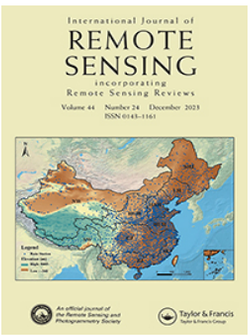

International Journal of Remote Sensing

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tres20

A multi-objective comparison of CNN architectures in Arctic human-built infrastructure mapping from sub-meter resolution satellite imagery

Elias Manos, Chandi Witharana, Amal S. Perera & Anna K. Liljedahl

To cite this article: Elias Manos, Chandi Witharana, Amal S. Perera & Anna K. Liljedahl (2023) A multi-objective comparison of CNN architectures in Arctic human-built infrastructure mapping from sub-meter resolution satellite imagery, International Journal of Remote Sensing, 44:24, 7670-7705, DOI: 10.1080/01431161.2023.2287563

To link to this article: https://doi.org/10.1080/01431161.2023.2287563

Published online: 11 Dec 2023.

INTERNATIONAL JOURNAL OF REMOTE SENSING
2023, VOL. 44, NO. 24, 7670–7705
https://doi.org/10.1080/01431161.2023.2287563

Taylor & Francis
Taylor & Francis Group

A multi-objective comparison of CNN architectures in Arctic human-built infrastructure mapping from sub-meter resolution satellite imagery

Elias Manos[a], Chandi Witharana[a,b], Amal S. Perera[a] and Anna K. Liljedahl[c]

[a]Department of Natural Resources and the Environment, University of Connecticut, Storrs, CT, USA; [b]Eversource Energy Center, University of Connecticut, Storrs, CT, USA; [c]Woodwell Climate Research Center, Falmouth, MA, USA

ABSTRACT

Risk assessment of infrastructure exposed to ice-rich permafrost hazards is essential for climate change adaptation in the Arctic. As this process requires up-to-date, comprehensive, high-resolution maps of human-built infrastructure, gaps in such geospatial information and knowledge of the applications required to produce it must be addressed. Therefore, this study highlights the ongoing development of a deep learning approach to efficiently map the Arctic built environment by detecting nine different types of structures (detached houses, row houses, multi-story blocks, non-residential buildings, roads, runways, gravel pads, pipelines, and storage tanks) from recently-acquired Maxar commercial satellite imagery (<1 m resolution). We conducted a multi-objective comparison, focusing on generalization performance and computational cost, of nine different semantic segmentation architectures. K-fold cross validation was used to estimate the average F1-score of each architecture and the Friedman Aligned Ranks test with the Bergmann-Hommel post-hoc procedure was applied to test for significant differences in generalization performance. ResNet-50-UNet++ performs significantly better than five out of the other eight candidate architectures; no significant difference was found in the pairwise comparisons of ResNet-50-UNet++ to ResNet-50-MANet, ResNet-101-MANet, and ResNet-101-UNet++. We then conducted a high-performance computing scaling experiment to compare the number of service units and runtime required for model inferencing on a hypothetical pan-Arctic scale dataset. We found that the ResNet-50-UNet++ model could save up to ~ 54% on service unit expenditure, or ~ 18% on runtime, when considering operational deployment of our mapping approach. Our results suggest that ResNet-50-UNet++ could be the most suitable architecture (out of the nine that were examined) for deep learning-enabled Arctic infrastructure mapping efforts. Overall, our findings regarding the differences between the examined CNN architectures and our methodological framework for multi-objective architecture comparison can provide a foundation that may propel future pan-Arctic GeoAI mapping efforts of infrastructure.

ARTICLE HISTORY
Received 21 August 2023
Accepted 12 November 2023

KEYWORDS
Segmentation; neural networks; buildings; Arctic; deep learning; very high spatial resolution

CONTACT Elias Manos elias.manos@uconn.edu Department of Natural Resources and the Environment, University of Connecticut, 1376 Storrs Road, Unit 4087, Storrs, CT 06269, USA

1. Introduction

Climate change has led to widespread warming across the Arctic (Biskaborn et al. 2019), where land surface temperatures are reported to have increased by more than 0.5°C per decade since 1981 (Comiso and Hall 2014), exceeding average global warming by a factor of between 2 and 3. Permafrost, defined as soil or bedrock at or below 0°C for at least two consecutive years (Dobinski 2011), underlies approximately 24% of the exposed land surface of the Northern Hemisphere ('Circum-Arctic Map of Permafrost and Ground-Ice Conditions, Version 2' 2022). As such, this warming is expected to promote thawing of near-surface permafrost and subsequent thickening of the active layer, which decreases the bearing capacity of the soil and results in ground subsidence in areas with ice-rich permafrost (Blunden and Arndt 2017; Nelson, Anisimov, and Shiklomanov 2001; Streletskiy et al. 2015, 2017). Both of these effects are major hazards to infrastructure (e.g. buildings, roads, airports, pipelines, industrial facilities) built on permafrost, as it relies on the mechanical strength and stability of the underlying frozen soils (Instanes and Anisimov 2016; Khrustalev, Parmuzin, and Emelyanova 2011).

Further contributing to this exposure are the estimated 1,162 settlements on permafrost in the Arctic, accommodating approximately 5 million inhabitants (Ramage et al. 2021). While this regional population composes only 0.07% of the world population, the Arctic plays a disproportionally large role in the global economy, contributing 0.6% to the global gross domestic product (Nymand Larsen 2014; J. N. Larsen and Huskey 2015), which is expected to grow larger due to increasing economic relevance in areas such as natural resource extraction (Gautier et al. 2009; Hossain 2017). Maintaining operational infrastructure is thus critical for the sustainable development of these communities and economies.

Recent reviews and benchmark reports have called for pan-Arctic permafrost hazard mapping and infrastructure risk assessments to quantify the socioeconomic impacts of permafrost degradation, which will inform effective adaptation and mitigation measures and future construction planning (Programme (AMAP) (2017; Hassol 2004; Hjort et al. 2022). Several such risk assessment studies have been conducted to quantify potential costs of infrastructure damage in the Russian Arctic (Badina 2020; Melnikov et al. 2022; Streletskiy et al. 2019), Alaska (Melvin et al. 2017; P. H.; Larsen et al. 2008), and Canada (Dore, Burton, and Dore 2001), while others have strictly focused on particular kinds of infrastructure, such as roads (Porfiriev, Eliseev, and Streletskiy 2019), housing (Porfiriev, Eliseev, and Streletskiy 2021), and healthcare facilities (Porfiriev, Eliseev, and Streletskiy 2021). Two studies have been conducted at the circumpolar scale to estimate the total cost of Arctic infrastructure damages due to permafrost degradation under different climate change scenarios, with estimates ranging from 20 billion USD to 276 billion USD (Streletskiy et al. 2023; Suter, Streletskiy, and Shiklomanov 2019). This disparity can be attributed to an obstacle that has generally challenged all the aforementioned risk assessment studies, that is, the lack of a comprehensive analysis-ready geospatial infrastructure inventory (Hjort et al. 2022). This exposure information is necessary to quantify the damage to assets (i.e. infrastructure) that are co-located with hazards, and can problematically lead to underestimated costs when limited (Suter, Streletskiy, and Shiklomanov 2019).

Satellite-based mapping can be used to improve the geospatial data record of pan-Arctic built infrastructure. While we cannot guarantee it is completely exhaustive, we conducted a literature survey on Arctic built infrastructure mapping efforts and found that the task is ill-addressed, with few studies existing and only one study addressing pan-Arctic mapping from recent satellite imagery. The survey results are summarized in Table 1. Based on this survey, it was found that most studies mapped built infrastructure across small geographic extents through manual digitization, with a focus on studying anthropogenic change in the Bovanenkovo gas field in the Yamal Peninsula, Russia. Bartsch et al (Bartsch et al. 2020, 2021). published the first and only pan-Arctic satellite-based record of infrastructure within 100 km of Arctic coasts, named the Sentinel-1/2 derived Arctic Coastal Human Impact (SACHI) dataset. Gradient Boosting Machine and U-Net, machine learning and deep learning algorithms, respectively, were used to automatically classify pixels in Sentinel-1 (synthetic aperture radar imagery) and Sentinel-2 (multispectral imagery) images as linear transport infrastructure (roads and railways), buildings, or other impacted area. In further confirmation of existing data gaps, the authors found that 40% of human-impacted area identified in the SACHI dataset was not yet included in OpenStreetMap (Ramm 2020). However, at 10 m spatial resolution, Sentinel imagery may not be able to fully address the infrastructure data gap. In general, very high spatial resolution (VHSR) (<5 m resolution) is crucial in providing the required level of detail for accurate detection and classification of individual built structures. This is particularly true in the Arctic as buildings are typically small, many roads are thin and unpaved, and characteristic features, such as pipelines, are difficult to resolve in 10 m resolution imagery (Bartsch et al. 2021; Kumpula et al. 2012; Kumpula, Forbes, and Stammler 2006). Preserving the semantic (e.g. building type) and geometric (e.g. area, length, shape) properties of individual structures mapped from imagery is imperative in enabling effective risk assessments and subsequent decision-making. Therefore, developing and testing an approach to map infrastructure from VHSR imagery is needed.

Further complicating this research, conspicuous shortfalls of traditional remote sensing image analysis when confronted with VHSR imagery (Blaschke 2010) have catalysed a migration towards computer vision-based algorithms, namely the convolutional neural network (CNN). High spatial resolution imagery presents scene objects much larger than the associated pixel size, introducing complex properties, such as geometry, context, pattern, and texture that compose objects at multiple levels (Blaschke 2010). Higher spatial resolution also significantly increases intra-class spectral variability, given the increased number of pixels constructing image features (Thomas Blaschke et al. 2014). As such, traditional image analysis methods, namely per-pixel-based approaches, are ill-equipped to handle VHSR imagery, whereas CNNs are better equipped.

Additionally, with the entire Arctic being imaged by Maxar commercial satellite sensors at a sub-metre resolution (Witharana et al. 2023), U.S. National Science Foundation Polar Program-funded researchers have access to free 'big' imagery data via the Polar Geospatial Center at the University of Minnesota. This has created unprecedented opportunities and challenges in producing circumpolar sub-metre resolution maps of the natural and built Arctic environments. Notably, our ongoing work has resulted in the novel Mapping Application for Arctic Permafrost Land Environment (MAPLE), an operational-scale GeoAI pipeline that harnesses AI and high-performance computing (HPC) resources for automated segmentation of tens of thousands of Maxar satellite images

Table 1. Literature survey on satellite-based Arctic built infrastructure mapping efforts.

Reference	Study Area	Data	Spatial Resolution (in Order of Listed Data, "Field Survey" Omitted)	Method	Feature(s) of Interest
Kumpula, Forbes, and Stammler (2006)	Bovanenkovo gas field, Yamal Peninsula (West Siberia)	Field survey, QuickBird-2 (panchromatic, multispectral), ASTER VNIR, Landsat (TM, MSS)	0.61 m, 2.5 m, 15 m, 30 m, 80 m	Manual digitization	Quarries, power lines, roads, winter roads, drill towers, barracks
T. Kumpula, Forbes, and Stammler (2010)	Bovanenkovo gas field, Yamal Peninsula (West Siberia)	Field survey, QuickBird-2 (pan, multi), ASTER VNIR, SPOT (pan, multi), Landsat (ETM7, TM, MSS)	0.63 m, 2.4 m, 15 m, 10 m, 20 m, 30 m, 30 m, 80 m	Manual digitization	Roads, impervious cover, barracks, winter roads, settlements, quarries
Timo Kumpula et al. (2011)	Bovanenkovo gas field and Toravei oil field, Yamal Peninsula (West Siberia)	Field survey, QuickBird-2 (pan, multi), ASTER VNIR, SPOT (multi), Landsat (ETM7, TM, MSS)	0.63 m, 2.4 m, 15 m, 10 m, 20 m, 30 m, 30 m	Manual digitization	Buildings, roads, sand quarries, pipelines
Timo Kumpula et al. (2012)	Bovanenkovo gas field, Yamal Peninsula (West Siberia)	Field survey, QuickBird-2 (pan, multi), GeoEye, ASTER VNIR, SPOT (multi), Landsat (ETM7, TM, MSS)	0.63 m, 2.4 m, 1.65 m, 15 m, 20 m, 30 m, 30 m, 70 m	Manual digitization	Pipelines, powerlines, drilling towers, roads, impervious cover, barracks, settlements, quarries
Raynolds et al. (2014)	Prudhoe Bay Oilfield, Alaska	Aerial photography (B&W, colour, colour infrared)	1 ft resolution for two images. Map scale was then used to describe the rest of the imagery. Scales are as follows: 1:3000, 1:6000, 1:12,000, 1:18,000, 1:24,000, 1:60,000, 1:68,000, 1:120,000	Manual digitization	Roads, gravel pads, excavations, pipelines, powerlines, fences, canals, gravel and construction debris
Gadal and Ouerghemmi (2019)	Yakutsk, Russia	SPOT-6 (pan, multi), Sentinel-2 (multi)	1.5 m, 6 m, 10 m	Semi-automated (object-based image analysis)	Houses, other structures
Ourng, Vaguet, and Derkacheva (2019)	Surgut, Russia	Sentinel-1 (SAR), Sentinel-2 (multi), Landsat (TM, MSS)	10 m, 10 m, 30 m, 60 m	Automated (machine learning)	Built-up area
Bartsch et al. (2020)	Pan-Arctic, within 100 km of the Arctic coast	Sentinel-1 (SAR) and Sentinel-2 (multi)	10 m, 10 m	Automated (machine learning and deep learning)	Buildings, roads, other human-impacted areas
Ardelean et al. (2020)	Bovanenkovo gas field, Yamal Peninsula (West Siberia)	QuickBird-2 (pan, multi), GeoEye-1 (pan, multi)	0.6 m, 2.4 m, 0.4 m, 1.8 m	Manual digitization	Buildings, roads
Manos et al. (2022)	North Slope of Alaska	WorldView-2, QuickBird-2	0.5 m – 0.87 m	Automated (deep learning)	Buildings (public, commercial, residential, industrial, roads)

(Udawalpola et al. 2021, 2022; Witharana, Abul Ehsan Bhuiyan, and Liljedahl 2020). MAPLE has been successfully deployed to produce the first pan-Arctic ice-wedge polygon map, with over 1 billion individual ice-wedge polygons detected and classified (Witharana et al. 2023). Additionally, Witharana et al. (2022) (Witharana et al. 2022) have recently explored extending MAPLE's capabilities into mapping another prominent permafrost landform known as the retrogressive thaw slump.

In expanding the capabilities of MAPLE into mapping the built environment, Manos et al. (2022) (Manos et al. 2022) conducted a pilot study to test the performance of a deep learning-based semantic segmentation workflow in mapping Arctic infrastructure from Maxar satellite imagery (<1 m resolution). As demonstrated in Table 1, this is the first and only study conducted on this particular application. The model, composed of a pre-trained ResNet-50 encoder and UNet++ decoder, was trained to detect buildings, classified based on their functional use, and roads, from WorldView-02 and QuickBird-02 scenes of a city and an industrial site on the Alaskan North Slope. The trained model attained a promising F1-score of 0.83 on the testing dataset. In this study, we build upon this previous work by expanding the geographic domain (i.e. more study sites) and thematic depth (i.e. more target classes) of the model training dataset. This is done in order to move closer to the pan-Arctic scale and better investigate how deep learning models would perform in infrastructure mapping at such a level.

However, a plethora of CNN architectures have been introduced and continue to emerge during this current 'golden age' of deep learning research. Selecting the proper architecture for a given application task is not necessarily a straightforward choice and is often times approached arbitrarily. Various textbooks (Japkowicz and Shah 2011), review articles (Raschka 2020; Santafé, Inza, and Lozano 2015) and research articles (Guerrero Vázquez et al. 2001; Pizarro, Guerrero, and Galindo 2002) in the machine learning literature have outlined a systematic process of statistically comparing the performance of multiple (more than two) learning algorithms on a given task. This process consists of the following recommended steps:

(1) Choose the learning algorithms to be evaluated.
(2) Select a performance measure of interest (e.g. F1-score for semantic segmentation).
(3) Select a resampling method (e.g. k-fold cross-validation, bootstrapping, randomization) with which the performance measure of interest will be reliably estimated.
(4) Test for significant differences in algorithmic performance with non-parametric statistical analysis, consisting of an omnibus test and subsequent pairwise post-hoc tests with adjustments for multiple comparisons (if the null hypothesis of the omnibus test was rejected).

We conducted a brief survey of the remote sensing literature and found that a number of recent studies (Duro, Franklin, and Dubé 2012; Li et al. 2016; López-Serrano et al. 2016; Nhu et al. 2020; Peña et al. 2014) have applied this general framework in conducting statistically rigorous comparisons of machine learning algorithm performance in image analysis. For example, Peña et al (Peña et al. 2014). compared decision tree, logistic regression, support vector machine, and multilayer perceptron in object-based image classification for summer crop

mapping from ASTER satellite imagery. The performances of these algorithms were estimated using 10-fold cross-validation and statistically compared with the Friedman test.

Therefore, by adopting this general framework that has been laid out in the literature, this study investigates one central question: given a set of candidate 'state-of-the-art' semantic segmentation architectures, is there any one architecture that significantly outperforms all others in the task of mapping Arctic human-built infrastructure from Maxar satellite imagery? Considering a set of nine different semantic segmentation architectures that combine three different encoders (ResNet-34, ResNet-50, ResNet-101) and decoders (UNet++, DeepLabV3+, MANet), we will answer this question through a multi-objective comparison procedure that focuses on the generalization performance and computational cost of each architecture. Firstly, we conduct model training and k-fold cross-validation-based performance estimation with each architecture. Secondly, based on these estimated performances, we conduct a non-parametric statistical analysis, using the Friedman aligned ranks test with the Bergmann-Hommel post-hoc procedure, to determine if any one architecture significantly outperforms the rest in terms of F1-score. Thirdly, we place this comparison in the context of an HPC environment by comparing expended computational resources of each architecture estimated through a scaling experiment. Finally, using the optimal architecture as determined through our comparison, this study also generally assesses the performance of deep learning-based semantic segmentation in detecting various kinds of structures across a diverse range of Arctic built environments (i.e. rural, medium-density, urban settlements, and industrial sites).

2. Materials and methods

2.1. Study sites

In constructing a dataset that accounts for regional variability in the natural and built environment across the Arctic, we selected nine different sites across Arctic Alaska, Canada, and Russia (Figure 1). Each of these sites was chosen primarily to represent a built environment setting (either a rural settlement, medium-density settlement, urban settlement, or industrial site), and secondarily to represent a particular climate setting (either tundra or boreal climate). These sites and the particular setting that they represent are listed in Table 2.

This diversity is imperative in ensuring that a CNN model is introduced to the full range of the major infrastructure types that exist in different settings (e.g. pipelines and gravel pads in the Prudhoe Bay Oil Field vs. multi-story apartment buildings in Norilsk, Russia). Moreover, each of these settings present unique challenges for detection. For example, detecting infrastructure in boreal climates might be more difficult than in tundra climates due to tree occlusion that blocks buildings or roads. Detection in urban settlements may be more difficult than in rural settlements for a multitude of reasons. For example, high building density can result in overlap between buildings. This makes distinguishing individual buildings difficult, which is further hampered by the irregular shapes that they often take on (Aytekın et al. 2012). Furthermore, complex urban scenes are characterized by heterogeneous development, consisting of a mixture of surface materials (e.g. concrete, brick, asphalt, metal, plastic, glass, shingles, and soil) that introduces spectral variability, which is problematic for detection (Aytekın et al. 2012).

Figure 1. Overview map of study sites.

Table 2. Attributes of each study site and satellite image(s) used for each site (WV02 – WorldView-02, WV03 – worldview-03, QB02 – QuickBird-02).

Study Area	Setting	Population	Sensor	Acquisition date	Spatial resolution (m)
Nuiqsut, AK	Rural-Tundra	512	WV02	12 August 2012	0.56 x 0.53
Point Hope, AK	Rural-Tundra	830	QB02	19 August 2006	0.62 x 0.60
Prudhoe Bay Oil Field, AK	Industrial-Tundra	N/A	QB02	21 August 2009	0.63 x 0.61
Utqiagvik, AK	Medium-Density-Tundra	4,927	QB02	1 August 2002	0.67 x 0.71
			WV02	1 September 2014	0.72 x 0.87
Wainwright, AK	Rural-Tundra	628	WV02	9 August 2011	0.49 x 0.49
Coral Harbour, CAN	Rural-Tundra	890	WV03	25 July 2020	0.34 x 0.37
Yellowknife, CAN	Urban-Boreal	20,340	WV03	17 August 2022	0.40 x 0.36
Norilsk, RU	Urban-Tundra	179,554	WV03	19 July 2019	0.33 x 0.32
Saskylakh, RU	Rural-Boreal	2,317	WV02	9 September 2014	0.50 x 0.50

2.2. Data

2.2.1. Maxar satellite imagery

We selected ten Maxar satellite images (WorldView-02 and –03, and QuickBird-02), one for each of the nine sites (with the exception of Utqiagvik, for which we selected two images to cover its full extent). We selected the most recent image, acquired during the summer, with minimal cloud cover. Furthermore, for model training, we only utilized the blue, green, and red bands of the imagery. This is because our model consists of a ResNet-50 encoder pre-trained on a three-channel input, therefore any new input imagery must only contain three channels. All of the pre-processed (pansharpened, orthorectified) images were provided by the Polar Geospatial Center at the University of Minnesota. The attributes of these images are given in Table 2.

2.2.2. Infrastructure classification scheme

Our infrastructure digitization process is based on the urban structure type (UST) classification scheme, as proposed by Lehner and Blaschke (2019) (Lehner and Blaschke 2019). As described by the authors, the UST scheme offers a generic structural- and object-based typology that exclusively focuses on the morphology of structures (i.e. general exterior appearance of buildings). It makes use of image object-related features, such as texture, patterns, shape, and leaves out all social or land use aspects of a structure. This concept is especially important when segmenting high-resolution remote sensing imagery of developed landscapes, as these common high-level semantic interpretations related to social/land use aspects do not always correspond to object types that can be drawn from low-level representations in satellite imagery (a phenomenon known as the *semantic gap*) (Li et al. 2022).

Therefore, we developed a nine-class infrastructure classification scheme, similar to the proposed UST scheme, that accounts for all the major built structures that compose most settlements or industrial sites across the Arctic. Our scheme follows a hierarchy (Figure 2) that generally considers all 'human-built structures' as either being 'buildings' or 'non-buildings'. It then considers 'buildings' as either 'residential'

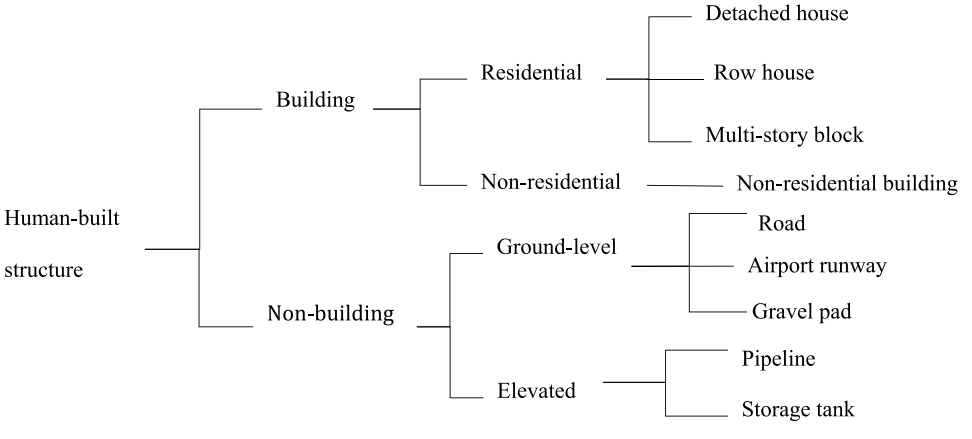

Figure 2. Arctic human-built infrastructure classification scheme hierarchy.

Table 3. Natural language description of the four 'building' types.

Class	Natural language description
Detached house	Small, rectangular objects covering a small area with pitched roofs and low height. Mostly occur in neighbourhoods adjacent to roads.
Row house	Spatially continuous collection of several dwellings attached together at sides. Repeating roof pattern, like chimneys or eaves (edges of the roof which overhang the face of a wall), signifies the presence of multiple connected houses.
Multi-story block	Elongated, rectangular (block) objects with low sloping/flat roofs. Much larger than detached houses and row houses. Height can typically be inferred from accompanying shadows.
Non-residential building	Typically cover largest area and can also display irregular geometries that are composed of multiple connected segments (e.g. L-shaped). Roofs are typically flat, which is indicative of non-residential development.

or 'non-residential'. We formulated natural language descriptions (Table 3) to define the semantics of the four resulting building types, which ultimately link human language to feature representations that consistently define each type in the digital image domain. As for the 'non-building' structures, these are further sub-categorized into 'ground-level' and 'elevated' structures. The former considers modifications of the earth's surface (i.e. impervious cover types). The latter considers those structures that can be distinguished from the former in that they are not surface-level modifications, but still cannot be considered 'buildings' that humans can populate. We did not formulate natural language descriptions for these classes since their human-assigned labels typically correspond to one distinct and intuitive visual appearance in imagery that most can recognize.

2.2.3. Infrastructure digitization

To produce data for CNN model training and testing, infrastructure features in the satellite imagery are digitized and labelled based on the aforementioned classification scheme (Figure 3). However, this is a time-consuming and laborious process. Therefore, we integrated multiple publicly-available geospatial data sources that offered high-quality geospatial data layers for infrastructure, which accounted for most of the major structures at each of our study sites. Data layers for buildings (i.e. detached house, row house, multi-story block, non-residential building), runways, gravel pads, and storage tanks consisted of polygons representing the footprint (i.e. physical border of a structure) of a given feature. Data layers for roads and pipelines consisted of lines that represent the centerline of a given feature. These sources and the digitized features that they offered for each site are listed in Table 4.

By making use of this existing data, we were able to mitigate the bottleneck imposed by performing on-screen digitization from scratch, which cannot be avoided in many cases of training data production. However, we still needed to ensure the quality of the features in these data layers and manipulate their attributes so that they adhered to our classification scheme. In addition, we still needed to account for the few features that were not represented in these data layers. This procedure, performed in the *ArcGIS Pro 3.0.3* software, consisted of the following points:

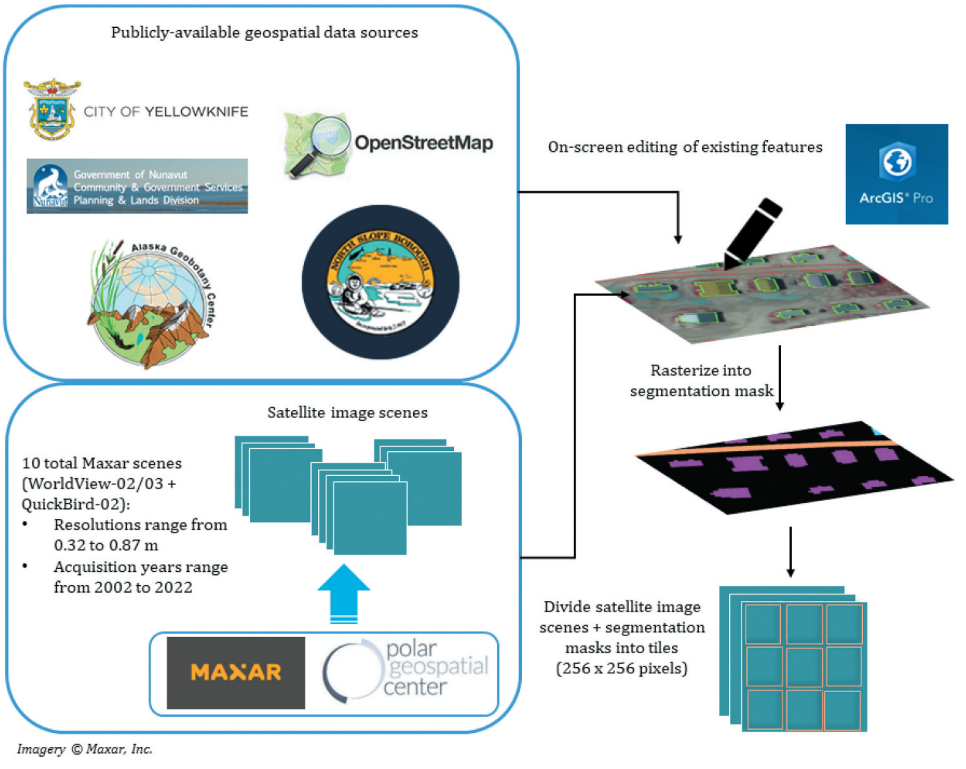

Figure 3. Graphical overview of the infrastructure digitization procedure.

Table 4. Publicly-available geospatial data sources used for infrastructure digitization.

Source	Study Site(s) Covered	Infrastructure Features Covered
North Slope Borough, Planning & Community Services, GIS ("NSB GIS Public" 2022)	Nuiqsut, Point Hope, Utqiagvik, Wainwright	Buildings, roads, storage tanks
Alaska Geobotany Center ("Alaska Geobotany Center - Organizations - Alaska Arctic Geoecological Atlas" 2022)	Prudhoe Bay Oil Field	Gravel pads
City of Yellowknife ("OpenDataTemplate" 2023)	Yellowknife	Buildings, roads
Government of Nunavut, Community & Government Services, Planning & Lands Division ("CGS Planning & Lands - GIS Data (ESRI Shapefile)" (2023)	Coral Harbour	Buildings, airport runways
OpenStreetMap(https://download.geofabrik.de/index.html)	All sites	Buildings (Russia and Prudhoe Bay), roads, airport runways

- Ensure that a polygon feature lines up with the intended corresponding building, road, gravel pad or runway as it appears in the image. If the alignment is poor, the whole feature must be shifted, or its geometry (e.g. vertices) must be modified so that the digital boundary and real-world boundary agree.
- Convert line features for roads and pipelines to polygon features by applying a buffer. If this new polygon does not match the road or pipeline well, either edit its geometry or create a new polygon feature from scratch.

Table 5. Summary statistics of the infrastructure polygon layers for each study site (entire dataset).

Study site	Detached house	Row house	Multi-story block	Non-residential building	Road	Airport runway	Gravel pad	Pipeline	Storage tank	Total features
Nuiqsut, AK	148	1	0	28	42	1	0	0	9	229
Point Hope, AK	269	0	0	40	81	1	0	7	8	406
Prudhoe Bay Oil Field, AK	0	0	0	161	117	0	255	152	29	714
Utqiagvik, AK	1371	7	6	115	240	2	0	0	17	1758
Wainwright, AK	203	0	0	36	63	1	0	0	10	313
Coral Harbour, CA	121	9	0	55	28	1	0	0	4	218
Yellowknife, CA	272	152	54	123	99	6	0	0	24	730
Norilsk, RU	0	0	525	115	82	0	0	0	0	722
Saskylakh, RU	233	0	0	9	27	1	2	0	18	290
Total features	2617	169	585	682	779	13	257	159	119	5380

- Create a new polygon from scratch for those real-world structures that are not digitally represented in the data layers.
- Label each feature with a class value based on the infrastructure classification scheme.

Finally, once this quality assurance procedure was completed, we merged all of the data layers for the different infrastructure features into one comprehensive polygon layer for each site (summarized in Table 5). These are rasterized to produce segmentation masks. Along with the corresponding satellite image, these are split into smaller tiles, sized at 256×256 pixels, to be used in CNN model training. In total, our dataset consists of 2,822 of these image and mask tiles.

2.3. Model training and evaluation

2.3.1. Development environment

Model training and evaluation was supported by HPC resources. We trained and evaluated models within a Conda environment set up on the Delta supercomputer ('Delta User Guide – Delta Supercomputer – NCSA Wiki' 2023), which is maintained by the National Center for Supercomputing Applications at the University of Illinois Urbana-Champaign (and supported by the National Science Foundation). To train each model, we strictly utilized one GPU on the Delta 4-way NVIDIA A100 GPU (40 GB memory) compute node, as opposed to distributing the training across all four GPUs. With *Python 3.7.16*, we built and developed each model using *PyTorch 1.10.1+cu113* (Paszke et al. 2019) and *Segmentation Models PyTorch 0.2.1* (Iakubovskii 2023). We set up k-fold cross validation using *scikit-learn 1.0.2*.

2.3.2. Semantic segmentation architectures

A plethora of CNN architectures designed for semantic segmentation, typically composed of an encoder and decoder, are currently circulating throughout deep learning research as 'state-of-the-art' techniques. By combining three of these encoders and decoders, we formulated nine different candidate architectures to

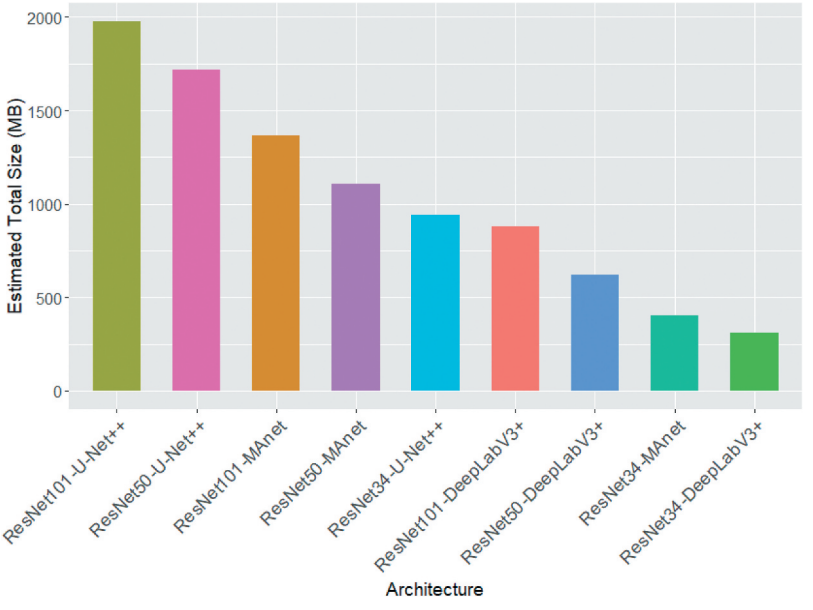

Figure 4. Estimated total size (MB) of each architecture, calculated as the sum of the parameter size and forward/backward pass size of the network.

assess in the task of Arctic infrastructure mapping. For encoders, we selected three ResNet (He et al. 2015) architectures of varying depth: ResNet-34, ResNet-50, ResNet-101. All encoders were pretrained on the ImageNet dataset (Deng et al. 2009). For decoders, we selected three decoders, each with their own unique architectural features and modules: DeepLabV3+ (encoder-decoder with dilated convolutions and multi-scale feature learning) (Chen et al. 2018), MANet (multi-scale attention network with spatial and channel attention modules) (Fan et al. 2020), and UNet++ (encoder-decoder with skip connections, extension of original UNet with a more complex decoder) (Zhou et al. 2018). To visualize the differences between these architectures, we plot their estimated total sizes in megabytes in Figure 4. The estimated total size, calculated as the sum of the parameter size and forward/backward pass size of the network, measures the amount of expected memory required to run a given model and gives insight into the computational complexity of a given algorithm.

2.3.3. Hyperparameter configuration

Since our variable of interest in this experiment was strictly CNN architecture, we deployed the architectures in a default format without manipulating any internal features (e.g. encoder/decoder depth) and held all hyperparameters constant:

- All models were trained for 80 epochs.
- Input dimensions were $256 \times 256 \times 3$, with a batch size of 16.
- We utilized the Focal loss function (Lin et al. 2018) in multilabel mode with gamma = 2.0. In the case of dense object detection or segmentation, as seen in this study with

target objects such as small houses and thin roads in rural Arctic communities, model training is challenged by an extreme foreground-background class imbalance. Focal loss was designed specifically to handle this issue by dampening the effect of easy negative examples dominated by background pixels.

- We utilized the Adam optimizer, with $\beta 1 = 0.9$, $\beta 2 = 0.999$ and $\varepsilon = 1e\text{-}08$ as suggested by Kingma and Ba (Kingma and Ba 2017). We set the initial learning rate to 1e-4, which was automatically reduced to 1e-5 after 35 epochs.

2.3.4. K-Fold cross-validation

As demonstrated by Japkowicz and Shah (2011) (Japkowicz and Shah 2011), the convergence of the empirical performance of a learning algorithm to the true performance depends on the size of the training and testing dataset. Naturally, the more data at hand, the closer a performance estimate will be to the true performance of a given algorithm. However, given that machine learning tasks often rely on insufficient amounts of data, resampling methods that make use of all the available data for both training and evaluation are recommended in order to generate accurate and reliable performance estimates (Japkowicz and Shah 2011; Molinaro, Simon, and Pfeiffer 2005; Raschka 2020; Santafé, Inza, and Lozano 2015). K-fold cross-validation is one of the most widely used resampling approaches in machine learning experiments.

As graphically described in Figure 5, k-fold cross-validation involves splitting the dataset into *k* equal-size and mutually exclusive folds. Different values for *k* can be chosen, but the most popular choice is ten (Kohavi 1995). $k-1$ of these folds are used to train the model and the remaining fold is used to estimate the performance measure of interest. This process is repeated *k* times until all the folds are used for testing. This repeated process obtains *k* classification models and therefore *k* estimations. Finally, the estimated k-fold cross-validation value of the performance measure is obtained by averaging over all the obtained values.

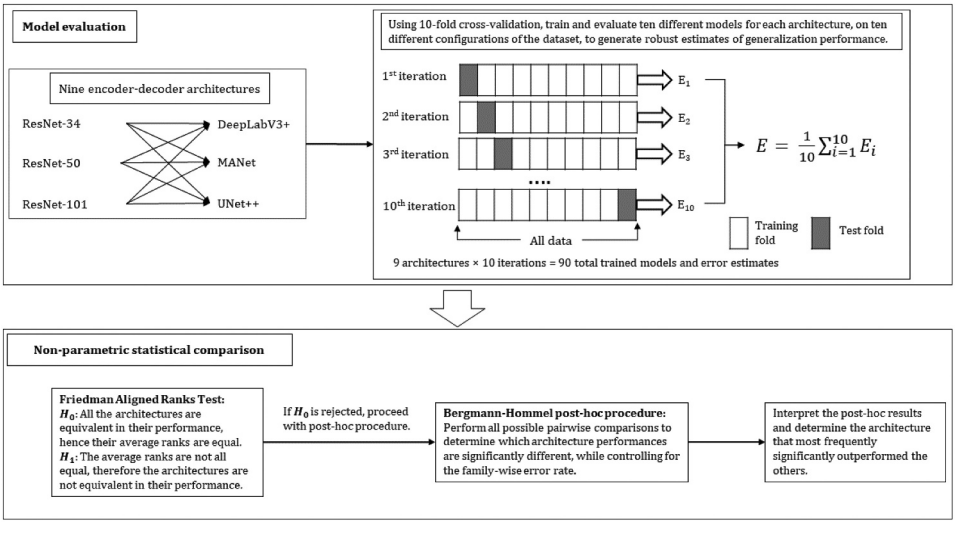

Figure 5. Graphical overview of our experimental framework for statistical comparison of architectures.

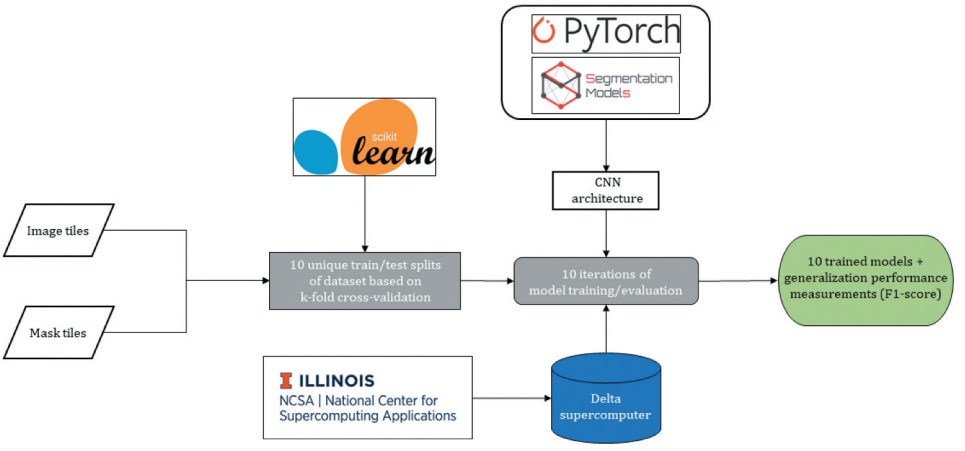

Figure 6. Graphical overview of the model training and evaluation process for one architecture.

Therefore, as displayed in Figure 6, we obtained a total of 90 trained models and 90 associated performance measurements and training runtimes (ten for each architecture). In particular, we measured generalization performance with F1-score, commonly used in the evaluation of semantic segmentation models, which is calculated as the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{1}$$

Precision is the fraction of relevant instances among the retrieved instances:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2}$$

Recall is the fraction of relevant instances that were retrieved:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{3}$$

The F1-scores achieved by each architecture over the ten folds are then analysed with non-parametric statistical significance testing to understand if the observed results can be attributed to real characteristics of the evaluated algorithms or if they were obtained by chance.

2.4. Non-parametric statistical analysis

Various studies have set guidelines for the use of statistical tests in determining whether learning algorithms exhibit significant differences in their performances, with distinctions made between different scenarios (e.g. pairwise comparisons, multiple comparisons with a control algorithm, multiple comparisons among all algorithms) (Japkowicz and Shah 2011; Santafé, Inza, and Lozano 2015; Demšar 2006; García and Herrera 2008; García et al. 2010). When comparing a set of

multiple algorithms, as is the case in this experiment, repeated-measures ANOVA is commonly used. However, it has generally been recommended to avoid the use of parametric tests (e.g. ANOVA) when analysing learning algorithm performance, since the assumptions of independency, normality, and homoscedasticity are unlikely to hold (Zar 1999). As Demsar (2006) (Demšar 2006) points out, while repeated-measures ANOVA is robust to the violations of the normality assumption given enough samples, verifying the homogeneity of variances is often difficult in most cases. Furthermore, independency is not truly verified when using resampling methods such as 10-fold cross validation (since a portion of the dataset could be used either for training and testing in different partitions), as we have used in this experiment (Demšar 2006).

As a result, a number of rank-based non-parametric tests that are not held to these assumptions have been proposed as alternatives. We have implemented the Friedman aligned ranks test (García et al. 2010; Hodges and Lehmann 1962), an extension of the Friedman test (Friedman 1937, 1940), which is a well-known non-parametric omnibus test that typically serves as an alternative to repeated-measures ANOVA. The null hypothesis for the Friedman aligned ranks test states equality of medians between the populations; the alternative hypothesis is defined as the negation of the null hypothesis.

To compute the test statistic, the performances achieved by all algorithms across all datasets are ranked relative to each other. These ranks are calculated in three steps. First, a value of location is computed as the average performance achieved by all algorithms in each data set. Then, the difference between the performance obtained by an algorithm and the value of location is calculated. This step is repeated for k algorithms and $n$datasets. Finally, these resulting differences, which are ultimately referred to as 'aligned observations', are ranked from 1 to kn relative to each other. The ranks assigned to the aligned observations are thus called 'aligned ranks'. The Friedman aligned ranks test statistic is then calculated as:

$$T = \frac{(k-1)\left[\sum_{j=1}^{k} \hat{R}_j^2 - \left(\frac{kn^2}{4}\right)(kn+1)^2\right]}{\{[kn(kn+1)(2kn+1)]/6\} - \left(\frac{1}{k}\right)\sum_{i=1}^{n} \hat{R}_{i.}^2} \tag{4}$$

where $\hat{R}_{i.}$ is equal to the rank total of the ith dataset and $\hat{R}_{.j}$ is the rank total of the jth algorithm. The test statistic T is then compared for significance with a chi-square distribution for $k-1$ degrees freedom.

If the null hypothesis of the test is rejected, one can proceed with a post-hoc test in order to determine which specific pairwise comparisons produced differences. Post-hoc tests are designed to adjust α to control the family-wise error rate, which is the probability that at least one Type I error is made among the multiple pairwise tests that are being conducted. We implemented the Bergmann-Hommel procedure (Bergmann and Hommel 1988), which is recommended when all possible pairwise comparisons must be considered (Garcıa and Herrera).

We carried out the Friedman aligned ranks test with the Bergmann-Hommel post-hoc procedure in *R version 4.2.2* using the *scmamp* package (Calvo and Santafé 2016). This package provides functions for various statistical tests and visualizations that can be used to compare multiple learning algorithms over multiple problems.

2.5. Computational cost of model inferencing in HPC environment

Finally, we aimed to determine which architecture would be most favourable in operational deployment for pan-Arctic scale infrastructure mapping using HPC resources. Based on the evaluation and statistical comparison of the nine architectures, we found that four architectures (ResNet-50-UNet++, ResNet-50-MANet, ResNet-101-MANet, and ResNet-101-UNet++) achieved high generalization performances which were not significantly different from each other. This is described in detail in section 3.2. However, the optimal model should employ the architecture that simultaneously achieves high generalization performance with minimal expenditure of computational resources on HPC systems.

In elucidating this tradeoff, we conducted a scaling experiment on the Delta supercomputer in order to extrapolate service units (SUs) and runtime required for model inferencing on a small sample area to the pan-Arctic scale. The SU is the basic unit used in HPC systems to measure an amount of computation. Considering the Delta 4-way NVIDIA A100 GPU compute node, 1 SU corresponds to the equivalent use of 1 GPU, or fractional GPU, using less than or equal to 62.5 GB of memory, or 16 cores for 1 hour. In this study, we define runtime as the elapsed time between the start of a submitted job to its completion, excluding any waiting time in the job queue. Furthermore, runtime measurements only correspond to CNN model inferencing on all input image tiles, excluding any pre-processing or post-processing tasks.

We separately applied four trained models (using the aforementioned architectures) in inferencing mode to map Utqiagvik from a 20 sq. km. (or 0.183 GB) subset of a full WorldView-02 scene (3 GB), with the same development environment described in section 2.3.1, and recorded the expended SUs and runtime. To extrapolate these expenses to the pan-Arctic scale, we relied on a set of simple assumptions. Firstly, it has been estimated that there are ~ 1,000 settlements built on permafrost across the circumpolar Arctic (Ramage et al. 2021). We assumed that, on average, two Maxar satellite image scenes are required to cover the full extent of a settlement and that each image is 5 GB in size, resulting in a total of 10,000 GB (10 TB) worth of imagery that must be processed by a given CNN model in order to map all Arctic settlements. The ratio of these dataset sizes is then:

$$\frac{10,000GB}{0.183GB} \cong 54,645 \tag{5}$$

Therefore, to estimate the number of SUs and runtime required to process the hypothetical 10 TB pan-Arctic scale dataset, we multiplied the number of SUs and runtime (in hours) required to process the 0.183 GB sample area by 54,645. To be clear, here we assumed that these requirements scale linearly with dataset size.

2.6. Assessment of final model

After determining the candidate architecture that outperforms the rest, we trained a model using this architecture and evaluated its performance on a single held-out test split. We conducted hyperparameter tuning through grid search cross-validation to determine the optimal combination of batch size and learning rate. Utilizing 5-fold cross-validation, we trained a ResNet-50-UNet++ model for each possible

combination of batch size (4, 8, 16, 32) and learning rate (1e-4, 1e-3, 5e-3, 1e-2). The default learning rate for the Adam optimizer is typically set at 1e-3 in popular deep learning libraries, such as PyTorch and Keras. Therefore, we decided to experiment with values below and above this. It was found that a batch size of 16 with a learning rate of 1e-4 yielded the highest average F1-score out of all possible combinations at 0.857. Therefore, the model was trained with all of the same hyperparameters described in section 2.3.3. Furthermore, we applied basic geometric augmentations (random 90° rotation, horizontal and vertical flipping, and transposition) to the input training data.

As described in section 2.3.4, we used precision, recall and F1-score to assess the generalization performance of the model, but we also individually computed these metrics for each of the nine infrastructure classes. We conducted this final step in order to gain deeper insights into the ability of semantic segmentation models to map Arctic infrastructure from sub-metre resolution satellite imagery. Specifically, this step aided us in understanding the feasibility of our classification scheme as described in section 2.2.2.

3. Results

3.1. Model training and evaluation

Table 6 contains the estimated generalization performance, measured with F1-score, achieved by each architecture on the test split of each fold in 10-fold cross-validation. This dataset was the input to the statistical analysis as described in section 2.4. Additionally, Table 7 contains the associated runtimes required for model training with each architecture on each fold. Figures 7 and 8 contain the box plot visualizations of these achieved generalization performances and associated training runtimes.

ResNet-50-UNet++ achieved the highest average F1-score, while ResNet-101-DLV3+ conversely achieved the lowest average F1-score (Table 6). Through all encoder-decoder combinations, the UNet++ decoder yields the highest average F1-score, while the DeepLabV3+ decoder yields the lowest average F1-score. In examining training runtimes, we can observe that ResNet-101-UNet++ requires the longest runtime, while ResNet-34-DLV3+ requires the lowest (Table 7). This is expected since these architectures represent the maximum and minimum in terms of expected total network size, respectively. Through all encoder-decoder combinations, the DeepLabV3+ decoder requires the lowest average runtime.

Table 6. Generalization performance (F1-score) of each architecture estimated through 10-fold cross-validation.

Architecture	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Average
ResNet-34-UNet++	0.849	0.801	0.796	0.750	0.799	0.804	0.764	0.773	0.801	0.787	0.792
ResNet-34-DLV3+	0.803	0.757	0.771	0.707	0.742	0.770	0.765	0.749	0.762	0.757	0.758
ResNet-34-MANet	0.871	0.705	0.749	0.748	0.786	0.804	0.810	0.784	0.789	0.783	0.783
ResNet-50-UNet++	0.875	0.833	0.829	0.787	0.816	0.836	0.831	0.817	0.831	0.817	0.827
ResNet-50-DLV3+	0.809	0.770	0.758	0.719	0.765	0.779	0.774	0.754	0.790	0.766	0.768
ResNet-50-MANet	0.883	0.838	0.748	0.757	0.735	0.809	0.826	0.795	0.755	0.804	0.795
ResNet-101-UNet++	0.868	0.842	0.824	0.720	0.819	0.832	0.827	0.799	0.838	0.817	0.819
ResNet-101-DLV3+	0.781	0.751	0.756	0.706	0.735	0.767	0.778	0.724	0.769	0.748	0.752
ResNet-101-MANet	0.885	0.733	0.817	0.770	0.788	0.797	0.806	0.791	0.821	0.802	0.801

Table 7. Runtimes (seconds) required for model training with each architecture on each fold of 10-fold cross-validation.

Architecture	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	Average
ResNet-34-UNet++	3767.58	3544.97	3469.77	3485.15	3462.36	3484.26	3477.27	3459.20	3466.47	3484.15	3510.12
ResNet-34-DLV3+	3177.40	2946.20	2885.11	2895.98	2906.16	2965.40	3034.77	3033.21	3037.34	3027.52	2990.91
ResNet-34-MANet	3312.88	3140.61	3112.71	3120.08	3103.90	3082.84	3083.68	3091.12	3078.79	3078.41	3120.50
ResNet-50-UNet++	4082.59	3861.86	3926.39	3919.68	3909.51	3869.34	3898.23	3893.58	3880.22	3853.59	3909.50
ResNet-50-DLV3+	3240.31	3049.66	3044.82	3021.87	2993.39	2987.12	2953.72	2958.11	2942.69	2963.75	3015.54
ResNet-50-MANet	3809.62	3807.37	4212.75	3877.79	3775.98	3819.67	4052.32	4054.61	4065.96	4243.02	3971.91
ResNet-101-UNet++	4466.64	4589.22	4845.79	4631.99	4440.29	4927.04	4893.85	4903.78	4918.58	4361.23	4697.84
ResNet-101-DLV3+	3185.20	3133.02	3789.90	3486.47	3376.31	3214.77	3316.26	3542.65	3633.02	6610.64	3728.82
ResNet-101-MANet	3761.70	3801.49	4306.87	4094.90	3828.82	4089.96	4321.76	4279.41	4414.14	4450.21	4134.93

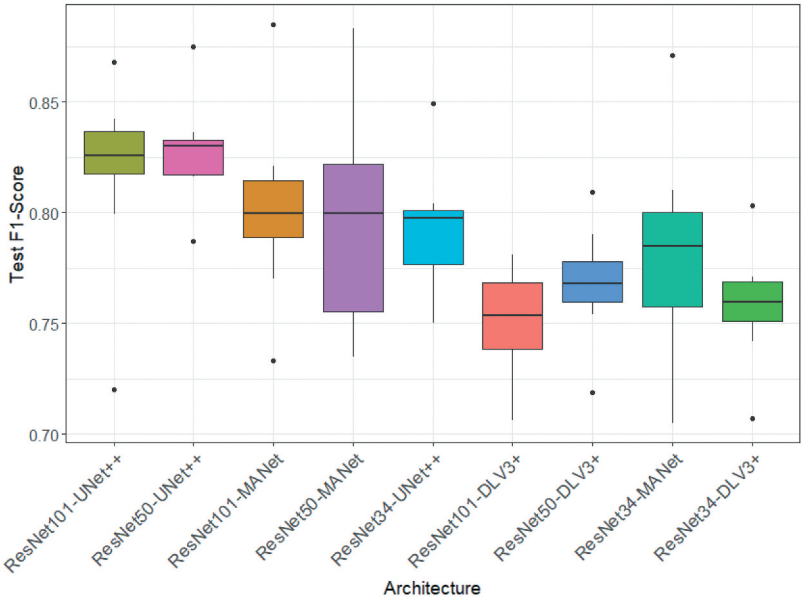

Figure 7. Boxplot of the generalization performance (F1-score) achieved by each architecture on the test split of each fold in 10-fold cross-validation.

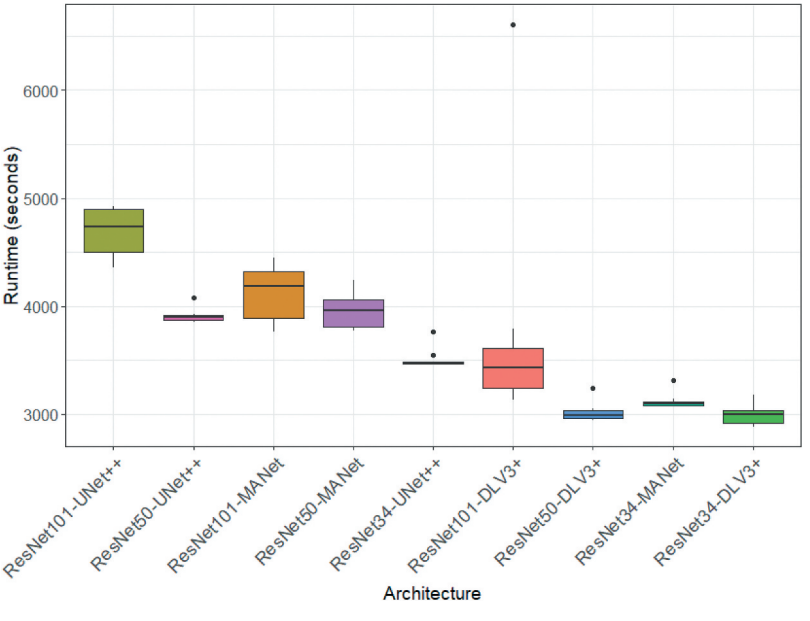

Figure 8. Boxplot of the model training runtimes (seconds) required by each architecture for each fold in 10-fold cross-validation.

There are outliers in the F1-scores achieved by all architectures except for ResNet-101-DLV3+ (Figure 7). The F1-scores of ResNet-50-MANet and ResNet-34-MANet exhibit notably high variances and are also both heavily skewed. This could suggest that these two architectures have poor generalization ability, as their performance across data folds is unstable. ResNet-50-UNet++ exhibits the highest median and low variance compared to the F1-scores of the other architectures. This is followed by ResNet-101-UNet++, which exhibits the second highest median F1-score, although a comparatively higher variance. This indicates that these two architectures have a high ability to generalize across unseen data for Arctic infrastructure mapping. The model training runtimes required by each architecture are generally stable (Figure 8), which is expected since the size of the training dataset is held constant throughout all simulations. However, ResNet-101-DLV3+ displays an extreme outlier. The reason for this is unclear but could possibly be attributed to some instability in the Delta supercomputer server.

3.2. Statistical comparison

The results of the Friedman aligned ranks test are given in Table 8. At $\alpha = 0.05$, we can reject the null hypothesis that the average ranks of each architecture are equal. Table 9 contains the adjusted p-values from subsequent post-hoc testing with the Bergmann-Hommel procedure. As shown in Table 9, the null hypotheses for all pairwise comparisons were accepted or rejected utilizing both $\alpha = 0.05$and $\alpha = 0.1$.

Furthermore, we visualize these post-hoc testing results in Figure 9, which represents the pairwise comparisons of all architectures as a network, where each node corresponds to an architecture and any connection between the nodes indicates that the null hypothesis was accepted (no significant difference in the performance of the architectures). The number displayed inside the node for a given architecture represents the F1-score ranking averaged across all folds. The architecture that achieved the lowest average F1-score ranking is highlighted in green. Figure 9a uses $\alpha = 0.05$while Figure 9b uses $\alpha = 0.1$to accept or reject the null hypothesis.

We observe that ResNet-50-UNet++ frequently significantly outperforms other architectures. At $\alpha = 0.05$, we can conclude that it performs significantly better than four out of the other eight architectures. At $\alpha = 0.1$, we can conclude that it performs significantly better than five out of the other eight architectures; no significant difference was found in the pairwise comparisons of ResNet-50-UNet++ to ResNet-50-MANet, ResNet-101-MANet, and ResNet-101-UNet++.

3.3. Computational cost of model inferencing in HPC environment

As determined by our scaling experiment, the number of SUs and inferencing time required for infrastructure mapping in a small target area and subsequent extrapolations

Table 8. Output of Friedman's aligned rank test for multiple comparisons.

T	df	p-value
50.331	8	3.53E–08

Table 9. Results of post-hoc test given as the adjusted p-values for the Friedman aligned ranks post-hoc test using the Bergmann-Hommel correction. Bolded p-value corresponds to rejected null hypothesis.

Index	Hypothesis	Bergmann-Hommel adjusted p-value
1	**ResNet-50-UNet++ vs. ResNet-101-DLV3+**	**0.0000014***
2	**ResNet-34-DLV3+ vs. ResNet-50-UNet++**	**0.0000086****
3	**ResNet-101-UNet++ vs. ResNet-101-DLV3+**	**0.0000387****
4	**ResNet-34-DLV3+ vs. ResNet-101-UNet++**	**0.0001796****
5	**ResNet-50-UNet++ vs. ResNet-50-DLV3+**	**0.0004280****
6	**ResNet-50-DLV3+ vs. ResNet-101-UNet++**	**0.0050247****
7	**ResNet-101-DLV3+ vs. ResNet-101-MANet**	**0.0116010****
8	**ResNet-34-MANet vs. ResNet-50-UNet++**	**0.0142748****
9	**ResNet-34-DLV3+ vs. ResNet-101-MANet**	**0.0320385****
10	**ResNet-50-MANet vs. ResNet-101-DLV3+**	**0.0686818***
11	**ResNet-34-UNet++ vs. ResNet-50-UNet++**	**0.0799039***
12	ResNet-34-UNet++ vs. ResNet-101-DLV3+	0.11515087
13	ResNet-34-MANet vs. ResNet-101-UNet++	0.11515087
14	ResNet-50-UNet++ vs. ResNet-50-MANet	0.115
15	ResNet-34-DLV3+ vs. ResNet-50-MANet	0.154
16	ResNet-34-UNet++ vs. ResNet-34-DLV3+	0.18750233
17	ResNet-50-DLV3+ vs. ResNet-101-MANet	0.29913884
18	ResNet-34-UNet++ vs. ResNet-101-UNet++	0.38844881
19	ResNet-34-MANet vs. ResNet-101-DLV3+	0.38844881
20	ResNet-50-MANet vs. ResNet-101-UNet++	0.42505906
21	ResNet-50-UNet++ vs. ResNet-101-MANet	0.425
22	ResNet-34-DLV3+ vs. ResNet-34-MANet	0.545
23	ResNet-50-DLV3+ vs. ResNet-50-MANet	0.76086623
24	ResNet-34-UNet++ vs. ResNet-50-DLV3+	0.85977882
25	ResNet-34-UNet++ vs. ResNet-34-MANet	1.000
26	ResNet-34-UNet++ vs. ResNet-50-MANet	1.000
27	ResNet-34-UNet++ vs. ResNet-101-MANet	1.000
28	ResNet-34-DLV3+ vs. ResNet-50-DLV3+	1.000
29	ResNet-34-DLV3+ vs. ResNet-101-DLV3+	1.000
30	ResNet-34-MANet vs. ResNet-50-DLV3+	1.000
31	ResNet-34-MANet vs. ResNet-50-MANet	1.000
32	ResNet-34-MANet vs. ResNet-101-MANet	1.000
33	ResNet-50-UNet++ vs. ResNet-101-UNet++	1.000
34	ResNet-50-DLV3+ vs. ResNet-101-DLV3+	1.000
35	ResNet-50-MANet vs. ResNet-101-MANet	1.000
36	ResNet-101-UNet++ vs. ResNet-101-MANet	1.000

(** rejected at α=0.05, * rejected at α=0.1).

to the pan-Arctic scale on the Delta supercomputer are shown in Figures 10 and 11. ResNet-50-UNet++ requires the least number of SUs to map infrastructure in Utqiagvik at 0.04, which extrapolates to 2,185.8 SUs at the pan-Arctic scale (Figure 10). ResNet-101-UNet++ requires 0.06 SUs to map Utqiagvik and 3,278.7 SUs at the pan-Arctic scale. Finally, both ResNet-50-MANet and ResNet-101-MANet require 0.07 SUs to map Utqiagvik and 3,825.15 SUs at the pan-Arctic scale.

In terms of model inferencing time, ResNet-50-UNet++ requires the least amount of time to map infrastructure in Utqiagvik at 41 seconds, which extrapolates to ~622 hours at the pan-Arctic scale (Figure 11). Both ResNet-101-UNet++ and ResNet-50-MANet take 43 seconds to map Utqiagvik and ~653 hours at the pan-Arctic scale. Finally, ResNet-101-MANet takes 49 seconds to map Utqiagvik and ~744 hours at the pan-Arctic scale. The differences amongst the architectures are proportional across the two scales as we have assumed the extrapolation follows a linear relationship.

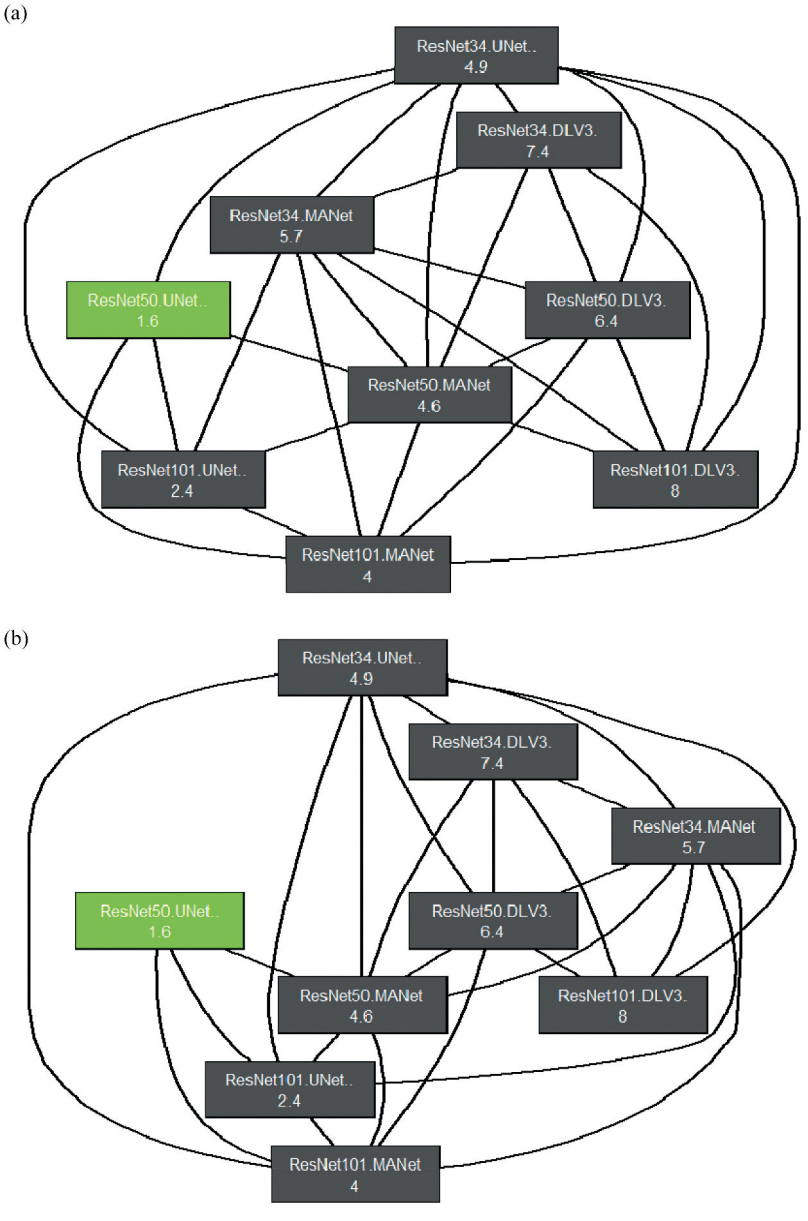

Figure 9. Post-hoc test represented as a network. Each node corresponds to an architecture. Connection between two nodes indicates that the null hypothesis is accepted at α=0.05 (a) or α=0.1 (b) and that there no significant difference in the performance of each architecture.

3.4. Assessment of final model

We trained and evaluated a final model with the ResNet-50-UNet++ architecture as our multi-objective comparison suggested that this architecture is the optimal candidate. In Figure 12, we provide the training F1-score and Focal loss curves. These show a fairly high degree of overlap between training and validation F1-score and loss, indicating that the

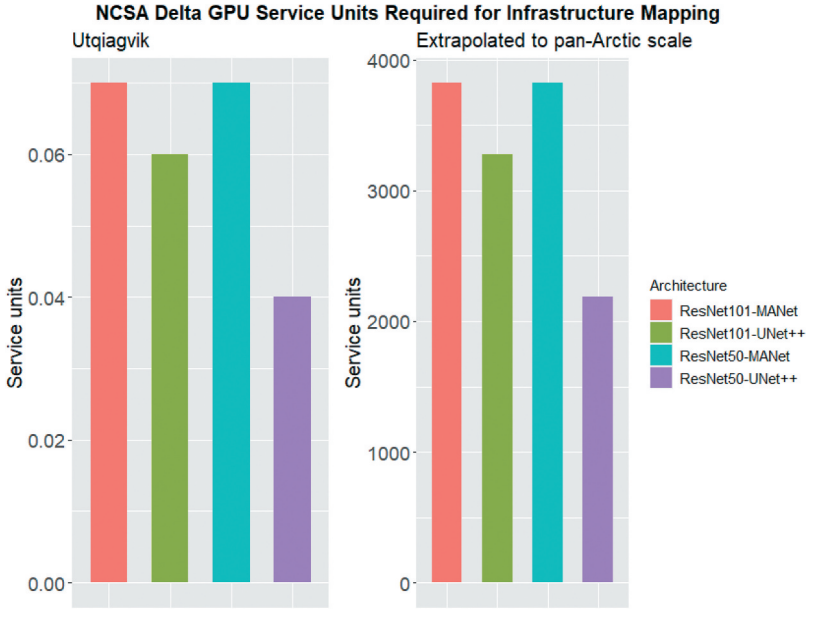

Figure 10. Required number of service units for Arctic infrastructure mapping in a small target area (Utqiagvik) and subsequent extrapolations to the pan-Arctic scale with four top-performing CNN architectures.

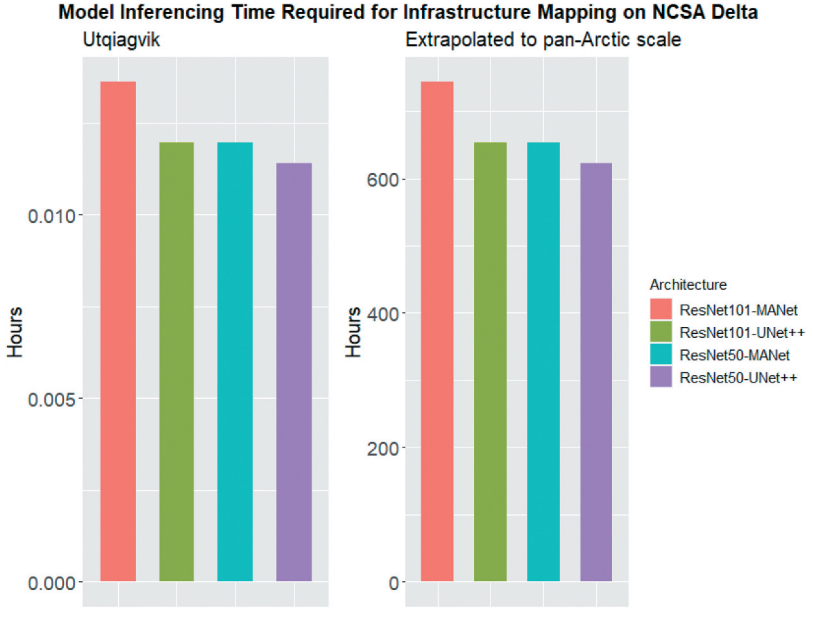

Figure 11. Model inferencing time in a small target area (Utqiagvik) and subsequent extrapolations to the pan-Arctic scale with four top-performing CNN architectures.

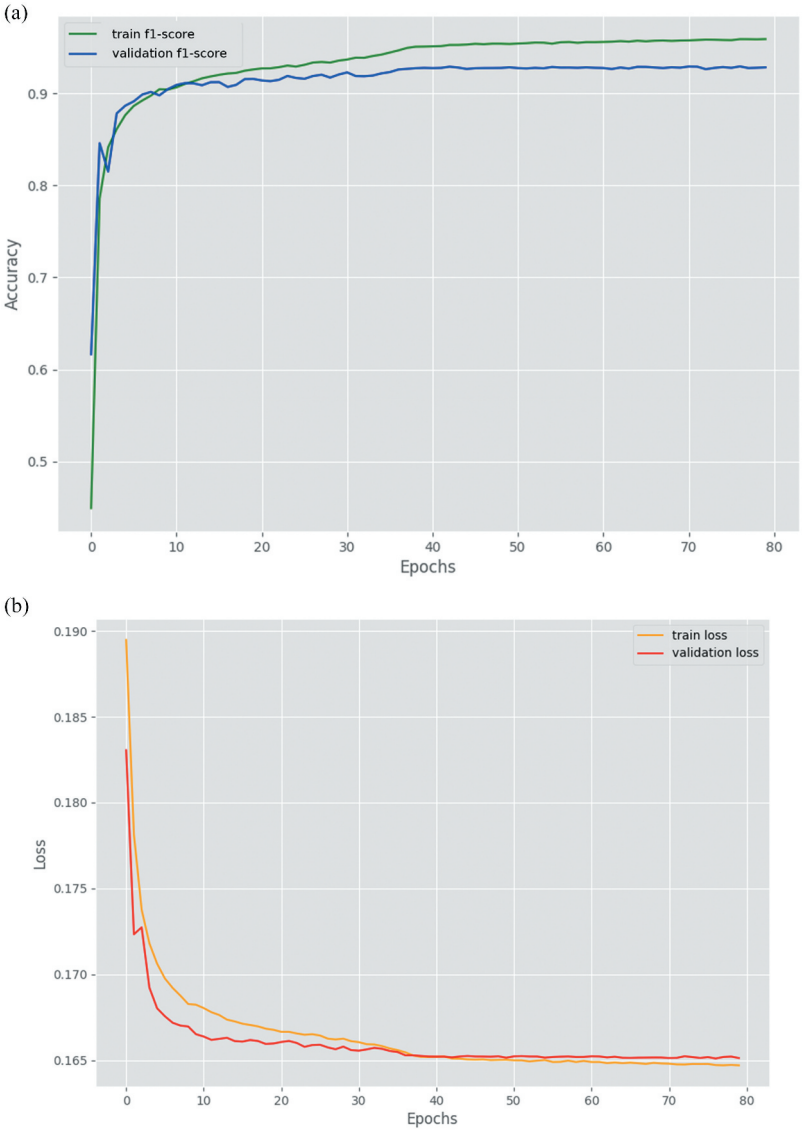

Figure 12. (a) training F1-score curve and (b) training Focal loss curve for final ResNet-50-UNet++ model.

model is able to generalize well on unseen data. Results of hyperparameter tuning through grid search cross-validation are shown in Figure 13. In Table 10, we provide the per-class and overall precision, recall, and F1-score that measure this model's performance on the test dataset. The overall metrics are provided as both unweighted averages and averages weighted based on the size of each class (measured as number of pixels).

Finally, Figure 14 displays several model predictions on input image tiles from each study site in the test dataset. These visual results demonstrate the model's capability to detect structures of various geometries, in a diverse range of environments, with accurate

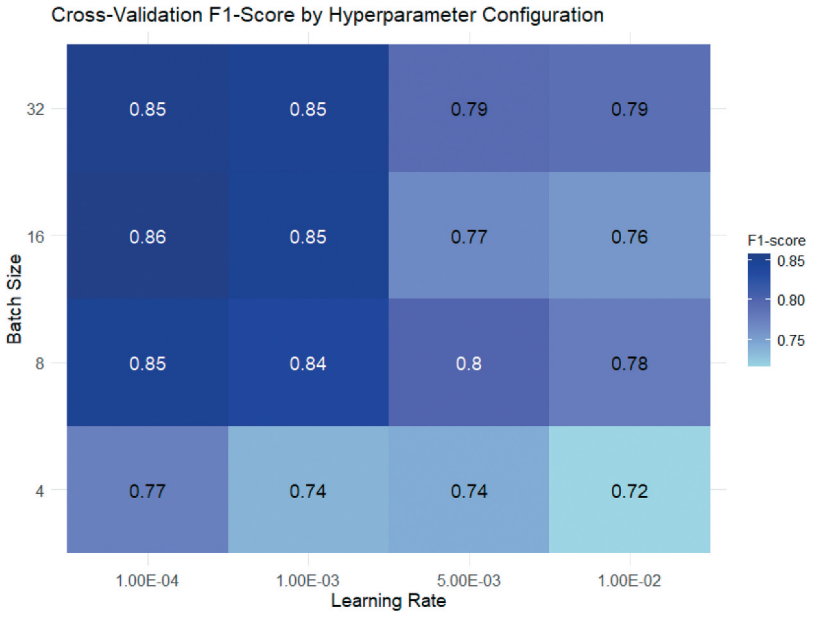

Figure 13. Average F1-scores achieved by the ResNet50-UNet++ model trained using different combinations of batch size and learning rate through grid search cross-validation.

Table 10. Generalization performance of final trained ResNet-50-UNet++ model, measured with per-class and overall precision, recall and F1-score.

Class	Precision	Recall	F1-score	Class size (# of pixels)
Background	0.96	0.97	0.97	14675456
Detached house	0.74	0.64	0.69	102741
Row house	0.00	0.00	0.00	40430
Multi-story block	0.88	0.82	0.85	459990
Non-residential	0.76	0.78	0.77	215343
Road	0.78	0.82	0.80	968426
Airport runway	0.92	0.84	0.88	672426
Gravel pad	0.86	0.80	0.83	928768
Pipeline	0.87	0.84	0.85	264014
Storage tank	0.00	0.00	0.00	22486
Unweighted average	**0.68**	**0.65**	**0.66**	
Weighted average	**0.93**	**0.94**	**0.93**	

delineation of boundaries and recognition of the structure type. For example, a comparison of Figure 14(e-h) demonstrate success in detecting unpaved roads with a subtle appearance in rural environments, as well as paved roads with a clear appearance in developed urban settings. Furthermore, Figure 14b, which contains larger houses that are clustered, and Figure 14e, which contains small houses that are sparsely distributed, demonstrate successful detection of detached houses of various sizes and spatial distributions. Finally, Figures 14a-d demonstrate the ability of the model to subcategorize building types; in each of these cases, the model recognizes the non-residential building among the detached housing units.

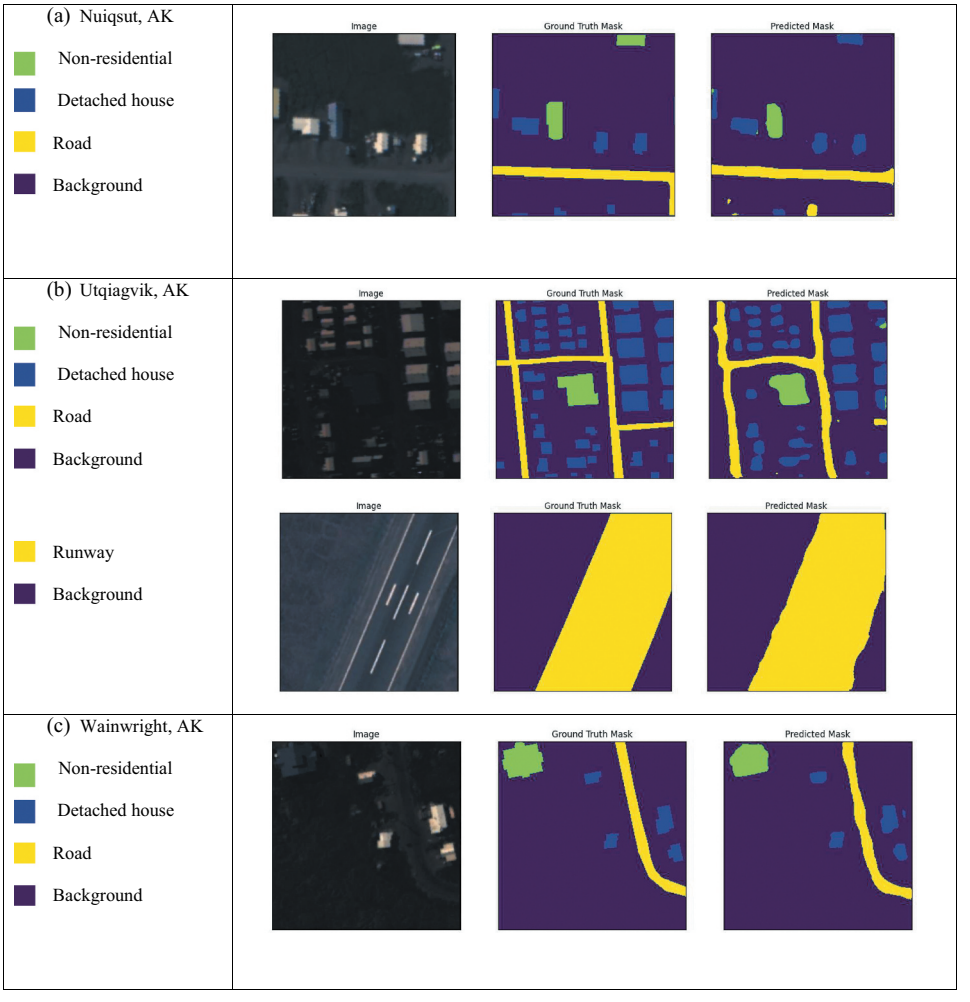

Figure 14. Model predictions on input image tiles from each study site in the test dataset. The model could not detect row houses and storage tanks; thus, they are omitted. In each row, the input image is on the left, the ground truth mask is in the centre, and the predicted output mask is on the right. The colour key for each predicted mask is also provided. *Imagery © Maxar, Inc.*

4. Discussion

4.1. Statistical comparison of semantic segmentation architectures

The Friedman aligned ranks test with the Bergmann-Hommel post-hoc procedure revealed that the ResNet-50-UNet++ architecture most frequently significantly outperforms the other architectures in the task of Arctic infrastructure mapping. Considering $\alpha = 0.1$, the generalization performance of ResNet-50-UNet++ was not significantly different from that achieved by either ResNet-50-MANet, ResNet-101-MANet, or ResNet-101-UNet++. As seen in the centre of each node in the network visualizations (Figure 9), these four architectures achieved the top four highest average ranks in terms of F1-score. Interestingly, these architectures are also the top four in terms of computational

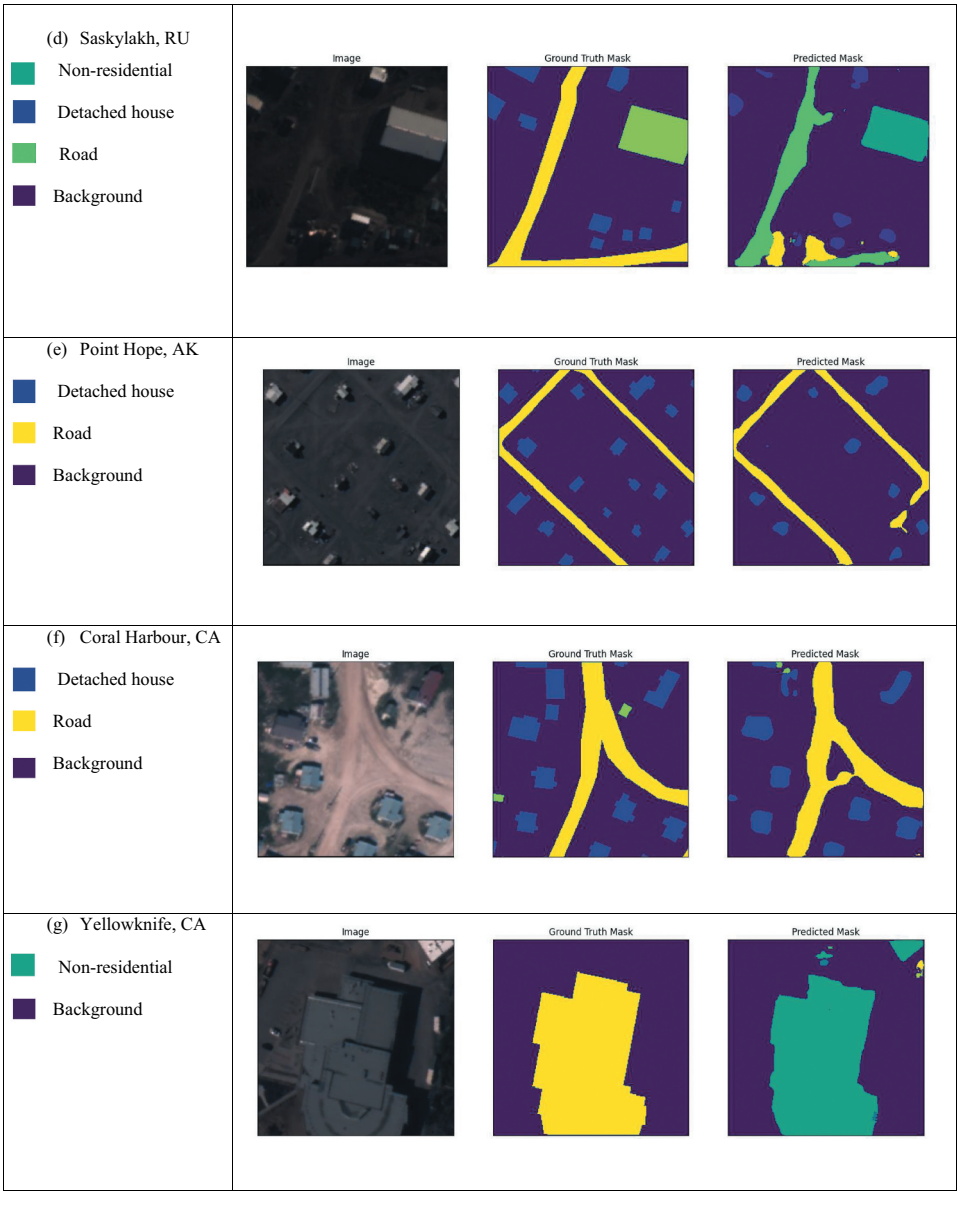

Figure 14. (Continued).

complexity. We hypothesize that this suggests a minimal amount of complexity required in a semantic segmentation architecture in order to effectively detect infrastructure based on our classification scheme. However, given the lack of significant differences between these four architectures, generalization performance likely plateaus once computational complexity increases past a certain point. For example, ResNet-50-UNet++ averaged an F1-score of 0.827 over 10-fold cross-validation, while ResNet-101-UNet++ averaged an F1-score of 0.819 (Table 6). While the difference is not significant, this could potentially be

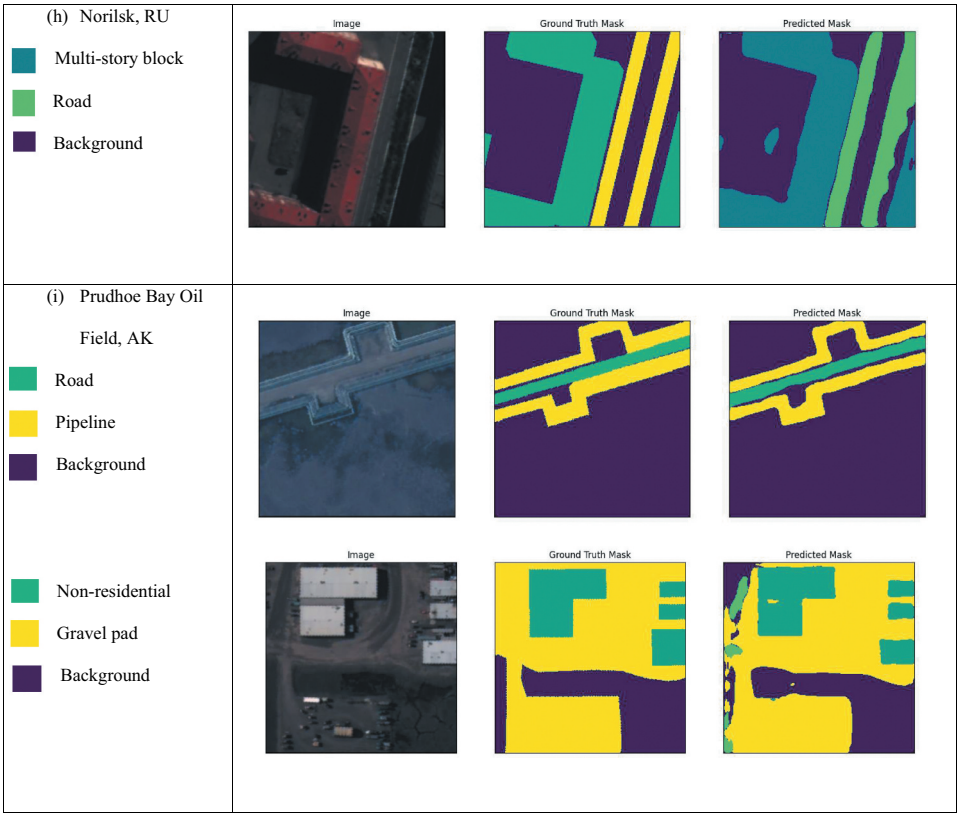

Figure 14. (Continued).

due to the fact that ResNet-101 causes slight overfitting, while ResNet-50 is more resilient to this phenomenon because of the comparatively lower number of parameters.

If we also consider the model training runtimes required by each architecture, (Table 7), it becomes clearer that ResNet-50-UNet++ might be the optimal architecture for this application task. With an average runtime of 3910 seconds, ResNet-50-UNet++ trains more quickly than ResNet-50-MANet, ResNet-101-MANet, and ResNet-101-UNet++. This difference is especially important when considering its implications for future model deployment on HPC clusters for large-scale mapping tasks, where we must be mindful of wall time and the amount of service units consumed. We did not statistically analyse these training runtimes, as they are expected to be directly related to the computational complexity of the given architectures that we have deliberately chosen ourselves.

At $\alpha = 0.05$, we additionally failed to reject the null hypothesis that the generalization performance of ResNet-50-UNet++ is significantly different from that of ResNet-34-UNet++. We suspect that this null hypothesis was likely not rejected at $\alpha = 0.05$due to a limited sample size or limited statistical power of the applied post-hoc procedure.

There are several other architectures that we have not tested, which could potentially perform better than ResNet-50-UNet++. For instance, we have only focused on encoders from the ResNet family, which only vary based on network size and not necessarily internal architecture. Other families of encoders, such as EfficientNet (Tan and Le 2020), DenseNet

(Huang et al. 2018), and VGG (Simonyan and Zisserman 2015) have achieved similar results as ResNet across many application domains. The same can be said for other decoders that we have not evaluated in this study, such as LinkNet (Chaurasia and Culurciello 2017). However, the challenge in comparing many architectures lies in the fact that a sufficient number of performance measurements must be taken in order to achieve meaningful results. If we decided to add the three aforementioned encoders to our pool of candidates in this study, the number of training/testing simulations would double from 90 to 180. This is likely logistically infeasible, so one must carefully select a limited number of candidate architectures to evaluate while also considering computational resources.

4.2. Computational cost of model inferencing in HPC environment

The difference between architectures grows to the order of thousands if we assume a linear extrapolation in required NCSA Delta GPU SUs for infrastructure mapping from a small target area (Figure 10). In our scaling experiment, we found that the ResNet-50-UNet++ model could expend 2,185.8 SUs if applied for inferencing on a pan-Arctic scale dataset. This is 1,092.9 SUs less than those required by the ResNet-101-UNet++ model and 1,639.35 SUs less than those required by both the ResNet-50-MANet and ResNet-101-MANet model. In terms of model inferencing time (Figure 11), the ResNet-50-UNet++ model can save up to 121 hours, or ~5 days. These findings are particularly significant given that SUs and HPC job runtime are limited resources. Deploying the ResNet-50-UNet ++ model for pan-Arctic scale infrastructure mapping as opposed to ResNet-101-MANet, for example, could ensure that we utilize our resource allocation efficiently.

4.3. Suitability of ResNet-50-UNet++ model as an infrastructure mapping tool

After determining that the ResNet-50-UNet++ architecture was most optimal for the task, we used it to train a final model and evaluate its performance in Arctic infrastructure mapping from sub-metre resolution satellite imagery. All nine infrastructure classes are detected with fairly high accuracy (Table 10). F1-scores for six out of nine classes were ≥ 0.80, with the F1-score for the 'non-residential building' class being a close exception at 0.77.

The 'detached house' class was detected with an F1-score 0.69, which indicates some room for improvement. Being small structures that are densely distributed in many areas naturally make detached houses difficult to accurately detect. Additionally, compared to the other classes that had F1-scores ≥0.80, detached houses were underrepresented in the training dataset. We did not account for the fact that because detached houses have small areas on average, many individual features must be digitized to balance their representation in the training dataset with other classes such as 'multi-story block', which are much larger structures that can quickly accumulate in the training dataset with comparatively fewer digitized features.

Row houses and storage tanks were not detected by the model at all, as they were severely underrepresented in the training dataset, with 169 and 119 polygon features in the dataset, respectively (see Table 5). Expanding our digitization efforts to new study sites would allow us to gather more samples for these two classes, since we have exhausted all of the existing samples of row houses and storage tanks in our nine study sites. Still, these kinds of structures do not naturally occur with the same

frequency as the other seven classes. Therefore, it may be worth considering discarding these classes in future training iterations if gathering more samples does not improve performance.

5. Conclusion

This study serves as a foundational step in the development of a deep learning-based approach to Arctic infrastructure mapping from sub-metre resolution satellite imagery. One of the most fundamental decisions in designing a deep learning pipeline is the choice of model architecture. Our multi-objective comparison approach identified the ResNet-50-UNet++ as the optimal CNN architecture based on both generalization performance and computational cost, which could potentially be used in the development of future pan-Arctic GeoAI mapping applications.

First, a statistically rigorous comparison of nine different semantic segmentation architectures found that the UNet++ decoder with the ResNet-50 encoder significantly outperformed five out of the other eight architectures (when $\alpha = 0.1$) in terms of F1-score. While not tested for statistical significance, we observed that this ResNet-50-UNet++ model trains quicker than those which it did not outperform. We then performed a scaling experiment to extrapolate SUs and model inferencing time expended for infrastructure mapping in Utqiagvik to a hypothetical pan-Arctic scale dataset. Results show that the ResNet-50-UNet++ model can save up to ~ 1,600 SUs (~54%), or ~120 hours of runtime (~18%), when compared to the other candidates, further supporting its suitability for this application task. However, there are many other 'state-of-the-art' encoders and decoders which we have not assessed in this study that may perform well in Arctic infrastructure mapping, so our conclusions can only safely apply to the set of candidate architectures that we have investigated.

We were able to map seven out of the nine types of infrastructure with high accuracy across all of our Arctic study sites; the model was not able to detect row houses or storage tanks due to data imbalance issues that can be addressed by expanding the training dataset as this work progresses. This suggests that deep learning-based mapping of major infrastructural features works well across a wide range of Arctic environments. However, we have not assessed the geographic transferability of our model, which would likely be one of the next steps in our ongoing research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is funded by the U.S. National Science Foundation's Office of Polar Programs (NSF-OPP) (grant No. 1927723, 1927872, and 2052107). Furthermore, this work used the Delta supercomputer at the National Center for Supercomputing Applications at the University of Illinois Urbana-Champaign through allocation #EES220055 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

ORCID

Elias Manos http://orcid.org/0000-0002-7350-0116

Data availability statement

The Python code used for deep learning model development will be made available at https://github.com/eliasm56/Arctic-Infrastructure-Detection-Paper. The training dataset used in this study is not available due to restrictions on sharing commercial satellite imagery.

References

"Alaska Geobotany Center - Organizations - Alaska Arctic Geoecological Atlas." 2022. Accessed December 11. https://arcticatlas.geobotany.org/catalog/organization/aaga.

Ardelean, F., A. Onaca, M.-A. Cheţan, A. Dornik, G. Georgievski, S. Hagemann, F. Timofte, and O. Berzescu. 2020. "Assessment of Spatio-Temporal Landscape Changes from VHR Images in Three Different Permafrost Areas in the Western Russian Arctic." *Remote Sensing* 12 (23): 3999. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/rs12233999.

Aytekın, Ö., A. Erener, İ. Ulusoy, and Ş. Düzgün. 2012. "Unsupervised Building Detection in Complex Urban Environments from Multispectral Satellite Imagery." *International Journal of Remote Sensing* 33 (7): 2152–2177. Taylor & Francis. https://doi.org/10.1080/01431161.2011.606852.

Badina, S. V. 2020. "Prediction of Socioeconomic Risks in the Cryolithic Zone of the Russian Arctic in the Context of Upcoming Climate Changes." *Studies on Russian Economic Development* 31 (4): 396–403. https://doi.org/10.1134/S1075700720040036.

Bartsch, A., G. Pointner, T. Ingeman-Nielsen, and L. Wenjun. 2020. "Towards Circumpolar Mapping of Arctic Settlements and Infrastructure Based on Sentinel-1 and Sentinel-2." *Remote Sensing* 12 (15): 2368. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/rs12152368.

Bartsch, A., G. Pointner, I. Nitze, A. Efimova, D. Jakober, S. Ley, E. Högström, G. Grosse, and P. Schweitzer. 2021. "Expanding Infrastructure and Growing Anthropogenic Impacts Along Arctic Coasts." *Environmental Research Letters* 16 (11): 115013. https://doi.org/10.1088/1748-9326/ac3176.

Bergmann, B., and G. Hommel. 1988. "Improvements of General Multiple Test Procedures for Redundant Systems of Hypotheses." In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, edited by P. Bauer, G. Hommel, and E. Sonnemann, 100–115. Berlin, Heidelberg: Medizinische Informatik und Statistik. https://doi.org/10.1007/978-3-642-52307-6_8.

Biskaborn, B. K., S. L. Smith, J. Noetzli, H. Matthes, G. Vieira, D. A. Streletskiy, P. Schoeneich, et al. 2019. "Permafrost is Warming at a Global Scale." *Nature Communications* 10 (1): 264. Nature Publishing Group. https://doi.org/10.1038/s41467-018-08240-4

Blaschke, T. 2010. "Object Based Image Analysis for Remote Sensing." *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (1): 2–16. https://doi.org/10.1016/j.isprsjprs.2009.06.004.

Blaschke, T., G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Queiroz Feitosa, et al. 2014. "Geographic Object-Based Image Analysis – Towards a New Paradigm." *ISPRS Journal of Photogrammetry and Remote Sensing* 87 (January): 180–191. https://doi.org/10.1016/j.isprsjprs.2013.09.014.

Blunden, J., and D. S. Arndt. 2017. "State of the Climate in 2016." *Bulletin of the American Meteorological Society* 98 (8): Si–S280. American Meteorological Society. https://doi.org/10.1175/2017BAMSStateoftheClimate.1.

Calvo, B., and G. Santafé. 2016. "Scmamp: Statistical Comparison of Multiple Algorithms in Multiple Problems." *The R Journal* 8 (1): 248. https://doi.org/10.32614/RJ-2016-017.

"CGS Planning & Lands - GIS Data (ESRI Shapefile)." 2023. Accessed April 12. https://cgs-pals.ca/downloads/gis/.

Chaurasia, A., and E. Culurciello. 2017. "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation." In *2017 IEEE Visual Communications and Image Processing* (*VCIP*), 1–4. https://doi.org/10.1109/VCIP.2017.8305148.

Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." *arXiv*. https://doi.org/10.48550/arXiv.1802.02611.

"Circum-Arctic Map of Permafrost and Ground-Ice Conditions, Version 2." 2022. *National Snow and Ice Data Center*. Accessed December 8. https://nsidc.org/data/ggd318/versions/2.

Comiso, J. C., and D. K. Hall. 2014. "Climate Trends in the Arctic as Observed from Space." *WIREs Climate Change* 5 (3): 389–409. https://doi.org/10.1002/wcc.277.

"Delta User Guide - Delta Supercomputer - NCSA Wiki." 2023. Accessed April 16. https://wiki.ncsa.illinois.edu/display/DSC/Delta+User+Guide#DeltaUserGuide-Table.4-wayNVIDIAA40GPUComputeNodeSpecifications.

Demšar, J. 2006. "Statistical Comparisons of Classifiers Over Multiple Data Sets." *The Journal of Machine learning research* 7:1–30.

Deng, J., W. Dong, R. Socher, L. Li-Jia, L. Kai, and L. Fei-Fei. 2009. "ImageNet: A Large-Scale Hierarchical Image Database." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Dobinski, W. 2011. "Permafrost." *Earth Science Review* 108 (3): 158–169. https://doi.org/10.1016/j.earscirev.2011.06.007.

Dore, M. H., I. Burton, and M. Dore. 2001. "The costs of adaptation to climate change in Canada: A stratified estimate by sectors and regions." *Canadian Climate Change Action Fund. St. Catherines, ON*.

Duro, D. C., S. E. Franklin, and M. G. Dubé. 2012. "A Comparison of Pixel-Based and Object-Based Image Analysis with Selected Machine Learning Algorithms for the Classification of Agricultural Landscapes Using SPOT-5 HRG Imagery." *Remote Sensing of Environment* 118 (March): 259–272. https://doi.org/10.1016/j.rse.2011.11.020.

Fan, T., G. Wang, L. Yan, and H. Wang. 2020. "MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation." *Institute of Electrical and Electronics Engineers Access* 8:179656–179665. https://doi.org/10.1109/ACCESS.2020.3025372.

Friedman, M. 1937. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance." *Journal of the American Statistical Association* 32 (200): 675–701. https://doi.org/10.1080/01621459.1937.10503522.

Friedman, M. 1940. "A Comparison of Alternative Tests of Significance for the Problem of m Rankings." *Annals of Mathematical Statistics* 11 (1): 86–92. Institute of Mathematical Statistics. https://doi.org/10.1214/aoms/1177731944.

Gadal, S., and W. Ouerghemmi. 2019. "Multi-Level Morphometric Characterization of Built-Up Areas and Change Detection in Siberian Sub-Arctic Urban Area: Yakutsk." *ISPRS International Journal of Geo-Information* 8 (3): 129. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/ijgi8030129.

García, S., A. Fernández, J. Luengo, and F. Herrera. 2010. "Advanced Nonparametric Tests for Multiple Comparisons in the Design of Experiments in Computational Intelligence and Data Mining: Experimental Analysis of Power." *Information Sciences* 180 (10): 2044–2064. https://doi.org/10.1016/j.ins.2009.12.010.

García, S., and F. Herrera. 2008. "An Extension on 'Statistical Comparisons of Classifiers Over Multiple Data Sets' for All Pairwise Comparisons." *Journal of machine learning research* 9 (12).

Gautier, D. L., K. J. Bird, R. R. Charpentier, A. Grantz, D. W. Houseknecht, T. R. Klett, T. E. Moore, et al. 2009. "Assessment of Undiscovered Oil and Gas in the Arctic." *Science* 324 (5931): 1175–1179. https://doi.org/10.1126/science.1169467.

Guerrero Vázquez, E., A. Yañez Escolano, P. Galindo Riaño, and J. Pizarro Junquera. 2001. "Repeated Measures Multiple Comparison Procedures Applied to Model Selection in Neural Networks." In *Bio-Inspired Applications of Connectionism* edited by, J. Mira and A. Prieto, Vol. 2085. 88–95. Berlin Heidelberg. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-45723-2_10 .

Hassol, S. J. 2004. *Impacts of a Warming Arctic: Arctic Climate Impact Assessment*. Cambridge, U.K.; New York, N.Y: Cambridge University Press.

He, K., X. Zhang, S. Ren, and J. Sun. 2015. "Deep Residual Learning for Image Recognition." *arXiv*. https://doi.org/10.48550/arXiv.1512.03385.

Hjort, J., D. Streletskiy, G. Doré, W. Qingbai, K. Bjella, and M. Luoto. 2022. "Impacts of Permafrost Degradation on Infrastructure." *Nature Reviews Earth and Environment* 3 (1): 24–38. Nature Publishing Group. https://doi.org/10.1038/s43017-021-00247-8.

Hodges, J. L., Jr, and E. L. Lehmann. 1962. "Rank Methods for Combination of Independent Experiments in Analysis of Variance." *Annals of Mathematical Statistics* 33 (2): 482–497. Institute of Mathematical Statistics. https://doi.org/10.1214/aoms/1177704575.

Hossain, K. 2017. "Arctic Melting: A New Economic Frontier and Global Geopolitics." *Current Developments in Arctic Law*: .

Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. 2018. "Densely Connected Convolutional Networks." *arXiv*. https://doi.org/10.48550/arXiv.1608.06993.

Iakubovskii, P. 2023. "Qubvel/Segmentation_models.Pytorch." Python. https://github.com/qubvel/segmentation_models.pytorch.

Instanes, A., and O. Anisimov. 2016. "Climate Change and Arctic Infrastructure." *Proceedings of the Ninth International Permafrost Conference* 1 (February): 779–784.

Japkowicz, N., and M. Shah. 2011. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511921803.

Khrustalev, L. N., S. Y. Parmuzin, and L. V. Emelyanova. 2011. *Reliability of Northern Infrastructure in Conditions of Changing Climate*. Moscow, RU: University Book Press Moscow.

Kingma, D. P., and J. Ba. 2017. "Adam: A Method for Stochastic Optimization." *arXiv*. http://arxiv.org/abs/1412.6980.

Kohavi, R. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, 1137–1143. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. IJCAI'95

Kumpula, T., B. C. Forbes, and F. Stammler. 2010. "Remote Sensing and Local Knowledge of Hydrocarbon Exploitation: The Case of Bovanenkovo, Yamal Peninsula, West Siberia, Russia." *ARCTIC* 63 (2): 165–178. https://doi.org/10.14430/arctic972.

Kumpula, T., B. Forbes, and F. Stammler. 2006. "Combining Data from Satellite Images and Reindeer Herders in Arctic Petroleum Development: The Case of Yamal, West Siberia." *Nordia Geographical Publications* 35 (2): 17–30.

Kumpula, T., B. C. Forbes, F. Stammler, and N. Meschtyb. 2012. "Dynamics of a Coupled System: Multi-Resolution Remote Sensing in Assessing Social-Ecological Responses During 25 Years of Gas Field Development in Arctic Russia." *Remote Sensing* 4 (4): Molecular Diversity Preservation International: 1046–1068. https://doi.org/10.3390/rs4041046.

Kumpula, T., A. Pajunen, E. Kaarlejärvi, B. C. Forbes, and F. Stammler. 2011. "Land Use and Land Cover Change in Arctic Russia: Ecological and Social Implications of Industrial Development." *Global Environmental Change* 21 (2): 550–562. https://doi.org/10.1016/j.gloenvcha.2010.12.010.

Larsen, P. H., S. Goldsmith, O. Smith, M. L. Wilson, K. Strzepek, P. Chinowsky, and B. Saylor. 2008. "Estimating Future Costs for Alaska Public Infrastructure at Risk from Climate Change." *Global Environmental Change* 18 (3): 442–457. https://doi.org/10.1016/j.gloenvcha.2008.03.005.

Larsen, J. N., and L. Huskey. 2015. "The Arctic Economy in a Global Context." In *The New Arctic*, edited by B. Evengård, J. N. Larsen, and Ø. Paasche, 159–174. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-17602-4_12.

Lehner, A., and T. Blaschke. 2019. "A Generic Classification Scheme for Urban Structure Types." *Remote Sensing* 11 (2): 173. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/rs11020173.

Li, J., X. Huang, L. Tu, T. Zhang, and L. Wang. 2022. "A Review of Building Detection from Very High Resolution Optical Remote Sensing Images." *GIScience & Remote Sensing* 59 (1): 1199–1225. Taylor & Francis. https://doi.org/10.1080/15481603.2022.2101727.

Li, M., L. Ma, T. Blaschke, L. Cheng, and D. Tiede. 2016. "A Systematic Comparison of Different Object-Based Classification Techniques Using High Spatial Resolution Imagery in Agricultural

Environments." *International Journal of Applied Earth Observation and Geoinformation* 49 (July): 87–98. https://doi.org/10.1016/j.jag.2016.01.011.

Lin, T.-Y., P. Goyal, R. Girshick, H. Kaiming, and P. Dollár. 2018. "Focal Loss for Dense Object Detection." *arXiv*. https://doi.org/10.48550/arXiv.1708.02002.

López-Serrano, P., M. Carlos, A. López-Sánchez, G.Á.-G. Juan, and J. García-Gutiérrez. 2016. "A Comparison of Machine Learning Techniques Applied to Landsat-5 TM Spectral Data for Biomass Estimation." *Canadian Journal of Remote Sensing* 42 (6): 690–705. https://doi.org/10.1080/07038992.2016.1217485.

Manos, E., C. Witharana, M. R. Udawalpola, A. Hasan, and A. K. Liljedahl. 2022. "Convolutional Neural Networks for Automated Built Infrastructure Detection in the Arctic Using Sub-Meter Spatial Resolution Satellite Imagery." *Remote Sensing* 14 (11): 2719. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/rs14112719.

Melnikov, V., V. Osipov, A. Brouchkov, A. Falaleeva, S. Badina, M. Zheleznyak, M. Sadurtdinov, et al. 2022. "Climate Warming and Permafrost Thaw in the Russian Arctic: Potential Economic Impacts on Public Infrastructure by 2050." *Natural Hazards* 112 (May): 1–21. https://doi.org/10.1007/s11069-021-05179-6.

Melvin, A. M., P. Larsen, B. Boehlert, J. E. Neumann, P. Chinowsky, X. Espinet, J. Martinich, et al. 2017. "Climate Change Damages to Alaska Public Infrastructure and the Economics of Proactive Adaptation." *Proceedings of the National Academy of Sciences* 114 (2): E122–E131. https://doi.org/10.1073/pnas.1611056113.

Molinaro, A. M., R. Simon, and R. M. Pfeiffer. 2005. "Prediction Error Estimation: A Comparison of Resampling Methods." *Bioinformatics* 21 (15): 3301–3307. https://doi.org/10.1093/bioinformatics/bti499.

Nelson, F. E., O. A. Anisimov, and N. I. Shiklomanov. 2001. "Subsidence Risk from Thawing Permafrost." *Nature* 410 (6831): 889–890. Nature Publishing Group. https://doi.org/10.1038/35073746.

Nhu, V.-H., A. Mohammadi, H. Shahabi, B. Bin Ahmad, N. Al-Ansari, A. Shirzadi, J. J. Clague, A. Jaafari, W. Chen, and H. Nguyen. 2020. "Landslide Susceptibility Mapping Using Machine Learning Algorithms and Remote Sensing Data in a Tropical Environment." *International Journal of Environmental Research and Public Health* 17:14. Multidisciplinary Digital Publishing Institute: 4933. https://doi.org/10.3390/ijerph17144933.

"NSB GIS Public." 2022. Accessed December 11. https://gis-public.north-slope.org/portal/home/.

Nymand Larsen, J., ed. 2014. "Arctic Human Development Report: Regional Processes and Global Linkages." In *TemaNord*, 567. Vol. 2014. Copenhagen: Nordic Council of Ministers.

"OpenDataTemplate." 2023. https://www.yellowknife.ca/en/inc/opendatatemplate.aspx.

Ourng, C., Y. Vaguet, and A. Derkacheva. 2019. "Spatio-Temporal Urban Growth Pattern in the Arctic: A Case Study in Surgut, Russia." *2019 Joint Urban Remote Sensing Event (JURSE)* 1–4. https://doi.org/10.1109/JURSE.2019.8809013.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." *C++ Advances in Neural Information Processing Systems 32* Curran Associates, Inc http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Peña, J. M., A. G. Pedro, C. Hervás-Martínez, J. Six, E. P. Richard, and F. López-Granados. 2014. "Object-Based Image Classification of Summer Crops with Machine Learning Methods." *Remote Sensing* 6 (6): 5019–5041. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/rs6065019.

Pizarro, J., E. Guerrero, and P. L. Galindo. 2002. "Multiple Comparison Procedures Applied to Model Selection." *Neurocomputing* 48 (1–4): 155–173. https://doi.org/10.1016/S0925-2312(01)00653-1.

Porfiriev, B. N., D. O. Eliseev, and D. A. Streletskiy. 2019. "Economic Assessment of Permafrost Degradation Effects on Road Infrastructure Sustainability Under Climate Change in the Russian Arctic." *Herald of the Russian Academy of Sciences* 89 (6): 567–576. https://doi.org/10.1134/S1019331619060121.

Porfiriev, B. N., D. O. Eliseev, and D. A. Streletskiy. 2021. "Economic Assessment of Permafrost Degradation Effects on Healthcare Facilities in the Russian Arctic." *Herald of the Russian Academy of Sciences* 91 (6): 677–686. https://doi.org/10.1134/S1019331621060113.

Porfiriev, B. N., D. O. Eliseev, and D. A. Streletskiy. 2021. "Economic Assessment of Permafrost Degradation Effects on the Housing Sector in the Russian Arctic." *Herald of the Russian Academy of Sciences* 91 (January): 17–25. https://doi.org/10.1134/S1019331621010068.

Programme (AMAP), Arctic Monitoring and Assessment. 2017. *Snow, Water, Ice and Permafrost in the Arctic (SWIPA) 2017*. Arctic Monitoring and Assessment Programme (AMAP). https://oaarchive.arctic-council.org/handle/11374/2105.

Ramage, J., L. Jungsberg, S. Wang, S. Westermann, H. Lantuit, and T. Heleniak. 2021. "Population Living on Permafrost in the Arctic." *Population and Environment* 43 (1): 22–38. https://doi.org/10.1007/s11111-020-00370-6.

Ramm, F. 2020. "OpenStreetMap Data in Layered GIS Format (Geofabrik)." Version 0.7.9. https://download.geofabrik.de/.

Raschka, S. 2020. "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning." *arXiv*. http://arxiv.org/abs/1811.12808.

Raynolds, M. K., D. A. Walker, K. J. Ambrosius, J. Brown, K. R. Everett, M. Kanevskiy, G. P. Kofinas, V. E. Romanovsky, Y. Shur, and P. J. Webber. 2014. "Cumulative Geoecological Effects of 62 Years of Infrastructure and Climate Change in Ice-Rich Permafrost Landscapes, Prudhoe Bay Oilfield, Alaska." *Global Change Biology* 20 (4): 1211–1224. https://doi.org/10.1111/gcb.12500.

Santafé, G., I. Inza, and J. Lozano. 2015. "Dealing with the Evaluation of Supervised Classification Algorithms." *Artificial Intelligence Review* 44 (June): 467–508. https://doi.org/10.1007/s10462-015-9433-y.

Simonyan, K., and A. Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *arXiv*. https://doi.org/10.48550/arXiv.1409.1556.

Streletskiy, D. A., S. Clemens, J.-P. Lanckman, and N. I. Shiklomanov. 2023. "The Costs of Arctic Infrastructure Damages Due to Permafrost Degradation." *Environmental Research Letters* 18 (1): 015006. https://doi.org/10.1088/1748-9326/acab18.

Streletskiy, D. A., A. B. Sherstiukov, O. W. Frauenfeld, and F. E. Nelson. 2015. "Changes in the 1963–2013 Shallow Ground Thermal Regime in Russian Permafrost Regions." *Environmental Research Letters* 10 (12): 125005. IOP Publishing. https://doi.org/10.1088/1748-9326/10/12/125005.

Streletskiy, D. A., N. I. Shiklomanov, J. D. Little, F. E. Nelson, J. Brown, K. E. Nyland, and A. E. Klene. 2017. "Thaw Subsidence in Undisturbed Tundra Landscapes, Barrow, Alaska, 1962–2015." *Permafrost and Periglacial Processes* 28 (3): 566–572. https://doi.org/10.1002/ppp.1918.

Streletskiy, D. A., L. J. Suter, N. I. Shiklomanov, B. N. Porfiriev, and D. O. Eliseev. 2019. "Assessment of Climate Change Impacts on Buildings, Structures and Infrastructure in the Russian Regions on Permafrost." *Environmental Research Letters* 14 (2): 025003. https://doi.org/10.1088/1748-9326/aaf5e6.

Suter, L., D. Streletskiy, and N. Shiklomanov. 2019. "Assessment of the Cost of Climate Change Impacts on Critical Infrastructure in the Circumpolar Arctic." *Polar Geography* 42 (4): 267–286. Taylor & Francis. https://doi.org/10.1080/1088937X.2019.1686082.

Tan, M., and Q. Le. 2020. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." *arXiv*. https://doi.org/10.48550/arXiv.1905.11946.

Udawalpola, M., A. Hasan, A. Liljedahl, A. Soliman, J. Terstriep, and C. Witharana. 2022. "An Optimal GeoAi Workflow for Pan-Arctic Permafrost Feature Detection from High-Resolution Satellite Imagery." *Photogrammetric Engineering & Remote Sensing* 88 (March): 181–188. https://doi.org/10.14358/PERS.21-00059R2.

Udawalpola, M., A. Hasan, A. K. Liljedahl, A. Soliman, and C. Witharana. 2021. "OPERATIONAL-SCALE GEOAI for PAN-ARCTIC PERMAFROST FEATURE DETECTION from HIGH-RESOLUTION SATELLITE IMAGERY." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (August): 175–180. XLIV-M-3–2021. https://doi.org/10.5194/isprs-archives-XLIV-M-3-2021-175-2021

Witharana, C., M. Abul Ehsan Bhuiyan, and A. Liljedahl. 2020. "BIG IMAGERY AND HIGH PERFORMANCE COMPUTING AS RESOURCES TO UNDERSTAND CHANGING ARCTIC POLYGONAL

TUNDRA." *ISPRS - International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences* XLIV-M-2-2020 (November): XLIV-M-2–2020. https://doi.org/10.5194/isprs-archives-XLIV-M-2-2020-111-2020.

Witharana, C., M. R. Udawalpola, A. K. Liljedahl, M. K. Jones, B. M. Jones, A. Hasan, D. Joshi, and E. Manos. 2022. "Automated Detection of Retrogressive Thaw Slumps in the High Arctic Using High-Resolution Satellite Imagery." *Remote Sensing* 14 (17): 4132. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/rs14174132.

Witharana, C., M. R. Udawalpola, A. S. Perera, A. Hasan, E. Manos, A. Liljedahl, M. Kanevskiy, et al. 2023. *Ice-Wedge Polygon Detection in Satellite Imagery from Pan-Arctic Regions, Permafrost Discovery Gateway, 2001-2021*. Arctic Data Center. https://doi.org/10.18739/A2KW57K57.

Zar, J. H. 1999. *Biostatistical Analysis*. 4th ed. Upper Saddle River, NJ: Prentice Hall.

Zhou, Z., M. Mahfuzur Rahman Siddiquee, N. Tajbakhsh, and J. Liang. 2018. "UNet++: A Nested U-Net Architecture for Medical Image Segmentation." *arXiv*. https://doi.org/10.48550/arXiv.1807.10165.