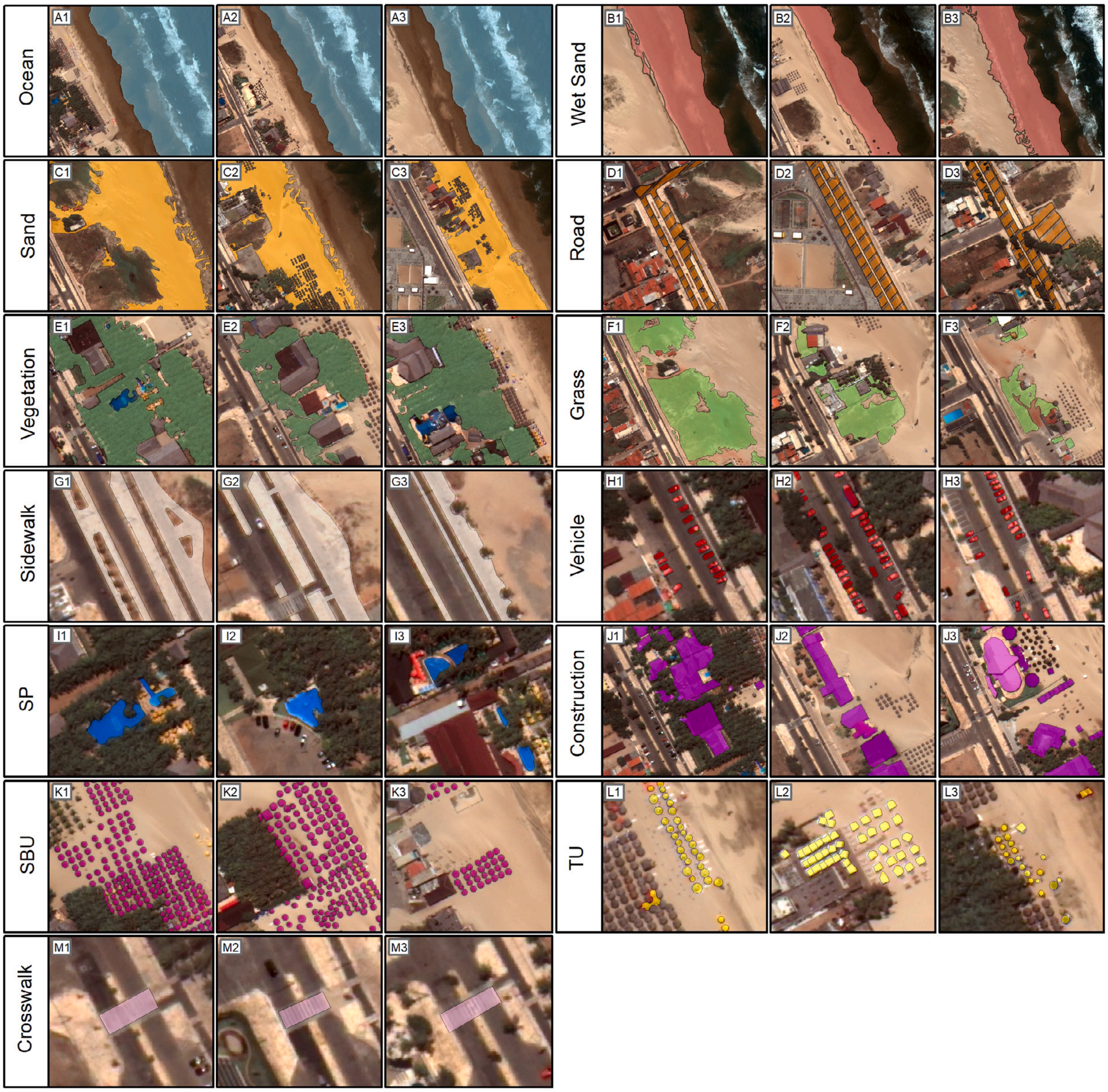


Fig. 2. Examples of annotations for each class. The highlighted segments show the class corresponding to the written labels, in which we considered: Ocean, Wet Sand, Sand, Road, Vegetation, Grass, Sidewalk, Vehicle, Swimming Pool (SP), Construction, Straw Beach Umbrella (SBU), Tourist Umbrella (TU), and Crosswalk. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.3.2. Semantic segmentation

Even though the panoptic and semantic tasks present different objectives, the comparison is valid for analyzing the pros and cons of each approach. The semantic segmentation field has been much more explored in the remote sensing community, and the various models and implementations are better documented. The way we have constructed our dataset enables researchers to use different tasks. In the semantic segmentation analysis, we compared five architectures (U-Net Ronneberger et al., 2015, U-Net++ Zhou et al., 2018, DeepLabv3+ Chen et al., 2018, FPN Lin et al., 2017, LinkNet Chaurasia and Culurciello, 2017) and three backbones (Efficient-net-B7 Tan and Le, 2019, ResNet-101 He et al., 2016, and ResNeXt-101 Xie et al., 2017). All models considered the same loss function (cross-entropy) and hyperparameter settings, including 0.0005 learning rate, batch size of 25, Adam optimizer, and 300 epochs.

2.4. Accuracy analysis

This study considers panoptic and semantic segmentation models that show remarkable differences. The Panoptic segmentation task involves three distinct types of evaluations: “stuff”, “thing”, and panoptic metrics. The per-pixel metrics suitable for semantic segmentation are the same as for the “stuff” evaluation. The difference is that the panoptic model only considers the “stuff” classes (for the stuff evaluation), and semantic segmentation considers all classes.

The “stuff” evaluation considered: (a) mean Intersection over Union, (b) frequency weighted IoU (fwIoU), (c) mean Accuracy (mAcc), and (d) pixel accuracy (pAcc). The mIoU corresponds to the mean average from all classes considering their area of intersection ($A \cap B$) divided by the area of union ($A \cup B$), in which A is the deep learning prediction and B is the ground truth. The fwIoU is similar but assigns weights according to the number of representations instead of a mean average. The pixel accuracy is simply the number of correctly classified pixels

Table 2
Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (RQ) results for the ResNet-50, ResNet-101, and ResNeXt-101 backbones. The best results are in bold.

Spectral bands	PQ	SQ	RQ
ResNeXt-101			
All	**65.90**	**81.23**	**80.83**
RGB+NIR1+NIR2	65.43	80.90	80.49
RGB+NIR1	64.82	80.67	79.94
RGB+NIR2	64.37	80.65	79.50
RGB	61.23	79.32	76.89
ResNet-101			
All	64.88	79.51	80.80
RGB+NIR1+NIR2	64.60	79.89	80.40
RGB+NIR1	64.41	80.10	79.94
RGB+NIR2	64.30	80.87	79.13
RGB	61.21	79.45	76.66
ResNet-50			
All	63.04	78.54	79.71
RGB+NIR1+NIR2	62.65	79.52	78.33
RGB+NIR1	62.52	78.91	78.73
RGB+NIR2	62.20	79.90	77.48
RGB	60.87	79.18	76.51

Table 3
Metric analysis for the "stuff" categories, considering Mean Intersection over Union (mIoU$_{stuff}$), frequency weighted (fwIoU$_{stuff}$), mean accuracy (mAcc$_{stuff}$), and pixel accuracy (pAcc$_{stuff}$) results for semantic segmentation in the Beach dataset. The best results for each class are in bold.

Spectral bands	mIoU$_{stuff}$	fwIoU$_{stuff}$	mAcc$_{stuff}$	pAcc$_{stuff}$
ResNeXt-101-32 × 8d				
8 bands	85.35	88.48	91.88	93.74
RGB+NIR1+NIR2	**85.58**	88.63	**92.02**	93.84
RGB+NIR1	85.10	88.34	91.91	93.63
RGB+NIR2	84.32	87.65	91.24	93.29
RGB	81.49	85.89	89.88	92.17
ResNet-101				
8 bands	84.76	88.23	91.41	93.57
RGB+NIR1+NIR2	85.53	88.78	91.93	93.90
RGB+NIR1	85.57	**88.83**	91.96	**93.93**
RGB+NIR2	85.10	88.47	91.92	93.73
RGB	82.27	86.44	90.02	92.54
ResNet-50				
8 bands	84.02	87.37	90.93	93.08
RGB+NIR1+NIR2	84.31	87.57	91.32	93.21
RGB+NIR1	83.58	86.88	91.47	92.74
RGB+NIR2	82.88	86.53	90.62	92.56
RGB	82.24	86.10	90.19	92.33

Table 4
Intersection over Union (IoU) results for the "stuff" categories per class in the Beach dataset, in which (1) ocean, (2) Wet Sand, (3) Sand, (4) Road, (5) Vegetation, (6) Grass, and (7) sidewalk. The best results for each class are in bold.

Spectral bands	1	2	3	4	5	6	7
ResNeXt-101							
All	96.40	87.72	92.49	**90.09**	83.84	74.00	76.54
RGB+NIR1+NIR2	96.73	89.04	92.46	89.48	83.74	**74.48**	**77.06**
RGB+NIR1	97.69	90.52	92.14	88.97	82.50	72.32	76.00
RGB+NIR2	95.06	85.20	92.24	89.19	83.56	72.46	76.05
RGB	94.93	80.84	91.49	83.29	82.48	63.62	74.75
ResNet-101							
All	**98.64**	90.91	91.83	87.01	82.34	71.31	74.13
RGB+NIR1+NIR2	98.63	**92.16**	92.22	87.01	83.39	72.90	75.83
RGB+NIR1	97.95	91.31	**92.71**	87.57	83.40	73.23	76.79
RGB+NIR2	96.78	88.06	92.65	87.46	**84.62**	73.39	75.92
RGB	96.33	84.21	91.67	80.97	82.40	69.60	72.54
ResNet-50							
All	97.96	90.30	90.97	85.81	81.84	72.19	73.81
RGB+NIR1+NIR2	97.79	87.87	91.05	86.04	82.82	73.70	75.21
RGB+NIR1	97.04	87.94	90.53	86.21	82.57	73.10	73.21
RGB+NIR2	95.74	86.29	91.34	85.44	82.56	71.10	73.51
RGB	95.51	84.67	90.94	83.06	82.59	68.57	74.54

Table 5
COCO metrics for the "thing" categories in the Beach Dataset considering the usage of different spectral bands: (1) all (eight spectral bands), (2) Red, Green, and Blue (RGB) with NIR1 and NIR2, (3) RGB with NIR1, (4) RGB with NIR2, and (5) only RGB. The best results for Box and Mask are in bold.

Spectral bands	Type	AP	AP_{50}	AP_{75}
ResNeXt-101				
All	Box	**60.05**	83.66	59.51
	Mask	53.39	82.23	55.35
RGB+NIR1+NIR2	Box	59.31	83.00	**63.21**
	Mask	**54.67**	81.85	56.38
RGB+NIR1	Box	59.48	82.89	66.35
	Mask	54.52	79.92	**59.30**
RGB+NIR2	Box	59.11	83.43	60.61
	Mask	54.19	81.82	55.45
RGB	Box	59.52	83.27	64.74
	Mask	53.70	81.19	58.70
ResNet-101				
All	Box	57.30	87.22	61.35
	Mask	50.49	85.18	54.05
RGB+NIR1+NIR2	Box	56.69	**87.30**	60.65
	Mask	51.23	**85.52**	54.23
RGB+NIR1	Box	58.38	85.01	61.96
	Mask	52.36	83.56	54.84
RGB+NIR2	Box	57.52	82.93	61.34
	Mask	53.13	80.25	58.95
RGB	Box	56.34	83.84	61.77
	Mask	48.97	82.55	50.82
ResNet-50				
All	Box	51.54	86.31	52.78
	Mask	46.24	83.92	45.30
RGB+NIR1+NIR2	Box	52.72	85.35	55.78
	Mask	44.87	83.13	45.77
RGB+NIR1	Box	50.85	82.67	53.78
	Mask	46.09	80.49	50.61
RGB+NIR2	Box	51.10	83.51	53.15
	Mask	45.82	82.41	49.62
RGB	Box	50.21	82.53	53.13
	Mask	44.43	78.60	47.84

divided by the total number of pixels, and the mean Accuracy is the average among the accuracies from all different classes. Due to the differences between the types of segmentation, we define mIoU to denote the mean across all categories (semantic segmentation) and mIoU$_{stuff}$ to denote the mean across the stuff categories (panoptic segmentation). The semantic segmentation evaluation considered the mIoU and mIoU$_{stuff}$ metrics.

In the "thing" evaluation (instance segmentation), the COCO Average Precision (AP) is the primary metric not only in the COCO challenge (Lin et al., 2014) but also in many studies (Bolya et al., 2020; Cai and Vasconcelos, 2018; Gao et al., 2021; He et al., 2020; Huang et al., 2019). The AP is a ranking metric expressed as the area under the precision–recall curve. In order to calculate precision and recall, we must identify the correctly predicted elements. Thus, the COCO metric uses different IoU thresholds between the predicted and ground truth bounding boxes. The primary AP metric considers 10 IoU thresholds, from 0.5 to 0.95, with 0.05 steps. Besides, to exclusively evaluate a more and less strict version, the AP50 and AP75 use the 0.5 and 0.75 thresholds.

Finally, the Panoptic metrics are: Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (de Carvalho et al., 2022b;

Table 6
COCO metrics for the “thing” categories in the Beach Dataset considering the usage of different spectral bands: (1) all (eight spectral bands), (2) Red, Green, and Blue (RGB) with NIR1 and NIR2, (3) RGB with NIR1, (4) RGB with NIR2, and (5) only RGB. The evaluated classes are: (8) vehicle, (9) SP, (10) construction, (11) straw beach umbrella, (12) tourist umbrella, and (13) crosswalk. The best results for Box and Mask are in bold.

Spectral bands	Type	8	9	10	11	12	13
ResNeXt-101							
All	Box	60.67	58.23	56.75	39.42	74.25	70.72
	Mask	54.47	49.17	54.61	33.75	63.65	64.71
RGB+NIR1+NIR2	Box	61.36	59.61	58.38	39.18	73.65	63.68
	Mask	56.31	49.71	57.13	35.38	66.80	62.68
RGB+NIR1	Box	61.05	57.43	58.38	35.58	71.19	73.26
	Mask	55.39	47.89	56.98	33.20	63.62	70.05
RGB+NIR2	Box	58.85	57.56	57.93	37.76	72.55	70.01
	Mask	53.98	46.84	55.56	34.38	67.70	66.65
RGB	Box	61.96	56.71	58.09	39.62	70.06	70.73
	Mask	55.10	46.03	55.50	35.86	65.96	63.77
ResNet-101							
All	Box	59.70	54.96	58.15	47.16	58.17	65.63
	Mask	52.63	44.63	54.69	41.21	53.52	56.22
RGB+NIR1+NIR2	Box	60.14	53.28	59.02	**48.74**	61.69	57.27
	Mask	52.38	45.91	57.13	**45.03**	55.84	51.12
RGB+NIR1	Box	58.44	56.08	58.84	44.65	66.00	**66.26**
	Mask	52.06	49.55	**58.00**	40.32	59.06	55.23
RGB+NIR2	Box	59.06	**56.42**	**59.68**	45.18	69.99	54.79
	Mask	53.02	50.66	56.90	41.44	63.45	53.34
RGB	Box	59.23	58.94	55.48	45.23	55.12	64.03
	Mask	49.36	50.45	53.86	42.08	47.51	50.55
ResNet-50							
All	Box	59.68	51.26	53.09	40.52	46.49	58.17
	Mask	51.74	42.16	50.50	32.28	40.12	**59.63**
RGB+NIR1+NIR2	Box	57.08	53.47	53.81	37.43	51.90	62.65
	Mask	48.82	41.97	50.68	31.94	43.65	52.15
RGB+NIR1	Box	59.08	55.46	52.16	29.12	49.89	59.37
	Mask	**53.84**	48.90	50.27	26.97	43.47	53.09
RGB+NIR2	Box	**62.48**	49.46	54.28	35.48	49.13	55.75
	Mask	53.51	48.18	52.01	28.31	41.63	51.28
RGB	Box	59.99	49.04	52.97	32.85	47.77	58.65
	Mask	50.47	44.34	49.38	28.98	37.70	55.60

Table 7
Semantic segmentation model metrics considering all spectral brands. The best results are in bold.

Architecture	Backbone	mIoU	mIoU_{stuff}
U-Net	Eff-B7	75.56	85.63
	ResNeXt-101	73.25	84.05
	ResNet-101	71.49	83.19
DeepLabv3+	Eff-B7	75.68	85.44
	ResNeXt-101	73.67	83.95
	ResNet-101	71.73	83.17
U-Net++	Eff-B7	75.71	85.35
	ResNeXt-101	73.47	83.84
	ResNet-101	69.13	82.09
LinkNet	Eff-B7	74.74	84.42
	ResNeXt-101	73.78	83.45
	ResNet-101	69.11	81.28
FPN	Eff-B7	**77.44**	**85.67**
	ResNeXt-101	73.52	84.21
	ResNet-101	72.61	83.17

Gao et al., 2021; Kirillov et al., 2019; Mohan and Valada, 2021). The PQ is the primary metric for this task, being the multiplication of the SQ by RQ, where SQ is $\frac{\sum_{(pred,GT)\in TP} IoU(pred,GT)}{|TP|}$ and RQ is $\frac{TP}{|TP|+\frac{1}{2}|FP|+\frac{1}{2}|FN|}$, in which pred, GT, TP, FN, and FP stand for the deep learning prediction, ground truth data, true positives (elements with an IoU greater than 0.5), false negatives, and false positives, respectively.

3. Results

This section is subdivided in three parts: (1) panoptic segmentation evaluation, (2) semantic segmentation evaluation, and (3) visual results that show the differences of semantic, instance, and panoptic segmentation.

3.1. Panoptic segmentation evaluation

3.1.1. Panoptic metrics

Table 2 lists the panoptic segmentation results for the PQ, SQ, and RQ metrics. Our models were selected from the validation set based on the best PQ performance, being the main metric for evaluation. ResNeXt-101 was the best backbone for all band compositions, followed by ResNet-101. Besides, using all spectral bands provided the best PQ, SQ, and RQ results. Nonetheless, the composition with only RGB bands obtained a significantly lower accuracy, showing that the NIR1 and NIR2 bands are significant for classifying typical targets in the beach scenario.

3.1.2. Stuff evaluation results

Table 3 lists the macro results for the stuff categories. In contrast to the panoptic metrics, using all bands did not yield the maximum results, even though they were very close. The panoptic models use a loss function that considers many elements, and there may be a tradeoff between some of them to yield the best results. Moreover, here, the RGB-only composition was considerably lower than the rest. When analyzing this behavior per class (Table 4), Wet Sand, Road, and Grass classes presented considerably lower results. The difference of RGB-only metrics considering the other classes did not show much difference, showing that depending on the classes being evaluated, the RGB bands can be satisfactory. Moreover, an exciting result is that most of the classes showed values above 80% in IoU, showing that the targets in the beach setting are suitable for expanding monitoring with deep learning methods.

3.1.3. Thing evaluation results

Table 5 lists the COCO metrics (AP, AP_{50}, AP_{75}) results for the “thing” classes, and Table 6 lists the AP results per class. The ResNeXt-101 presented the best values for all combinations considering the AP metric. Nonetheless, even though the ResNet-101 and ResNet-50 presented the worst values for the main metric, the AP_{50} was superior for many configurations. The influence of band composition on the things classes was much less significant, in which none of the classes had a significantly worst behavior. The main factor for results for the thing classes is the backbone.

3.2. Semantic segmentation results

This section shows the benchmark results for the Beach Dataset considering the semantic segmentation task. Table 7 lists the mIoU metrics for each model. Note that the metrics shown here are different from the mIoU_{stuff} since we are now considering all classes. For an easier comparison of this section with the panoptic models, we also incorporated the mIoU_{stuff}. The FPN model presented the highest mIoU (77.44) and mIoU_{stuff} (85.67) results. These results are slightly higher than the mIoU_{stuff} from the best panoptic model (85.58). Within the different architectures, the Efficient-net-B7 was the best backbone, followed by ResNeXt-101. The differences across architectures were smaller than across backbones.

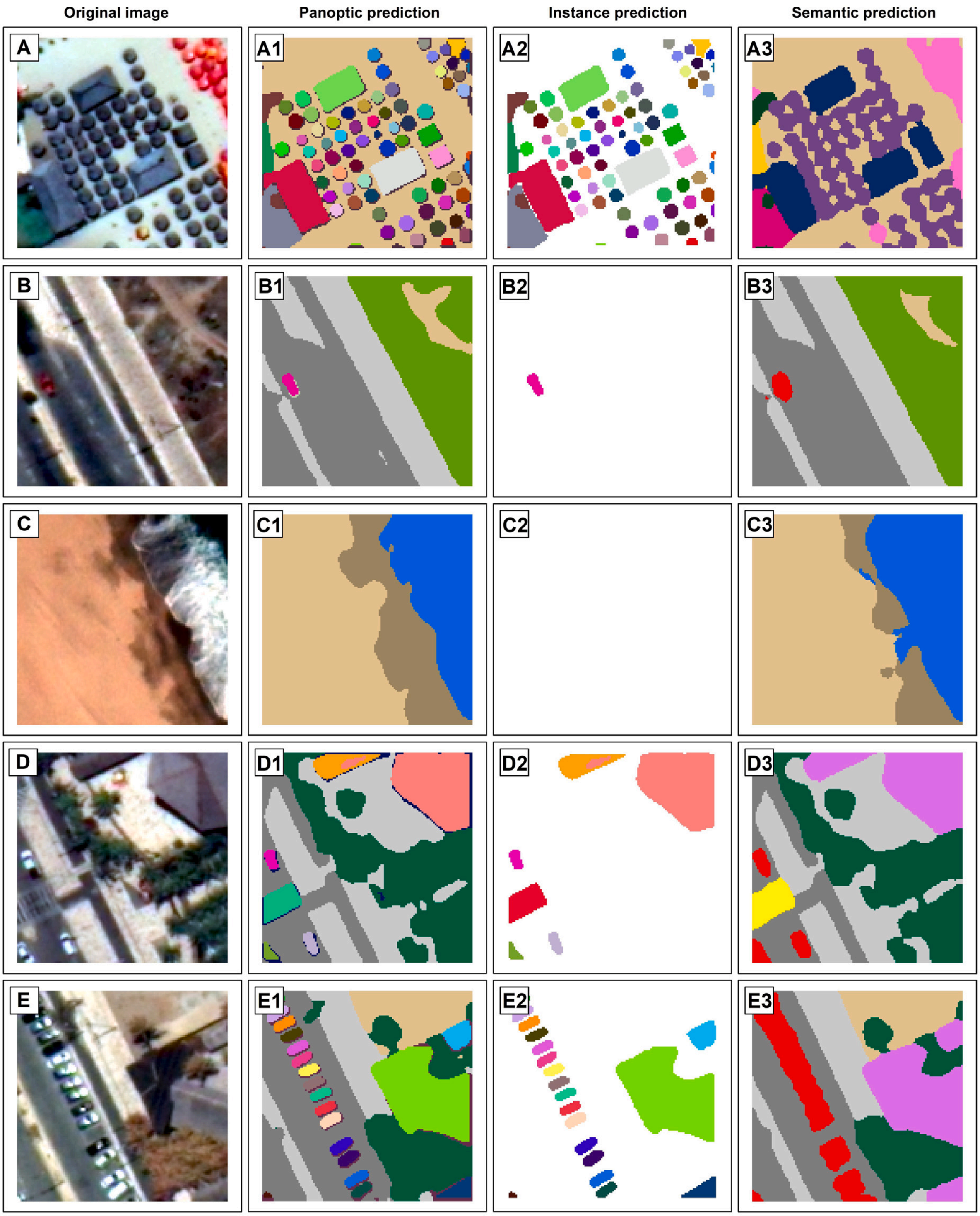


Fig. 3. Six examples with the original image (considering the RGB bands) and the prediction. The predictions maintained the same colors for the stuff classes, and each thing class presents a varying color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. *Visual results*

Fig. 3 shows five examples from the test set considering the original image, and the panoptic, instance, and semantic predictions. The concept of panoptic segmentation changes the presentation configuration where each stuff category has a unique color, while thing categories has unique values, and consequently colors for each object. Therefore, this technique brings a new approach to the cartographic representation of land use/land cover maps, which usually adopts a pixel classification. Besides, the Panoptic-FPN model generates a JSON for each predicted image, retrieving more attributes of each element, such as the bounding box and the class, favoring other ways of visualizing the data. We removed the bounding boxes for visual purposes, as overlapping information would make the image cluttered. Even though some metrics do not seem very high, mainly related to the nature of small objects, the model can predict crowded correctly and numerous elements in a single image from a visual perspective. The results show that “thing” targets close to each other tend to merge the predictions, making it very difficult to separate different instances, as shown in Fig. 3A3 and E3. Moreover, the beach setting has many amorphous elements, which are disconsidered by the instance predictions, demonstrating that the panoptic segmentation aggregates the benefits of both methods.

4. Discussion

This research adapted the original Panoptic-FPN code of the panoptic segmentation to perform the joint processing of all available multispectral bands and to couple the ResNeXt-101 backbone, considering multiclass “thing” and “stuff”. The panoptic segmentation task presents a much larger complex in the model design compared to the instance and semantic segmentation task. The loss function encompasses both instance and semantic segmentation losses. The instance segmentation also presents a complexity since it involves segmentation loss, bounding box regression loss, and classification loss. All of those elements associated with a large number of classes may present fluctuations when comparing the models. For example, the AP metrics for the individual targets may show a higher metric for a single target given a model, this is why it is essential to analyze the macro metrics since they provide an overall view of the model, and they should primarily focus on identifying and building the best model.

The investigation used the Panoptic-FPN model of Detectron2 software proposed by Meta Artificial Intelligence Research and used and tested globally. Unlike semantic segmentation, which has several models developed, panoptic segmentation has a restricted number of models that still lack detailed documentation. Thus, we compared the Panoptic-FPN architecture with three different backbones (ResNeXt-101, ResNet-101, and ResNet-50). Besides, this study evaluates different compositions of spectral bands within the panoptic segmentation. This approach allows quantifying the gain in accuracy with the use of multispectral data. The Panoptic-FPN model using ResNeXt-101 backbone and all bands obtained better results in all panoptic metrics (PQ, SQ, and RQ). Even though the panoptic segmentation results were similar using all bands, $RGB + NIR1 + NIR2$, $RGB + NIR1$, and $RGB + NIR2$, the spectral characteristics from remotely sensed data can enhance the results significantly when compared to the traditional RGB channels with nearly 5% worse than the rest in the PQ. This deep learning survey was also the first to use multiple remote sensing targets in the beach setting.

The results can guide other researchers in selecting bands for future studies in a beach scenario. Similar studies achieved some complementary results to our findings. Carvalho et al. (2021) compared RGB and all bands from the Landsat −8 sensor for center pivot mapping using instance segmentation models, where the authors found that using all bands had a 3% increase in the metric AP. In a semantic segmentation study, Barros et al. (2022) found that the NIR band was almost sufficient to map vineyards. Furthermore, the remote sensing field has many opportunities for studies using multichannel inputs, especially considering time series and multispectral data (de Albuquerque et al., 2021b; Carvalho et al., 2021). Recent studies have used a time-series sequence as the input, in which each time represents a different channel (de Albuquerque et al., 2021a; de Bem et al., 2021; Li et al., 2020). In many of these studies, we can see that introducing new information is complementary to deep learning studies until it reaches the point when new information is redundant. This analysis is significant, allowing primary bands to be selected instead of all available bands, reducing the computational cost.

However, the panoptic segmentation task is challenging to compare with other deep learning and machine learning methods because the evaluation criteria are very different from other methodologies, such as instance and semantic segmentation that do not have the categories together of “stuff” and “thing”. Recently, De Carvalho et al. (2022) proposed a novel way to approach panoptic segmentation with semantic segmentation models, which could also be an alternative way to address a more robust model comparison in future studies. Nonetheless, despite the difficulty in a metric-wise comparison, we can evaluate the benefits of each task, especially when using a visual comparison. As shown in the results section, the IoU results for semantic segmentation models exhibit to be very accurate for some classes. However, the beach targets are crowded in many cases, such as the beach straw umbrellas (the most numerous category). For targets like this, it is very hard to retrieve relevant information, such as the number of individual elements, since the semantic predictions tend to aggregate many of the targets that are close to each other, similar to what happens in vehicle detection (de Carvalho et al., 2022a; Mou and Zhu, 2018). On the other hand, the instance segmentation also has limitations to important classes such as sand, sidewalks, and roads. The panoptic segmentation emerges as a viable and interesting solution for handling the beach setting with many objects and backgrounds.

Our novel proposed dataset presents different characteristics than other panoptic segmentation datasets, formed by RGB and ground-level images (Cityscapes, COCO, and Mapillary Vistas). Although the BSB Aerial Dataset (de Carvalho et al., 2022b) consists of aerial photos for panoptic segmentation, the available channels are RGB. The present dataset contains orbital multispectral images, considering images composed of up to 8 bands from the WV-3 sensor. Changing the RGB inputs to include all available spectral bands produces better results. Using ground-level RGB dataset transfer learning for multispectral imaging can still provide some leverage to reduce the training period as low-level features such as corners are similarly represented. Even so, transfer learning between different sensor datasets is complicated as each sensor has different amounts of spectral bands with different characteristics of the spectrum range, providing less accurate transferability. A possible solution and future studies would include building a dataset using various sensors simultaneously, covering a wide range of spectral and spatial behaviors.

Most of the classes are in the range of small objects. The small objects are a great difficulty since their representation is much less significant, bringing difficulties for classification. In many datasets such as COCO, the AP_{small} metrics is much lower than the rest. The results may be affected by the size of the objects. In a previous study in this same region, de Carvalho et al. (2021) analyzed only the class Straw Beach Umbrella with different scaling dimensions, in which they upscaled the image up to 8 times the original size. The AP metric nearly doubled by a simple operation. One of the augmentation steps in our research was to resize the image to 800 × 800 spatial dimensions, a typical and default setting in the Detectron2 software. The nature of the metrics is not favorable for achieving high results since small displacements in the bounding boxes significantly affect the metric, which does not happen for larger objects. Tong et al. (2020) made a review article on small objects, in which they stated that increasing the dimensions is one of the simplest forms of increasing results. Kisantal et al. (2019) created a method for increasing the representation of small objects. Even though this solution is up-and-coming and can indeed increase the results, in some situations, the fact that the objects are small is enough to bring the metric down, and sometimes in visual results, the prediction is very accurate.

Finally, the results proved to be entirely satisfactory in the landscape of Praia do Futuro, an important area for government inspection for having high tourist activity on public lands. Therefore, the model has immediate application in the periodic monitoring of this urban beach with constant misuse of public property. However, a limitation of the present investigation is that the model is only suitable for beaches with similar characteristics (same composition of sediments, vegetation, and tourist infrastructure). Therefore, future research should include a wider variety of beach settings, including different regions and countries. In addition, future research may test other images, such as drones.

5. Conclusion

This study was a pioneer in exploiting the panoptic segmentation tasks in the beach setting, considering high-resolution WorldView-3 images with 0.31-meter resolution and a multispectral dataset with "things" and "stuff" classes. Since most computer vision developments use RGB image datasets, we evaluated different configurations on band arrangement and found that the combination of the near infra-red (NIR) and the RGB bands can significantly improve results. In the beach setting, the main panoptic metric (PQ) differed by nearly 5%. Besides, the difference in using all eight multispectral bands with $RGB + NIR1$ and NIR2 is very shallow, and in situations with less computational resources, using fewer bands will not affect the results. Moreover, the Panoptic-FPN architecture with the ResNeXt-101 backbone performed better for panoptic metrics than the ResNet-101 and ResNet-50.

The panoptic segmentation presents a new possibility for developing inspection solutions in the beach areas. We can simultaneously retrieve crucial information about individual objects, such as their size and abundance. Future research aims to develop methodologies for mapping large regions, encompassing sliding window methods, which are still unexplored for panoptic segmentation, and using other images such as aerial images and unmanned aerial vehicles.

CRediT authorship contribution statement

Osmar Luiz Ferreira de Carvalho: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Osmar Abílio de Carvalho Júnior:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Anesmar Olino de Albuquerque:** Writing – review & editing, Visualization, Investigation, Data curation, Resources. **Nickolas Castro Santana:** Writing – review & editing, Visualization, Investigation, Data curation, Resources. **Díbio Leandro Borges:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Argelica Saiaka Luiz:** Writing – review & editing, Visualization, Investigation, Data curation, Resources. **Roberto Arnaldo Trancoso Gomes:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Renato Fontes Guimarães:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

The dataset and code can be obtained by contacting the corresponding author upon reasonable request.

Acknowledgment

This research was funded by the Conselho Nacional de Pesquisa e Desenvolvimento (grant numbers 434838/2018-7 and 305769/2017-0) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (grant number 001).

References

de Albuquerque, A.O., de Carvalho, O.L.F., e Silva, C.R., de Bem, P.P., Gomes, R.A.T., Borges, D.L., Guimarães, R.F., Pimentel, C.M.M., de Carvalho Júnior, O.A., 2021a. Instance segmentation of center pivot irrigation systems using multi-temporal SENTINEL-1 SAR images. Remote Sens. Appl.: Soc. Environ. 23, 100537. http://dx.doi.org/10.1016/j.rsase.2021.100537.

de Albuquerque, A.O., de Carvalho, O.L.F., e Silva, C.R., Luiz, A.S., Pablo, P., Gomes, R.A.T., Guimarães, R.F., de Carvalho Júnior, O.A., 2021b. Dealing with clouds and seasonal changes for center pivot irrigation systems detection using instance segmentation in sentinel-2 time series. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 8447–8457. http://dx.doi.org/10.1109/JSTARS.2021.3104726.

Barros, T., Conde, P., Gonçalves, G., Premebida, C., Monteiro, M., Ferreira, C.S., Nunes, U., 2022. Multispectral vineyard segmentation: A deep learning comparison study. Comput. Electron. Agric. 195, 106782. http://dx.doi.org/10.1016/j.compag.2022.106782.

de Bem, P.P., de Carvalho Júnior, O.A., de Carvalho, O.L.F., Gomes, R.A.T., Guimarães, R.F., Pimentel, C.M.M., 2021. Irrigated rice crop identification in Southern Brazil using convolutional neural networks and Sentinel-1 time series. Remote Sens. Appl.: Soc. Environ. 24, 100627. http://dx.doi.org/10.1016/j.rsase.2021.100627.

Bolya, D., Zhou, C., Xiao, F., Lee, Y.J., 2020. YOLACT++: Better real-time instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 1. http://dx.doi.org/10.1109/TPAMI.2020.3014297.

Caesar, H., Uijlings, J., Ferrari, V., 2018. COCO-stuff: Thing and stuff classes in context. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, UT, USA, pp. 1209–1218. http://dx.doi.org/10.1109/CVPR.2018.00132.

Cai, Z., Vasconcelos, N., 2018. Cascade R-CNN: Delving into high quality object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, UT, USA, pp. 6154–6162. http://dx.doi.org/10.1109/CVPR.2018.00644.

Carvalho, O.L.F.d., de Carvalho Júnior, O.A., Albuquerque, A.O.d., Bem, P.P.d., Silva, C.R., Ferreira, P.H.G., Moura, R.d.S.d., Gomes, R.A.T., Guimarães, R.F., Borges, D.L., 2021. Instance segmentation for large, multi-channel remote sensing imagery using mask-RCNN and a mosaicking approach. Remote Sens. 13 (1), 39. http://dx.doi.org/10.3390/rs13010039.

de Carvalho, O.L.F., de Carvalho Júnior, O.A., de Albuquerque, A.O., Santana, N.C., Guimarães, R.F., Gomes, R.A.T., Borges, D.L., 2022a. Bounding box-free instance segmentation using semi-supervised iterative learning for vehicle detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 3403–3420.

de Carvalho, O.L.F., de Carvalho Júnior, O.A., Silva, C.R.e., de Albuquerque, A.O., Santana, N.C., Borges, D.L., Gomes, R.A.T., Guimarães, R.F., 2022b. Panoptic segmentation meets remote sensing. Remote Sens. 14 (4), 965. http://dx.doi.org/10.3390/rs14040965.

de Carvalho, O.L.F., de Moura, R.d.S., de Albuquerque, A.O., de Bem, P.P., Pereira, R.d.C., Weigang, L., Borges, D.L., Guimarães, R.F., Gomes, R.A.T., de Carvalho Júnior, O.A., 2021. Instance segmentation for governmental inspection of small touristic infrastructure in beach zones using multispectral high-resolution WorldView-3 imagery. ISPRS Int. J. Geo-Inf. 10 (12), 813. http://dx.doi.org/10.3390/ijgi10120813.

Cha, J.-Y., Yoon, H.-I., Yeo, I.-S., Huh, K.-H., Han, J.-S., 2021. Panoptic segmentation on panoramic radiographs: Deep learning-based segmentation of various structures including maxillary sinus and mandibular canal. J. Clin. Med. 10 (12), 2577.

Chaurasia, A., Culurciello, E., 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing. VCIP, IEEE, pp. 1–4.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. In: Lecture Notes in Computer Science, vol. 11211, Springer, Cham, pp. 833–851. http://dx.doi.org/10.1007/978-3-030-01234-2_49.

Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.-C., 2020. 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. pp. 12475–12485.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, vol. 29, IEEE, Las Vegas, NV, USA, pp. 3213–3223. http://dx.doi.org/10.1109/CVPR.2016.350.

De Carvalho, O.L.F., De Carvalho Júnior, O.A., De Albuquerque, A.O., Santana, N.C., Borges, D.b.L., 2022. Rethinking panoptic segmentation in remote sensing: A hybrid approach using semantic segmentation and non-learning methods. IEEE Geosci. Remote Sens. Lett. 1. http://dx.doi.org/10.1109/LGRS.2022.3172207.

Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P., 2021. Res2Net: A new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. 43 (2), 652–662. http://dx.doi.org/10.1109/TPAMI.2019.2938758.

Garnot, V.S.F., Landrieu, L., 2021. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4872–4881. http://dx.doi.org/10.1109/ICCV48922.2021.00483.

Guo, Y., Liu, Y., Georgiou, T., Lew, M.S., 2018. A review of semantic segmentation using deep neural networks. Int. J. Multimedia Inf. Retr. 7 (2), 87–93.

Hafiz, A.M., Bhat, G.M., 2020. A survey on instance segmentation: state of the art. Int. J. Multimedia Inf. Retr. 9 (3), 171–189.

He, K., Gkioxari, G., Dollar, P., Girshick, R., 2020. Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 386–397. http://dx.doi.org/10.1109/TPAMI.2018.2844175.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, vol. 45, IEEE, Las Vegas, NV, USA, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Hua, X., Wang, X., Rui, T., Shao, F., Wang, D., 2021. Cascaded panoptic segmentation method for high resolution remote sensing image. Appl. Soft Comput. 109, 107515. http://dx.doi.org/10.1016/j.asoc.2021.107515.

Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X., 2019. Mask scoring R-CNN. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, Long Beach, CA, USA, USA, pp. 6402–6411. http://dx.doi.org/10.1109/CVPR.2019.00657.

Khoshboresh-Masouleh, M., Shah-Hosseini, R., 2021. Building panoptic change segmentation with the use of uncertainty estimation in squeeze-and-attention CNN and remote sensing observations. Int. J. Remote Sens. 42 (20), 7798–7820. http://dx.doi.org/10.1080/01431161.2021.1966853.

Kim, D., Woo, S., Lee, J.-Y., Kweon, I.S., 2020. Video panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9859–9868.

Kirillov, A., He, K., Girshick, R., Rother, C., Dollar, P., 2019. Panoptic segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, Long Beach, CA, USA, USA, pp. 9396–9405. http://dx.doi.org/10.1109/CVPR.2019.00963.

Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., Cho, K., 2019. Augmentation for small object detection. arXiv preprint arXiv:1902.07296.

Li, Z., Chen, G., Zhang, T., 2020. A CNN-transformer hybrid approach for crop classification using multitemporal multisensor images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 847–858. http://dx.doi.org/10.1109/JSTARS.2020.2971763.

Li, J., Meng, L., Yang, B., Tao, C., Li, L., Zhang, W., 2021. LabelRS: An automated toolbox to make deep learning samples from remote sensing images. Remote Sens. 13 (11), 2064. http://dx.doi.org/10.3390/rs13112064.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. In: Fleet, D., Tomas, P., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. In: Lecture Notes in Computer Science, vol. 8693, Springer Cham, Zurich, Switzerland, pp. 740–755. http://dx.doi.org/10.1007/978-3-319-10602-1_48.

Mohan, R., Valada, A., 2021. EfficientPS: Efficient panoptic segmentation. Int. J. Comput. Vis. 129 (5), 1551–1579. http://dx.doi.org/10.1007/s11263-021-01445-z.

Mou, L., Zhu, X.X., 2018. Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network. IEEE Trans. Geosci. Remote Sens. 56 (11), 6699–6711. http://dx.doi.org/10.1109/TGRS.2018.2841808.

Neuhold, G., Ollmann, T., Bulo, S.R., Kontschieder, P., 2017. The mapillary vistas dataset for semantic understanding of street scenes. In: 2017 IEEE International Conference on Computer Vision (ICCV). vol. 2017-Octob, IEEE, pp. 5000–5009. http://dx.doi.org/10.1109/ICCV.2017.534,

Qiao, S., Zhu, Y., Adam, H., Yuille, A., Chen, L.-C., 2021. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3997–4008.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (Eds.), Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 9351, Springer, Cham, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Tan, M., Le, V.Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, California, USA, pp. 6105–6114.

Tong, K., Wu, Y., Zhou, F., 2020. Recent advances in small object detection based on deep learning: A review. Image Vis. Comput. 97, 103910. http://dx.doi.org/10.1016/j.imavis.2020.103910.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R., 2019. Detectron2. [Online]. URL: https://github.com/facebookresearch/detectron2. (Accessed 3 March 2021).

Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, Honolulu, HI, USA, pp. 5987–5995. http://dx.doi.org/10.1109/CVPR.2017.634.

Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., Urtasun, R., 2019. Upsnet: A unified panoptic segmentation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8818–8826.

Yu, X., Lou, B., Zhang, D., Winkel, D., Arrahmane, N., Diallo, M., Meng, T., Busch, H.v., Grimm, R., Kiefer, B., et al., 2020. Deep attentive panoptic model for prostate cancer detection using biparametric MRI scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 594–604.

Zhang, D., Song, Y., Liu, D., Jia, H., Liu, S., Xia, Y., Huang, H., Cai, W., 2018. Panoptic segmentation with an end-to-end cell R-CNN for pathology image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 237–244.

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 3–11.