Environmental Data Science (2023), 2: e42, 1–18
doi:10.1017/eds.2023.30

CAMBRIDGE
UNIVERSITY PRESS

METHODS PAPER

Extending scene-to-patch models: Multi-resolution multiple instance learning for Earth observation

Joseph Early[1,2], Ying-Jung Chen Deweese[3,4], Christine Evers[1] and Sarvapali Ramchurn[1,2]

[1]Agents, Interaction, and Complexity Group, University of Southampton, Southampton, United Kingdom
[2]The Alan Turing Institute, London, United Kingdom
[3]School of Computer Science, Georgia Institute of Technology, Atlanta, Georgia, USA
[4]Descartes Labs, Santa Fe, New Mexico, USA
Corresponding author: Joseph Early; Email: joseph.early.ai@gmail.com

Received: 28 February 2023; **Revised:** 28 June 2023; **Accepted:** 01 September 2023

Keywords: Earth observation; remote sensing; multi-resolution; multiple instance learning; scene-to-patch models

Abstract

Land cover classification (LCC) and natural disaster response (NDR) are important issues in climate change mitigation and adaptation. Existing approaches that use machine learning with Earth observation (EO) imaging data for LCC and NDR often rely on fully annotated and segmented datasets. Creating these datasets requires a large amount of effort, and a lack of suitable datasets has become an obstacle in scaling the use of machine learning for EO. In this study, we extend our prior work on Scene-to-Patch models: an alternative machine learning approach for EO that utilizes Multiple Instance Learning (MIL). As our approach only requires high-level scene labels, it enables much faster development of new datasets while still providing segmentation through patch-level predictions, ultimately increasing the accessibility of using machine learning for EO. We propose new multi-resolution MIL architectures that outperform single-resolution MIL models and non-MIL baselines on the DeepGlobe LCC and FloodNet NDR datasets. In addition, we conduct a thorough analysis of model performance and interpretability.

Impact Statement

This method paper proposes a new method for segmenting Earth observation (EO) images without requiring expensive labeling processes. While it is demonstrated in two domains (land cover classification from satellite images and flooding detection from aerial images), the method is intended to be generally applicable and lays the foundation for expanding the use of EO imaging data in technology, government, and academia.

1. Introduction

To tackle critical problems such as climate change and natural disaster response (NDR), it is essential to collect information and monitor ongoing changes in Earth systems (Oddo and Bolten, 2019). Remote sensing (RS) for Earth observation (EO) is the process of acquiring data using instruments equipped with sensors, for example, satellites or unmanned aerial vehicles (UAVs). Due to ever-expanding volumes of

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

EO data, machine learning (ML) has been widely used in automated RS monitoring for a range of applications. However, curation of the large labeled EO datasets required for ML is expensive—accurately labeling the data takes a lot of time and often requires specific domain knowledge. This results in bottlenecks and concerns about a lack of suitable EO datasets for ML (Hoeser and Kuenzer, 2020; CCAI, 2022).

In this work, we focus on two EO RS applications: land cover classification (LCC) and NDR. LCC is an important task in EO and can be applied to agricultural health/yield monitoring, deforestation, sustainable development, urban planning, and water availability (Demir et al., 2018; Hoeser et al., 2020). Natural disasters (floods, earthquakes, etc.) are dynamic processes that require rapid response to save assets and lives. ML approaches can provide near real-time monitoring and change detection for NDR, as well as mitigate supply chain issues (Oddo and Bolten, 2019). A common objective in both LCC and NDR is image segmentation (Hoeser et al., 2020; Munawar et al., 2021). This involves separating images into different classes (e.g., urban land, forest land, agricultural land, etc.) or objects (e.g., houses, vehicles, people, etc.). Training ML models to perform automated segmentation typically requires an EO dataset that has already been segmented, that is, each image has a corresponding set of annotations that mark out the different objects or class regions. The annotation process often has to be done by hand and is very expensive and time-consuming. For example, each image in the FloodNet dataset (used in this work) took approximately 1 hr to annotate (Rahnemoonfar et al., 2021).

In our prior work (Early et al., 2022b), we proposed a new approach to segmentation for EO, in which cost- and time-intensive segmentation labels are not required. Instead, only scene-level summaries are needed (which are much quicker to curate). This was achieved by reframing segmentation as a Multiple Instance Learning (MIL) regression problem. Our previously proposed Scene-to-Patch (S2P) MIL models produce segmented outputs without requiring segmentation labels and are also able to preserve high-resolution data during training and inference. They are effectively able to transform the low-resolution scene labels used in training into high-resolution patch predictions. However, our prior S2P approach only operated at a single resolution, and we only studied LCC using satellite data. In this work, we extend S2P models to use multiple resolutions and also apply them to NDR from aerial imagery. Specifically, our contributions are as follows:

1. We extend our prior S2P approach to utilize multi-resolution inputs and make multi-resolution predictions.
2. We apply our novel multi-resolution approach to LCC (satellite data) and disaster response monitoring (aerial imagery), comparing it to single-resolution baselines.
3. We investigate how changing the configuration of our approach affects its performance.
4. We show how our approach is inherently interpretable at multiple resolutions, allowing segmentation of EO images without requiring pixel-level labels during training.

The rest of this paper is laid out as follows. Section 2 details the background literature and related work. Section 3 reintroduces our S2P approach and extends it to a novel multi-resolution method. Our experiments are presented in Section 4, with a further discussion in Section 5. Section 6 concludes the paper.

2. Background and Related Work

In this section, we provide the background literature to our work, covering LCC, NDR, and their existing ML solutions. We also discuss MIL and its multi-resolution extensions.

LCC. In EO, ML can be applied to monitoring and forecasting land surface patterns, involving ecological, hydrological, agricultural, and socioeconomic domains (Liu et al., 2018, 2017). For example, satellite images can be used to track carbon sequestration and emission sources, which is useful for monitoring greenhouse gas (GHG) levels (Rolnick et al., 2022). The way land is used is both impacted by and contributes to climate change—it is estimated that land use is responsible for around a quarter of global GHG emissions, and improved land management could lead to a reduction of about a third of emissions (Rolnick et al., 2022). Furthermore, changes in land use can have significant effects on the carbon balance within ecosystem services that contribute to climate change mitigation (Friedlingstein et al., 2020). In order to work towards UN Net Zero emission targets (Sadhukhan, 2022), there is a need for improved understanding and monitoring of how land is used, that is, real-time monitoring that facilitates better policy design, planning, and enforcement (Kaack et al., 2022). For example, automated LCC with ML can be used to determine the effect of regulation or incentives to drive better land use practices (Rolnick et al., 2022), and for monitoring the amount of acreage in use for farmland, allowing the assessment of food security and GHG emissions (Ullah et al., 2022).

NDR. The increasing magnitude and frequency of extreme weather events driven by climate change is accelerating the change of suitable land areas for cropland and human settlement (Elsen et al., 2022), and raising the need for timely and effective NDR. With recent advances in EO, RS is a valuable approach in NDR efforts, providing near real-time information to help emergency responders execute their response efforts, plan emergency routes, and identify effective lifesaving strategies. For instance, Sentinel-2 data has been used for flood response (Caballero et al., 2019), synthetic aperture radar (SAR) imagery has been used for mapping landslides (Burrows et al., 2019) and wildfire patterns (Ban et al., 2020), and helicopter and UAV imagery has been used to identify damaged buildings and debris following hurricanes (Pi et al., 2020). RS is only one aspect of utilizing automated data processing techniques to facilitate improved NDR; other paradigms include digital twins (Fan et al., 2021) and natural language processing (Zhang et al., 2019).

Existing approaches. A common problem type in both LCC and NDR is image segmentation. Here, the aim is to assign each pixel in an input image to a class, such that different objects or regions in the original image are separated and classified. For example, LCC segmentation typically uses classes such as urban, agricultural, forest, and so forth. In image segmentation settings, most models require the original training images to be annotated with segmentation labels, that is, all pixels are labelled with a ground-truth class. There are several existing approaches to image segmentation, such as Fully Convolutional Networks (Long et al., 2015), U-Net (Ronneberger et al., 2015), and Pyramid Networks (Lin et al., 2017). Existing works have applied these or similar approaches to LCC (Kuo et al., 2018; Rakhlin et al., 2018; Seferbekov et al., 2018; Tong et al., 2020; Wang et al., 2020; Karra et al., 2021); we refer readers to Hoeser and Kuenzer (2020) for a more in-depth review of existing work. For NDR, examples of existing work include wildfire segmentation with U-Net (Khryashchev and Larionov, 2020), and flooding segmentation using Multi3Net (Rudner et al., 2019).

MIL. In conventional supervised learning, each piece of data is given a label. However, in MIL, data are grouped into bags of instances, and only the bags are labeled, not the instances (Carbonneau et al., 2018). This reduces the burden of labeling, as only the bags need to be labeled, not every instance. In this work, we utilize MIL neural networks (Wang et al., 2018). Extensions such as attention (Ilse et al., 2018), graph neural networks (Tu et al., 2019), and long short-term memory (Wang et al., 2021, 2020; Early et al., 2022a) exist, but these are not explored in this work. MIL has previously been used in EO data, for example, fusing panchromatic and multi-spectral images (Liu et al., 2018), settlement extraction (Vatsavai et al., 2013), landslide mapping (Zhang et al., 2021), crop yield prediction (Wang et al., 2012), and scene classification (Wang et al., 2022). However, to the best of the authors' knowledge, our prior work (Early et al., 2022b) was the first to study the use of MIL for generic multi-class LCC.

Multi-resolution MIL. When using MIL in imaging applications, patches are extracted from the original images to form bags of instances. As part of this patch extraction process, it is necessary to choose the number and size of patches, which determines the effective resolution at which the MIL model is operating (and affects the overall performance of the model). There is an inherent trade-off when choosing the patch size: patches must be large enough to capture meaningful information, but small enough such that they can be processed with ML (or that downsampling does not lead to a loss of detail). To overcome this trade-off, prior research has investigated the use of multi-resolution approaches, where several different patch sizes are used (if only a single patch size is used, the model is considered to be operating at a single resolution). Existing multi-resolution MIL approaches typically focus on medical imaging (Hashimoto et al., 2020; Li et al., 2021; Li et al., 2020; Marini et al., 2021) as opposed to EO data (we discuss these methods further in Section 3.3). Non-MIL multi-resolution approaches have been applied in EO domains such as ship detection (Wang et al., 2019) and LCC (Robinson et al., 2019), but to the best of the authors' knowledge, we are the first to propose the use of multi-resolution MIL methods for generic EO. In the next section, we reintroduce our prior MIL S2P method (Early et al., 2022b), and extend it to a multi-resolution approach.

3. Methodology

In this section, we give our S2P methodology. First, in Sections 3.1 and 3.2, we reintroduce and provide further detail on our prior single-resolution S2P approach (Early et al., 2022b). We then propose our S2P extension to multi-resolution settings, explaining our novel model architecture (Section 3.3), and how it combines information from multiple resolutions (Section 3.4).

3.1. Scene-to-patch overview

Following the nomenclature of our prior work, there are three tiers of operation in EO data:

1. **Scene level**: The image is considered as a whole. Classic convolutional neural networks (CNNs) are examples of scene-level models, where the entire image is used as input, and each image has a single class label, for example, brick kiln identification (Lee et al., 2021).
2. **Patch level**: Images are split into small patches (typically tens or hundreds of pixels). This is a standard approach in MIL, where a group of patches extracted from the same image are a bag of instances. In most cases, labeling and prediction occur only at the bag (scene) level. However, depending on the dataset and MIL model, labeling and prediction can occur at the patch level.
3. **Pixel level**: Labelling and predictions are performed on the scale of individual pixels. Segmentation models (e.g., U-Net; Ronneberger et al., 2015) are pixel-level approaches. They typically require pixel-level labels (i.e., segmented images) to learn to make pixel-level predictions. A notable exception in EO research is Wang et al. (2020), where U-Net models are trained from scene-level labels without requiring pixel-level annotations.

In our prior work, we proposed a novel approach for EO, where both scene- and patch-level predictions can be made while only requiring scene-level labels for training. This is achieved by reframing pixel-level segmentation as scene-level regression, where the objective is to predict the coverage proportion of each segmentation class. The motivation for such an approach is that scene-level labels are easy to procure (as they are summaries of the content of an image as opposed to detailed pixel-level annotations), which helps expedite the labeling process. Furthermore, it also reduces the likelihood of label errors, which often occur in EO segmentation datasets (Rakhlin et al., 2018). However, scene-level predictions cannot be used for segmentation. Therefore, it is necessary to have a model that learns from scene-level labels but can produce patch- or pixel-level predictions—our prior S2P approach achieved this by using inherently interpretable MIL models. We provide a high-level overview of our S2P approach in Figure 1 and discuss our model architecture in more detail in the next section.

3.2. Single resolution scene-to-patch approach

Our original S2P approach utilized an Instance Space Neural Network (known as mi-Net in Wang et al., 2018), which operates at a single input and output resolution. In Figure 2, we give an example S2P architecture for the DeepGlobe dataset (Section 4.1), which has seven classes. The model takes a bag containing b instance patches as input, where each patch has three channels (red, green, and blue) and is 102 × 102 px. The patches are independently passed through a feature extractor, resulting in b

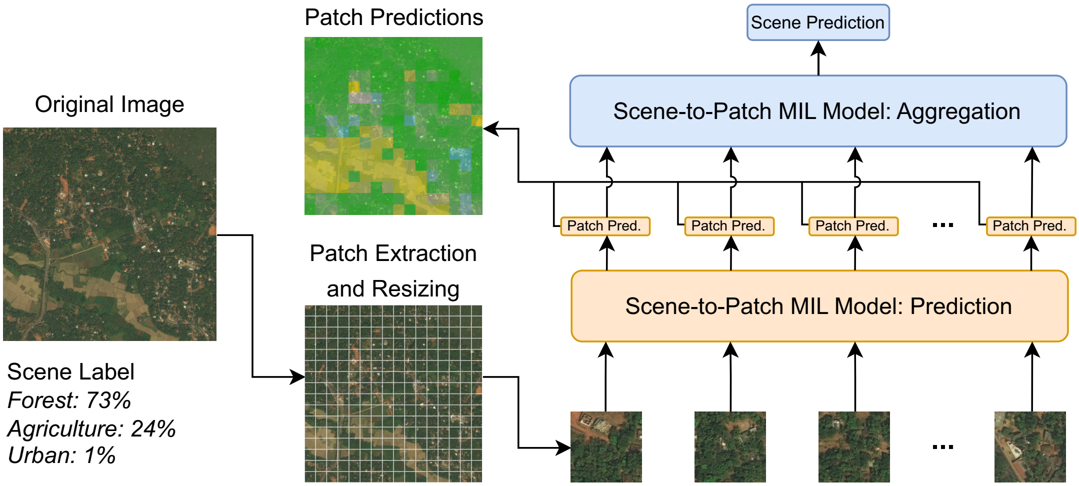

Figure 1. *MIL scene-to-patch overview. The model produces both instance (patch) and bag (scene) predictions but only learns from scene-level labels. Example from the DeepGlobe dataset (Section 4.1).*

patch embeddings; each of length 128. These patch embeddings are then classified, and the mean of the patch predictions is used as the overall bag prediction. Note these models are trained end-to-end and only learn from scene-level (bag) labels—no patch-level labels are used during training.

Formally, for a collection of n EO images $\mathcal{X} = \{X_1, \dots, X_n\}$, each image $X_i \in \mathcal{X}$ has a corresponding scene-level label $Y_i \in \mathcal{Y}$, where $Y_i = \left\{Y_i^1, \dots, Y_i^C\right\}$. C is the number of classes (e.g., types of land cover), and Y_i^c is the coverage proportion for class c in image X_i, such that $\sum_{c=1}^{C} Y_i^c = 1$. For each EO image $X_i \in \mathcal{X}$, a set of b patches $\left\{x_i^1, \dots, x_i^b\right\}$ is extracted, and our S2P models make a prediction $\hat{y}_i^j$ for each patch. The mean of these patch-level predictions is then taken to make an overall scene-level prediction $\hat{Y}_i = \frac{1}{b}\sum_{j=1}^{b} \hat{y}_i^j$.

The configuration of our S2P model shown in Figure 2 uses three convolutional layers in the feature extractor model, but we also experiment with only using two convolution layers (see Supplementary Appendix C.3 for further details). The S2P models are relatively small (in comparison to baseline architectures such as ResNet18 and U-Net; see Supplementary Appendix C.1), which means they are less expensive to train and run. This reduces the GHG emission effect of model development and deployment, which is an increasingly important consideration for the use of ML (Kaack et al., 2022). Furthermore, the S2P approach makes patch predictions inherently—post-hoc interpretability methods such as MIL local interpretations (Early et al., 2022c) are not required.

Reframing EO segmentation as a regression problem does not necessitate a MIL approach; it can be treated as a traditional supervised regression problem. However, as EO images are often very high resolution, the images would have to be downsampled, and, as such, important data would be lost. With MIL, it is possible to operate at higher resolutions than a purely supervised learning approach. In our prior work, we experimented with different resolutions, but the S2P models were only able to operate at a single resolution. In the next section, we discuss our extension to S2P models which allows processing of multiple resolutions within a single model.

3.3. Multi-resolution scene-to-patch approaches

The objective of our S2P extension is to utilize multiple resolutions within a single model. While our prior work investigated the use of different resolutions, it only did so with separate models for each resolution. This does not allow information to be shared between different resolutions, which may hinder performance. For example, in aerial imagery, larger features such as a house might be easy to classify at lower

https://doi.org/10.1017/eds.2023.30 Published online by Cambridge University Press

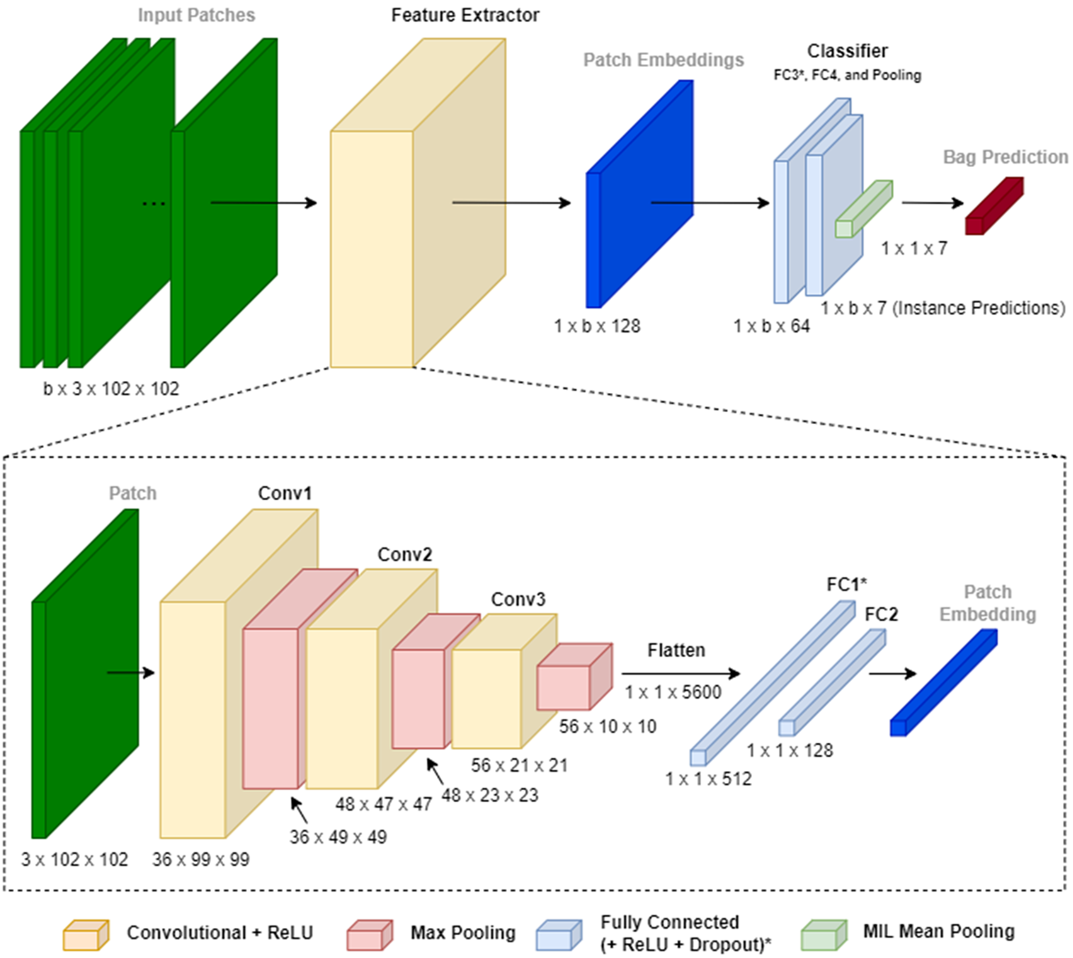

Figure 2. *S2P single-resolution model architecture. Note that some fully connected (FC) layers use ReLU and Dropout (denoted with *) while some do not, and b denotes bag size (number of patches).*

resolutions (when the entire house fits in a single patch), but higher resolutions might be required to separate similar classes, such as the difference between trees and grass. Furthermore, a multi-resolution approach incorporates some notion of spatial relationships, helping to reduce the noise in high-resolution predictions (e.g., when a patch prediction is different from its neighbors). Combining multiple resolutions into a single model facilitates stronger performance and multi-resolution prediction (which aids interpretability) without the need to train separate models for different resolutions.

Several existing works have investigated multi-resolution MIL models, which we use as inspiration for our multi-resolution S2P models. Most notably, we base our approach on Multi-scale Domain-adversarial MIL (MS-DA-MIL; Hashimoto et al., 2020) and Dual Stream MIL (DSMIL; Li et al., 2021). MS-DA-MIL uses a two-stage approach, where the latter stage operates at multiple resolutions and makes a combined overall prediction. DSMIL also uses a two-stage approach and provides a unique approach to combining feature embeddings extracted at different resolutions. Our novel method differs from these approaches in that it only uses a single stage of training. We extract patches at different resolutions and transform these patches into embeddings at each resolution using separate feature extractors (i.e., distinct convolutional layers for each resolution). This allows better extraction of features specific to each resolution (in comparison to using a single feature extractor for all resolutions, as done by Marini et al., 2021). We propose two different approaches for classification:

1. **Multi-resolution single-out (MRSO)**—This model uses multiple resolutions as input, but only makes predictions at the highest resolution. However, this high-resolution output uses information from all the input resolutions in its decision-making process.
2. **Multi-resolution multi-out (MRMO)**—This model uses multiple-resolution inputs and makes predictions at multiple resolutions. Given s_n input resolutions, it makes $s_n + 1$ sets of predictions: one independent set for each input resolution, and one main set, utilizing information from all input resolutions (as in the MRSO model).

We give an example of the MRSO/MRMO models in Figure 3. Here, we utilize $s_n = 3$ input resolutions: $s = 0$, $s = 1$, and $s = 2$, where each resolution is twice that of the previous. For MRMO, the individual resolution predictions are made independently (i.e., not sharing information between the different resolutions), but the main prediction, $s = m$, utilizes information from all input resolutions to make predictions at the $s = 2$ scale. During MRMO training, we optimize the model to minimize the root mean square error (RMSE) averaged over all four outputs, aiming to achieve strong performance at each independent resolution as well as the combined resolution. MRSO does not make independent predictions at each scale, only the main $s = m$ predictions, meaning it can be trained using standard RMSE. In the next section, we discuss how we combine information across different resolutions for MRSO and MRMO.

3.4. Multi-resolution MIL concatenation

In our multi-resolution models MRSO and MRMO, we combine information between different input resolutions to make more accurate predictions. To do so, we use an approach similar to Li et al. (2021). A fundamental part of our approach is how patches are extracted at different resolutions: we alter the grid size, doubling it at each increasing resolution. This means a single patch of size p x p px at resolution $s = 0$ is represented with four patches at resolution $s = 1$, each also of size p x p px. As such, the same area is represented at twice the original resolution, that is, $2p$ x $2p$ px at $s = 1$ compared to p x p px at $s = 0$. Extending this to $s = 2$ follows the same process, resulting in 16 patches for each patch at $s = 0$, that is, at four times the original resolution ($4p$ x $4p$ px). This process is visually represented on the left of Figure 4. As such, given a bag of size b at $s = 0$, the equivalent bags at $s = 1$ and $s = 2$ are of size $4b$ and $16b$ respectively, as shown in Figure 3.

Given this multi-resolution patch extraction process, we then need to combine information across the different resolutions. We do so by using independent feature extraction modules (one per resolution, see Figure 3), which give a set of patch embeddings, one per patch, for each resolution. Therefore, we will have b embeddings at scale $s = 0$, $4b$ embeddings at $s = 1$, and $16b$ embeddings at $s = 2$. To combine these embeddings, we repeat the $s = 0$ and $s = 1$ embeddings to match the number of embeddings at $s = 2$ (16 repeats per embedding for $s = 0$ and 4 repeats per embedding for $s = 1$). This repetition preserves spatial relationships, such that the embeddings for lower resolutions ($s = 0$ and $s = 1$) cover the same image regions as captured by the higher resolution $s = 2$ embeddings. Given the repeated embeddings, it is now possible to concatenate the embeddings from different resolutions, resulting in an extended embedding for each $s = 2$ embedding that includes information from $s = 0$ and $s = 1$. Note the embeddings extracted at each resolution are the same size ($l = 128$), so the concatenated embeddings are three times as long ($3l = 384$). This process is summarized in Figure 4.

4. Experiments

In this section, we detail our experiments, first describing the datasets (Section 4.1) and models (Section 4.2). We then provide our results (Section 4.3).

https://doi.org/10.1017/eds.2023.30 Published online by Cambridge University Press

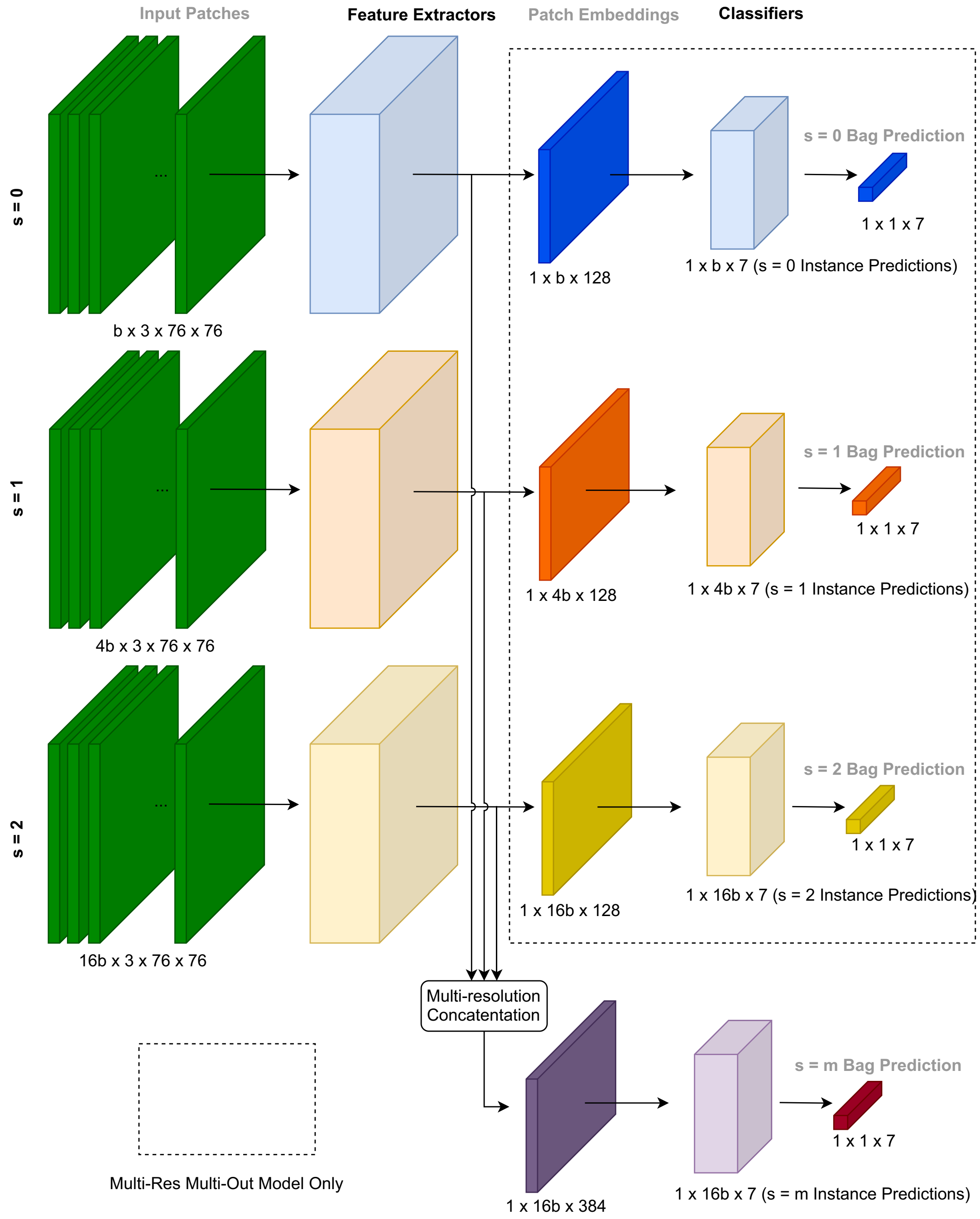

Figure 3. *S2P multi-resolution architecture. The embedding process uses independent feature extraction modules (CNN layers, see Figure 2), allowing specialized feature extraction for each resolution. The MRMO configuration produces predictions at $s=0$, $s=1$, $s=2$, and $s=m$ resolutions (indicated by the dashed box); MRSO only produces $s=m$ predictions.*

4.1. Datasets

We conduct experiments on two datasets: DeepGlobe (Demir et al., 2018) and FloodNet (Rahnemoonfar et al., 2021). Both datasets are described in detail below.

https://doi.org/10.1017/eds.2023.30 Published online by Cambridge University Press

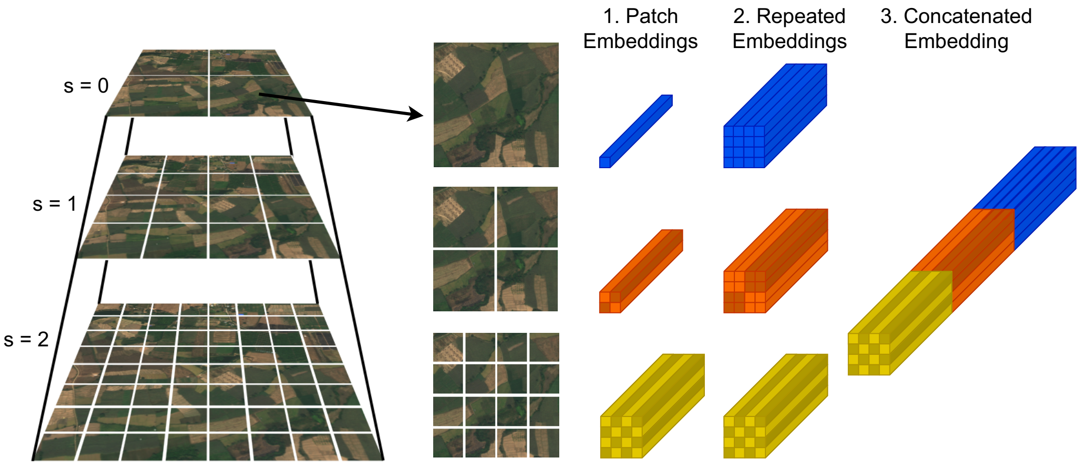

Figure 4. *Multi-resolution patch extraction and concatenation. Left: Patches are extracted at different resolutions, where each patch at scale $s = 0$ has four corresponding $s = 1$ patches and 16 corresponding $s = 2$ patches. Right: The $s = 0$ and $s = 1$ embeddings are repeated to match the number of $s = 2$ embeddings, and then concatenated to create multi-resolution embeddings.*

DeepGlobe. Following our prior work (Early et al., 2022b), we use the DeepGlobe-LCC dataset (Demir et al., 2018). It consists of 803 satellite images with three channels: red, green, and blue (RGB). Each image is 2448 × 2448 pixels with a 50 cm pixel resolution. All images were sourced from the WorldView3 satellite covering regions in Thailand, Indonesia, and India. Each image has a corresponding annotation mask, separating the images into 7 different land cover classes such as urban, agricultural, and forest. There are also validation and test splits with 171 and 172 images respectively, but as these splits did not include annotation masks, they were not used in this work.

FloodNet. To assess our S2P approach on EO data captured with aerial imagery, we use the FloodNet dataset (Rahnemoonfar et al., 2021). This dataset consists of 2343 high-resolution (4000 × 3000 px) RGB images of Ford Bend County in Texas, captured between August 30 and September 4, 2017 after Hurricane Harvey. Images were taken with DJI Mavic Pro quadcopters at 200 feet above ground level (flown by emergency responders as part of the disaster response process). The dataset aims to capture post-disaster flooding. Each image is segmented into 10 different classes, including flood-specific annotations such as flooded/non-flooded buildings, and flooded/non-flooded roads.

While both the DeepGlobe and FloodNet datasets provide pixel-level annotations, these segmentation labels are only used to generate the regression targets for training and for evaluating the predicted patch segmentation, that is, they are not directly used during training. However, we would like to stress that these segmentation labels are not strictly required for our approach, that is, the scene-level regression targets can be created without having to perform segmentation.

For the DeepGlobe dataset, due to its limited size (only 803 images), we used 5-fold cross-validation rather than the standard 10-fold approach. With this configuration, each fold had an 80/10/10 split for train/validation/test. The FloodNet dataset comes with pre-defined dataset splits: 1445 ($\sim 61.7\%$) train, 450 ($\sim 19.2\%$) validation, and 448 ($\sim 19.1\%$) test, which were used consistently in five training repeats (i.e., no cross-validation). For more details on both datasets, see Supplementary Appendix B.

4.2. Model configurations

To apply MIL at different resolutions, we alter the grid size applied over the image (i.e., the number of extracted patches). For the DeepGlobe dataset (square images), we used three grid sizes: 8 × 8, 16 × 16, and 32 × 32. For the FloodNet dataset (non-square images), we used 8 × 6, 16 × 12, and 32 × 24. These grid sizes are chosen such that they can be used directly for three different resolutions in the multi-

https://doi.org/10.1017/eds.2023.30 Published online by Cambridge University Press

resolution models, for example, 32 × 32 ($s = 2$) is twice the resolution of 16 × 16 ($s = 1$), which is twice the resolution of 8 × 8 ($s = 0$).

We also compare our MIL S2P models to a fine-tuned ResNet18 model (He et al., 2016), and two U-Net variations operating on different image sizes (224 × 224 px and 448 × 448 px). These baseline models are trained in the same manner as the S2P models, that is, using scene-level regression. Although the ResNet model does not produce patch- or pixel-level predictions, we use it as a scene-level baseline as many existing LCC approaches utilize ResNet architectures (Hoeser et al., 2020; Hoeser and Kuenzer, 2020; Rahnemoonfar et al., 2021). For the U-Net models, we follow the same procedure as Wang et al. (2020) and use class activation maps to recover segmentation outputs. This means U-Net is a stronger baseline than ResNet as it can be used for both scene- and pixel-level prediction. For more details on the implementation, models, and training process, see Supplementary Appendices A and C.

4.3. Results

We evaluate performance on both datasets using four metrics, covering scene-, patch-, and pixel-level prediction. Scene-level performance is scored using root mean square error (RMSE) and mean absolute error (MAE), where lower values are better. Both of these metrics compare the scene-level (bag) predictions with the true scene-level coverage labels. For patch-level predictions, we report the patch-level mean Intersection over Union (mIoU; Everingham et al., 2010; Minaee et al., 2021), where larger values are better.[1] Patch labels (only used during evaluation; not during training) are derived from the true segmentation masks, where the labels are the class that has maximum coverage in each patch. For pixel-level prediction, we compute pixel-level mIoU using the original ground-truth segmentation masks.[2] For S2P models, this is achieved by resizing the patch-level segmentation output to the same size as the ground-truth segmentation mask. Pixel-level mIoU is the primary metric for evaluation as it best determines the ability of the models to segment the input images—patch-level evaluation is dependent on the grid size so is not a consistent metric across different resolutions (but can be useful to compare patch-level predictions at the same resolution). Strong models should perform well at both scene- and pixel-level prediction, that is, low scene RMSE, low scene MAE, and high pixel mIoU.

Our results are given in Tables 1 and 2 for DeepGlobe and FloodNet respectively. We compare the three non-MIL baselines[3] (ResNet18, U-Net 224, and U-Net 448) with S2P single resolution (SR) models operating at three different scales ($s = 0$, $s = 1$, and $s = 2$), the S2P MRSO approach, and the S2P MRMO approach (evaluating the MRMO outputs at all scales). Our first observation is that, for both datasets, all of our S2P models outperform the baseline approaches at both scene- and pixel-level prediction, showing the efficacy of our S2P approach. We also find that, out of all the methods, the MRMO approach achieves the best performance when using its combined $s = m$ output, demonstrating the performance gain of using multiple resolutions. Note the MRMO $s = m$ performance is better than the MRSO $s = m$ performance, suggesting that optimizing the model for independent scale predictions ($s = 0$, $s = 1$, and $s = 2$) as well as the combined output ($s = m$) is beneficial for learning—the multi-output nature potentially leads to better representation learning at each individual scale and could be helpful in avoiding overfitting. However, the MRSO model still outperforms its single-resolution equivalent (S2P SR $s = 2$), again demonstrating the efficacy of using multiple resolutions.

To compare the change in performance caused by using different/multiple resolutions, we visualize our S2P results in Figure 5. For the DeepGlobe dataset, when using single-resolution models, we observe that increasing the resolution leads to worse performance (greater Scene RMSE and MAE, and lower Pixel mIoU). In our prior work, we identified this as a trade-off between performance and segmentation resolution. However, with the extension to multi-resolution models, it is possible to achieve strong

[1] Patch-level evaluation is only applicable for our S2P models as the ResNet18 and U-Net approaches do not use patches.

[2] Pixel-level evaluation is not possible for the ResNet18 baseline as it does not produce any segmentation output.

[3] The original works reported baseline pixel mIoU scores of 0.43 for DeepGlobe and 0.43 to 0.80 for FloodNet. However, those baselines are trained against segmentation maps rather than coverage labels, so are expected to perform better.

https://doi.org/10.1017/eds.2023.30 Published online by Cambridge University Press

Table 1. *DeepGlobe results*

Model	Scene RMSE	Scene MAE	Patch mIoU	Pixel mIoU
ResNet18	0.218 ± 0.008	0.128 ± 0.004	N/A	N/A
U-Net 224	0.136 ± 0.008	0.075 ± 0.004	N/A	0.245 ± 0.008
U-Net 448	0.134 ± 0.008	0.076 ± 0.004	N/A	0.272 ± 0.008
S2P SR $s=0$	0.090 ± 0.005	0.047 ± 0.002	**0.439** ± **0.014**	0.397 ± 0.014
S2P SR $s=1$	0.097 ± 0.003	0.051 ± 0.001	0.404 ± 0.016	0.384 ± 0.014
S2P SR $s=2$	0.104 ± 0.008	0.055 ± 0.004	0.353 ± 0.018	0.345 ± 0.018
S2P MRSO $s=m$	0.093 ± 0.004	0.051 ± 0.002	0.394 ± 0.014	0.388 ± 0.014
S2P MRMO $s=0$	0.087 ± 0.003	0.048 ± 0.001	0.432 ± 0.012	0.389 ± 0.013
S2P MRMO $s=1$	0.087 ± 0.004	0.046 ± 0.002	0.412 ± 0.010	0.391 ± 0.010
S2P MRMO $s=2$	0.095 ± 0.004	0.050 ± 0.002	0.370 ± 0.011	0.362 ± 0.011
S2P MRMO $s=m$	**0.084** ± **0.004**	**0.045** ± **0.002**	0.425 ± 0.013	**0.417** ± **0.013**

Note: We give the mean performance averaged over five repeat training runs, with bounds using standard error of the mean. Best performance for each metric is indicated in bold.

Table 2. *FloodNet results*

Model	Scene RMSE	Scene MAE	Patch mIoU	Pixel mIoU
ResNet18	0.145 ± 0.001	0.077 ± 0.002	N/A	N/A
U-Net 224	0.083 ± 0.001	0.039 ± 0.001	N/A	0.193 ± 0.003
U-Net 448	0.080 ± 0.002	0.037 ± 0.002	N/A	0.206 ± 0.004
S2P SR $s=0$	0.070 ± 0.001	**0.028** ± **0.000**	0.273 ± 0.002	0.233 ± 0.002
S2P SR $s=1$	0.073 ± 0.000	0.030 ± 0.001	0.269 ± 0.002	0.248 ± 0.002
S2P SR $s=2$	0.072 ± 0.001	0.030 ± 0.001	0.263 ± 0.003	0.251 ± 0.003
S2P MRSO $s=m$	0.070 ± 0.001	0.029 ± 0.000	0.273 ± 0.004	0.264 ± 0.004
S2P MRMO $s=0$	0.070 ± 0.001	0.029 ± 0.000	0.273 ± 0.003	0.234 ± 0.002
S2P MRMO $s=1$	0.070 ± 0.001	0.029 ± 0.000	0.277 ± 0.002	0.256 ± 0.001
S2P MRMO $s=2$	0.070 ± 0.000	0.030 ± 0.000	0.272 ± 0.001	0.260 ± 0.001
S2P MRMO $s=m$	**0.069** ± **0.001**	**0.028** ± **0.000**	**0.279** ± **0.002**	**0.271** ± **0.002**

Note: We give the mean performance averaged over five repeat training runs, with bounds using standard error of the mean. Best performance for each metric is indicated in bold.

performance while operating at higher resolutions, overcoming this trade-off. For the FloodNet dataset, we observe the opposite trend for pixel mIoU—increasing resolution leads to better segmentation performance, even for the single resolution models. This is likely explained by the different spatial modalities of the two datasets: DeepGlobe images cover much larger spatial areas (e.g., entire towns and fields in a single image) compared to FloodNet (e.g., a handful of buildings or a single road per image). As such, the FloodNet segmentation masks identify individual objects (e.g., a single tree), which require high-resolution inputs and outputs to correctly separate (as opposed to broadly segmenting a forest region as in DeepGlobe). However, we still observe that utilizing multiple resolutions gives better performance for FloodNet. We also find that the MRMO independent resolution predictions ($s=0$, $s=1$, and $s=2$) typically outperform the equivalent single-resolution model predictions.

5. Discussion

In this section, we further discuss our findings. First, we analyze the model predictions in more detail (Section 5.1), then investigate the interpretability and segmentation outputs of the models (Section 5.2).

https://doi.org/10.1017/eds.2023.30 Published online by Cambridge University Press

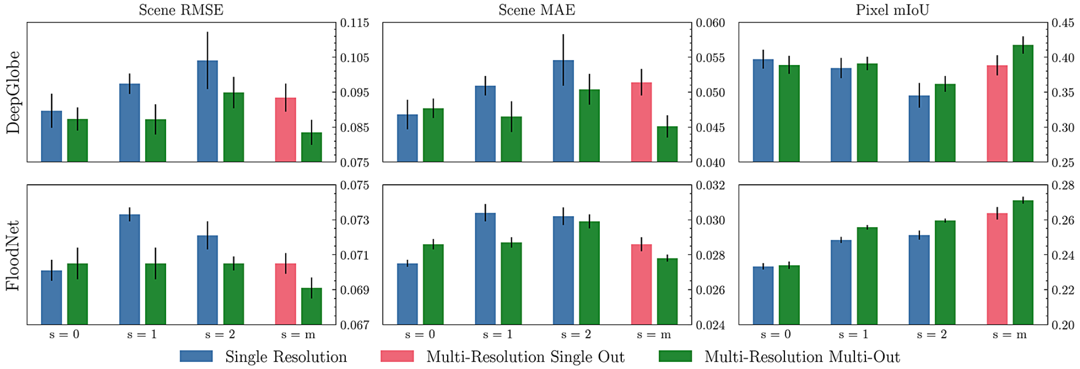

Figure 5. *S2P resolution comparison. We compare model performance for scene RMSE (left), scene MAE (middle), and pixel mIoU (right) for both DeepGlobe (top) and FloodNet (bottom).*

Next, we conduct an ablation study into model architectures (Section 5.3). Finally, we discuss the limitations of this study and outline areas for future work (Section 5.4).

5.1. *Model analysis*

We conduct an additional study to further analyze and compare the effectiveness of the different approaches. We use confusion matrices along with precision and recall analysis to investigate the performance of the models on particular classes. The results for the DeepGlobe dataset are presented in Figure 6. We observe that all of the analyzed models perform best on the agricultural, urban, and forest classes, but struggle on rangeland and unknown classes. For the MRMO $s = m$ model (the best-performing approach) rangeland is most often predicted as agricultural, and unknown is most often predicted as water or agricultural. Rangeland is defined as any green land or grass that is not forest or farmland, so is often difficult to separate from agricultural land. The unknown class is very

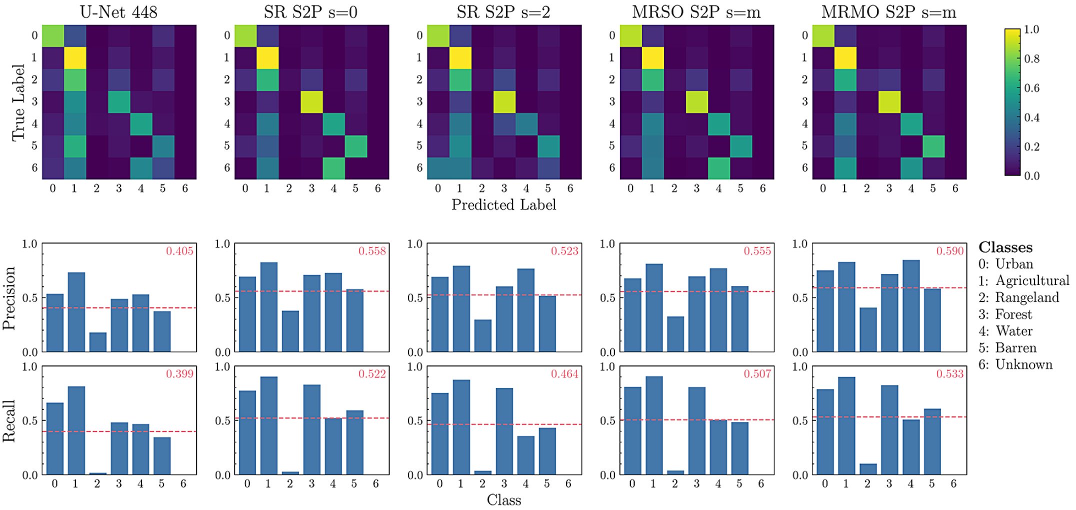

Figure 6. *DeepGlobe model analysis. Top: Row normalized confusion matrices. Bottom: Classwise precision and recall, where the dotted line indicates the macro-averaged performance (which is also denoted in the top right corner of each plot).*

https://doi.org/10.1017/eds.2023.30 Published online by Cambridge University Press

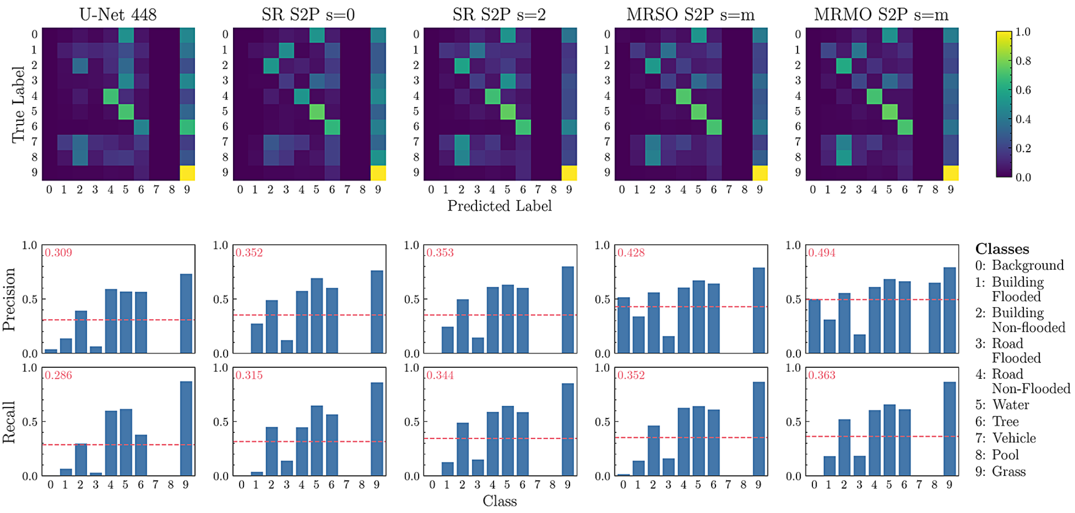

Figure 7. *FloodNet model analysis. Top: Row normalized confusion matrices. Bottom: Classwise precision and recall, where the dotted line indicates the macro-averaged performance (which is also denoted in the top left corner of each plot).*

underrepresented in the dataset compared to the other classes (see Supplementary Appendix B.1). We note that the MRMO $s = m$ model also achieves the best macro-averaged precision and recall.

In a similar study for FloodNet (Figure 7), the models perform best at identifying the road non-flooded, water, tree, and grass classes, but struggle on the background, vehicle, and pool classes. The vehicle and pool classes represent relatively small objects (compared to buildings and roads) and are also very underrepresented (see Supplementary Appendix B.2), making them hard to identify. However, they are most often classified as building non-flooded, which is an appropriate alternative prediction (i.e., better than classifying them as natural objects such as trees). The background category is a catch-all for any regions that do not belong to the other classes, but it often contains elements of the other classes, making it difficult to segment. Again, we find that the MRMO $s = m$ model has the best precision and recall performance.

5.2. *Model interpretability*

As our S2P models inherently produce patch-level predictions, we can directly interpret their segmentation outputs. In Figure 8, we do so for the MRMO model on the FloodNet dataset. We observe that the model is able to accurately separate the non-flooded building and road, and also identifies the grass and trees. Furthermore, we find the $s = m$ predictions are an improvement over the independent predictions, notably that the $s = m$ output is less noisy than the $s = 2$ predictions. We also note that the model is also to support and refute different classes, which further aids interpretability.

5.3. *Ablation study: Single resolution S2P configurations*

In our prior work (Early et al., 2022b), we experimented with changing the patch size: the dimensions to which each cell from the grid extraction process is resized. This influences the effective resolution (i.e., the level of downsampling) at which the S2P models are operating, and also impacts the MIL model architectures. We continue this analysis here, applying it to both DeepGlobe and FloodNet. We conducted an ablation study using three different patch sizes (small, medium, and large). This was done for each of three resolutions $s = 0$, $s = 1$, and $s = 2$, resulting in nine different S2P single-resolution configurations (for further details, see Supplementary Appendix C.1). As shown in Figure 9, we find that the large model

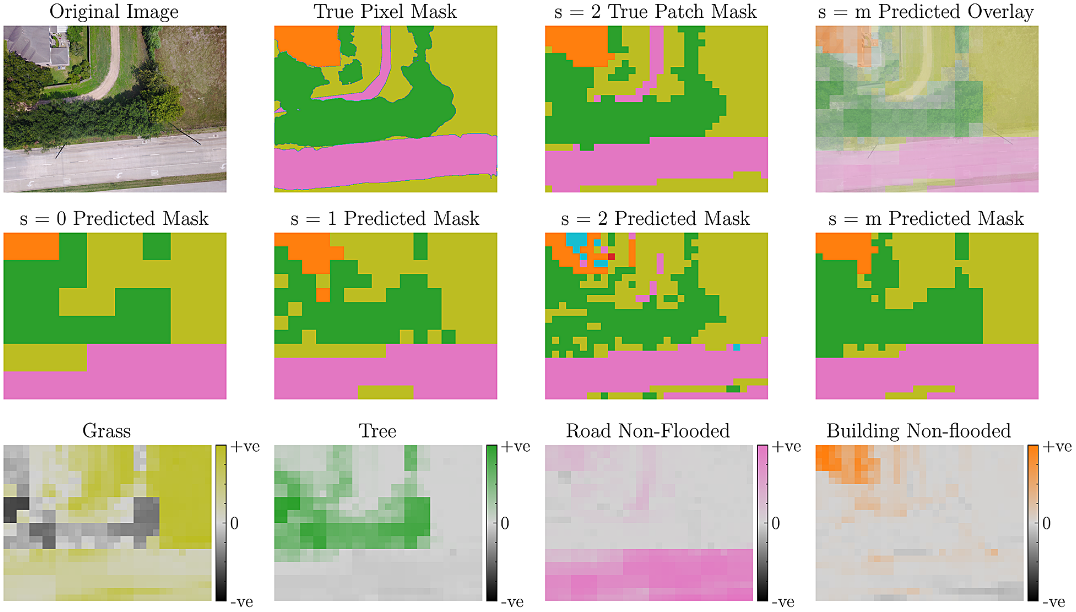

Figure 8. *MRMO interpretability example. Top (from left to right): the original dataset image; the true pixel-level mask; the true patch-level mask at resolution $s = 2$; the predicted mask from the MRMO model overlaid on the original image. Middle: Predicted masks for each of the MRMO outputs. Bottom: MRMO $s = m$ predicted masks for each class, showing supporting (+ve) and refuting regions (−ve).*

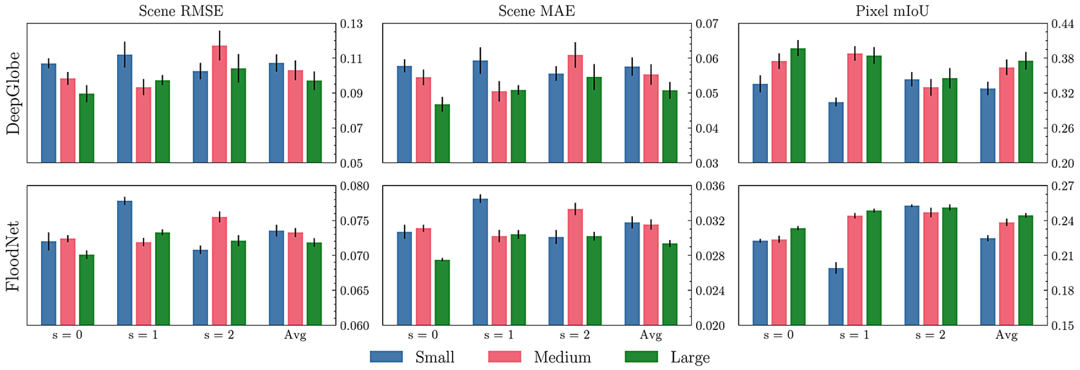

Figure 9. *Single-resolution S2P ablation study. We observe that, on average, the large configuration achieves the best performance (lowest RMSE, lowest MAE, and highest mIoU) on both datasets. Error bars are given representing the standard error of the mean from five repeats.*

configuration has the best average performance on all metrics for both datasets. This matches our intuition that operating at higher resolutions leads to better performance. Given these results, we used the large configuration as the backbone of the multi-resolution S2P models, and as the choice of single-resolution model in our main results (Tables 1 and 2).

5.4. Limitations and future work

While our new multi-resolution S2P approach outperforms existing methods, there are still limitations to the method. For the DeepGlobe dataset, the model struggles to separate rangeland and agricultural

https://doi.org/10.1017/eds.2023.30 Published online by Cambridge University Press

regions, and for the FloodNet dataset, it struggles to correctly separate flooded regions and identify smaller objects such as pools and vehicles. These problems mostly stem from dataset imbalance; future work could investigate dataset balancing or augmentation to overcome these class imbalances. It could also be beneficial to extend the MIL models to incorporate additional spatial information—this could help separate small objects from the larger objects they often appear next to, for example, swimming pools from houses. Potential solutions include utilizing positional embeddings (Vaswani et al., 2017), MIL attention (Ilse et al., 2018), or graph neural networks (Tu et al., 2019).

When considering the application of our work to other datasets, there are two points to consider. First, while our solution will scale to datasets containing more images of a similar size to the ones studied in this work, there could be resource issues when using datasets containing larger images (i.e., images larger than 4000 × 3000 px). The computational resources required are determined by the image size, the model's architecture, and the effective resolution at which the model is operating (see Supplementary Appendix C). Reducing the model complexity or lowering the image resolution would overcome potential issues. However, we did not run into any such issues in our work (see Supplementary Appendix A for details on our compute resources). Second, in this work, we used datasets with existing segmentation labels that are converted into scene-level coverage labels. An area for future work is to investigate how to generate coverage labels for datasets without segmentation labels. This could also incorporate an analysis of model performance in the presence of label noise, that is, imperfect coverage labels.

Finally, we only tuned training hyperparameters on the DeepGlobe dataset (see Supplementary Appendix C.2 for more details). This was an intentional choice, as it allowed us to demonstrate that the hyperparameters found through tuning were robust (i.e., useful default parameters for training on other datasets). However, potentially better performance could be achieved on FloodNet by tuning the training hyperparameters specifically for that dataset and optimizing the model architectures as well as the training hyperparameters.

6. Conclusion

In this work, we presented a novel method for segmenting EO RS images. Our method extends our previous single-resolution S2P approach to multiple resolutions, which leads to improved performance and interpretability. We demonstrated the efficacy of our approach on two datasets: DeepGlobe (LCC from satellite imagery) and FloodNet (NDR from aerial imagery). As our S2P approaches do not require segmentation labels, faster and easier curation of larger and more diverse datasets can be achieved, facilitating the wider use of ML for climate change mitigation in technology, government, and academia.

Acknowledgements. We would like to thank the University of Southampton and the Alan Turing Institute for their support. The authors acknowledge the use of the University of Southampton IRIDIS high-performance computing facilities and associated support services. Finally, this work was supported by the Climate Change AI (CCAI) Mentorship Program.

Author contribution. Conceptualization: J.E.; Methodology: J.E.; Implementation: J.E.; Data visualization: J.E.; Writing (original draft): J.E.; Writing (editing and review): J.E. and Y.-J. C.D.; Supervision: Y.-J. C.D., C.E., and S.R.

Competing interest. The authors declare no competing interests exist.

Data availability statement. Code and trained models are available at https://github.com/JAEarly/MIL-Multires-EO (for more details, see Supplementary Appendix A). All data are openly available: DeepGlobe at https://www.kaggle.com/datasets/balraj98/deepglobe-land-cover-classification-dataset and FloodNet at https://github.com/BinaLab/FloodNet-Supervised_v1.0 (for more details, see Supplementary Appendix B).

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This work was funded by the AXA Research Fund and the UKRI TAS Hub (EP/V00784X/1). Both funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2023.30.

https://doi.org/10.1017/eds.2023.30 Published online by Cambridge University Press

References

Ban Y, **Zhang P**, **Nascetti A**, **Bevington AR and Wulder MA** (2020) Near real-time wildfire progression monitoring with Sentinel-1 SAR time series and deep learning. *Scientific Reports 10*(1). https://doi.org/10.1038/s41598-019-56967-x

Burrows K, **Walters RJ**, **Milledge D**, **Spaans K and Densmore AL** (2019) A new method for large-scale landslide classification from satellite radar. *Remote Sensing 11*(3), 237. https://doi.org/10.3390/rs11030237

Caballero I, **Ruiz J and Navarro G** (2019) Sentinel-2 satellites provide near-real time evaluation of catastrophic floods in the west mediterranean. *Water 11*(12), 2499. https://doi.org/10.3390/w11122499

Carbonneau M-A, **Cheplygina V**, **Granger E and Gagnon G** (2018) Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition 77*, 329–353. https://doi.org/10.1016/j.patcog.2017.10.009

CCAI (2022) Climate Change AI Dataset Wishlist. Available at https://www.climatechange.ai/dataset-wishlist.pdf (accessed 15 Feb 2023).

Demir I, **Koperski K**, **Lindenbaum D**, **Pang G**, **Huang J**, **Basu S**, **Hughes F**, **Tuia D and Raskar R** (2018) DeepGlobe 2018: A challenge to parse the Earth through satellite images. In *2018 IEEE/CVF Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://doi.org/10.1109/cvprw.2018.00031

Early J, **Bewley T**, **Evers C and Ramchurn S** (2022a) Non-Markovian reward modelling from trajectory labels via interpretable multiple instance learning. *Advances in Neural Information Processing Systems 35*, 27652–27663.

Early J, **Deweese Y-J**, **Evers C and Ramchurn S** (2022b) Scene-to-Patch Earth observation: Multiple instance learning for land cover classification. In *NeurIPS Workshop: Tackling Climate Change with Machine Learning*.

Early J, **Evers C and Ramchurn S** (2022c) Model agnostic interpretability for multiple instance learning. In *International Conference on Learning Representations*.

Elsen PR, **Saxon EC**, **Simmons BA**, **Ward M**, **Williams BA**, **Grantham HS**, **Kark S**, **Levin N**, **Perez-Hammerle K-V**, **Reside AE and Watson JEM** (2021) Accelerated shifts in terrestrial life zones under rapid climate change. *Global Change Biology 28* (3), 918–935. https://doi.org/10.1111/gcb.15962E

Everingham M, **Gool LV**, **Williams CKI**, **Winn J and Zisserman A** (2009) The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision 88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Fan C, **Zhang C**, **Yahja A and Mostafavi A** (2021) Disaster city digital twin: A vision for integrating artificial and human intelligence for disaster management. *International Journal of Information Management 56*, 102049. https://doi.org/10.1016/j.ijinfomgt.2019.102049

Friedlingstein P, **O'Sullivan M**, **Jones MW**, **Andrew RM**, **Hauck J**, **Olsen A**, **Peters GP**, **Peters W**, **Pongratz J**, **Sitch S**, **Quéré CL**, **Canadell JG**, **Ciais P**, **Jackson RB**, **Alin S**, **Aragão LEOC**, **Arneth A**, **Arora V**, **Bates NR**, **Becker M**, **Benoit-Cattin A**, **Bittig HC**, **Bopp L**, **Bultan S**, **Chandra N**, **Chevallier F**, **Chini LP**, **Evans W**, **Florentie L**, **Forster PM**, **Gasser T**, **Gehlen M**, **Gilfillan D**, **Gkritzalis T**, **Gregor L**, **Gruber N**, **Harris I**, **Hartung K**, **Haverd V**, **Houghton RA**, **Ilyina T**, **Jain AK**, **Joetzjer E**, **Kadono K**, **Kato E**, **Kitidis V**, **Korsbakken JI**, **Landschützer P**, **Lefèvre N**, **Lenton A**, **Lienert S**, **Liu Z**, **Lombardozzi D**, **Marland G**, **Metzl N**, **Munro DR**, **Nabel JEMS**, **Nakaoka S-I**, **Niwa Y**, **O'Brien K**, **Ono T**, **Palmer PI**, **Pierrot D**, **Poulter B**, **Resplandy L**, **Robertson E**, **Rödenbeck C**, **Schwinger J**, **Séférian R**, **Skjelvan I**, **Smith AJP**, **Sutton AJ**, **Tanhua T**, **Tans PP**, **Tian H**, **Tilbrook B**, **van der Werf G**, **Vuichard N**, **Walker AP**, **Wanninkhof R**, **Watson AJ**, **Willis D**, **Wiltshire AJ**, **Yuan W**, **Yue X and Zaehle S** (2020) Global carbon budget 2020. *Earth System Science Data 12*(4), 3269–3340. https://doi.org/10.5194/essd-12-3269-2020

Hashimoto N, **Fukushima D**, **Koga R**, **Takagi Y**, **Ko K**, **Kohno K**, **Nakaguro M**, **Nakamura S**, **Hontani H and Takeuchi I** (2020) Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In *2020 IEEE/CVF Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.00391

He K, **Zhang X**, **Ren S and Sun J** (2016) Deep residual learning for image recognition. In *2016 IEEE Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.90

Hoeser T, **Bachofer F and Kuenzer C** (2020) Object detection and image segmentation with deep learning on Earth observation data: A review—Part II: Applications. *Remote Sensing 12*(18), 3053. https://doi.org/10.3390/rs12183053

Hoeser T and Kuenzer C (2020) Object detection and image segmentation with deep learning on Earth observation data: A review —Part I: Evolution and recent trends. *Remote Sensing 12*(10), 1667. https://doi.org/10.3390/rs12101667

Ilse M, **Tomczak J and Welling M** (2018) Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pp. 2127–2136.

Kaack LH, **Donti PL**, **Strubell E**, **Kamiya G**, **Creutzig F and Rolnick D** (2022) Aligning artificial intelligence with climate change mitigation. *Nature Climate Change 12*, 518–527. https://doi.org/10.1038/s41558-022-01377-7

Karra K, **Kontgis C**, **Statman-Weil Z**, **Mazzariello JC**, **Mathis M and Brumby SP** (2021) Global land use/land cover with Sentinel 2 and deep learning. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. https://doi.org/10.1109/igarss47720.2021.9553499

Khryashchev V and Larionov R (2020) Wildfire segmentation on satellite images using deep learning. In *2020 Moscow Workshop on Electronic and Networking Technologies (MWENT)*. https://doi.org/10.1109/mwent47943.2020.9067475

Kuo T-S, **Tseng K-S**, **Yan J-W**, **Liu Y-C and Frank Wang Y-CF** (2018) Deep aggregation net for land cover classification. In *2018 IEEE/CVF Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://doi.org/10.1109/cvprw.2018.00046

https://doi.org/10.1017/eds.2023.30 Published online by Cambridge University Press

Lee J, **Brooks NR**, **Tajwar F**, **Burke M**, **Ermon S**, **Lobell DB**, **Biswas D and Luby SP** (2021) Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences 118*(17), https://doi.org/10.1073/pnas.2018863118

Li B, **Li Y and Eliceiri KW** (2021) Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *2021 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr46437.2021.01409

Li X, **Shen L**, **Xie X**, **Huang S**, **Xie Z**, **Hong X and Yu J** (2020) Multi-resolution convolutional networks for chest x-ray radiograph based lung nodule detection. *Artificial Intelligence in Medicine 103*, 101744. https://doi.org/10.1016/j.artmed.2019.101744

Lin T-Y, **Dollar P**, **Girshick R**, **He K**, **Hariharan B and Belongie S** (2017) Feature pyramid networks for object detection. In *2017 IEEE Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2017.106

Liu X, **Jiao L**, **Zhao J**, **Zhao J**, **Zhang D**, **Liu F**, **Yang S and Tang X** (2018) Deep multiple instance learning-based spatial–spectral classification for PAN and MS imagery. *IEEE Transactions on Geoscience and Remote Sensing 56*(1), 461–473. https://doi.org/10.1109/tgrs.2017.2750220

Long J, **Shelhamer E and Darrell T** (2015) Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2015.7298965

Marini N, **Otálora S**, **Ciompi F**, **Silvello G**, **Marchesin S**, **Vatrano S**, **Buttafuoco G**, **Atzori M and Müller H** (2021) Multi-scale task multiple instance learning for the classification of digital pathology images with global annotations. In *MICCAI Workshop on Computational Pathology*, pp. 170–181.

Minaee S, **Boykov YY**, **Porikli F**, **Plaza AJ**, **Kehtarnavaz N and Terzopoulos D** (2021) Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/tpami.2021.3059968

Munawar HS, **Ullah F**, **Qayyum S**, **Khan SI and Mojtahedi M** (2021) UAVs in disaster management: Application of integrated aerial imagery and convolutional neural network for flood detection. *Sustainability 13*(14), 7547. https://doi.org/10.3390/su13147547

Oddo PC and Bolten JD (2019) The value of near real-time Earth observations for improved flood disaster response. *Frontiers in Environmental Science 7*. https://doi.org/10.3389/fenvs.2019.00127

Pi Y, **Nath ND and Behzadan AH** (2020) Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics 43*, 101009. https://doi.org/10.1016/j.aei.2019.101009

Rahnemoonfar M, **Chowdhury T**, **Sarkar A**, **Varshney D**, **Yari M and Murphy RR** (2021) FloodNet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access 9*, 89644–89654. https://doi.org/10.1109/access.2021.3090981

Rakhlin A, **Davydow A and Nikolenko S** (2018) Land cover classification from satellite imagery with U-Net and Lovász Softmax loss. In *2018 IEEE/CVF Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://doi.org/10.1109/cvprw.2018.00048

Robinson C, **Hou L**, **Malkin K**, **Soobitsky R**, **Czawlytko J**, **Dilkina B and Jojic N** (2019) Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2019.01301

Rolnick D, **Donti PL**, **Kaack LH**, **Kochanski K**, **Lacoste A**, **Sankaran K**, **Ross AS**, **Milojevic-Dupont N**, **Jaques N**, **Waldman-Brown A**, **Luccioni AS**, **Maharaj T**, **Sherwin ED**, **Mukkavilli SK**, **Kording KP**, **Gomes CP**, **Ng AY**, **Hassabis D**, **Platt JC and Creutzig F**, **Chayes J and Bengio Y** (2022) Tackling climate change with machine learning. *ACM Computing Surveys (CSUR) 55*(2), 1–96. https://doi.org/10.1145/3485128

Ronneberger O, **Fischer P and Brox T** (2015) U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Rudner TGJ, **Rußwurm M**, **Fil J**, **Pelich R**, **Bischke B**, **Kopačková V and Biliński P** (2019) Multi3Net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. *Proceedings of the AAAI Conference on Artificial Intelligence 33*(1), 702–709. https://doi.org/10.1609/aaai.v33i01.3301702

Sadhukhan J (2022) Net-zero action recommendations for scope 3 emission mitigation using life cycle assessment. *Energies 15* (15), 5522. https://doi.org/10.3390/en15155522

Seferbekov S, **Iglovikov V**, **Buslaev A and Shvets A** (2018) Feature pyramid network for multi-class land segmentation. In *2018 IEEE/CVF Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://doi.org/10.1109/cvprw.2018.00051

Tong X-Y, **Xia G-S**, **Lu Q**, **Shen H**, **Li S**, **You S and Zhang L** (2020) Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment 237*, 111322. https://doi.org/10.1016/j.rse.2019.111322

Tu M, **Huang J**, **He X and Zhou B** (2019) Multiple instance learning with graph neural networks. In *ICML Workshop: Learning and Reasoning with Graph-Structured Representations.*

Ullah F, **Liu J**, **Shafique M**, **Ullah S**, **Rajpar MN**, **Ahmad A and Shahzad M** (2022) Quantifying the influence of Chashma Right Bank Canal on land-use/land-cover and cropping pattern using remote sensing. *Ecological Indicators 143*, 109341. https://doi.org/10.1016/j.ecolind.2022.109341

Vaswani A, **Shazeer N**, **Parmar N**, **Uszkoreit J**, **Jones L**, **Gomez AN**, **Kaiser Ł and Polosukhin I** (2017) Attention is all you need. *Advances in Neural Information Processing Systems 30*, 5998–6008.

Vatsavai RR, **Bhaduri B and Graesser J** (2013) Complex settlement pattern extraction with multi-instance learning. *Joint Urban Remote Sensing Event 2013*. https://doi.org/10.1109/jurse.2013.6550711

Wang K, **Oramas J and Tuytelaars T** (2021) In defense of LSTMs for addressing multiple instance learning problems. In *Proceedings of the Asian Conference on Computer Vision– ACCV 2020*, 444–460. https://doi.org/10.1007/978-3-030-69544-6_27

Wang S, **Chen W**, **Xie SM**, **Azzari G and Lobell DB** (2020) Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sensing 12*(2), 207. https://doi.org/10.3390/rs12020207

Wang X, **Xu H**, **Yuan L**, **Dai W and Wen X** (2022) A remote-sensing scene-image classification method based on deep multiple-instance learning with a residual dense attention ConvNet. *Remote Sensing 14*(20), 5095. https://doi.org/10.3390/rs14205095

Wang X, **Yan Y**, **Tang P**, **Bai X and Liu W** (2018) Revisiting multiple instance neural networks. *Pattern Recognition 74*, 15–24. https://doi.org/10.1016/j.patcog.2017.08.026

Wang Y, **Wang C**, **Zhang H**, **Dong Y and Wei S** (2019) Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery. *Remote Sensing 11*(5), 531. https://doi.org/10.3390/rs11050531

Wang Z, **Lan L and Vucetic S** (2012) Mixture model for multiple instance regression and applications in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing 50*(6), 2226–2237. https://doi.org/10.1109/tgrs.2011.2171691

Zhang C, **Fan C**, **Yao W**, **Hu X and Mostafavi A** (2019) Social media for intelligent public information and warning in disasters: An interdisciplinary review. *International Journal of Information Management 49*, 190–207. https://doi.org/10.1016/j.ijinfomgt.2019.04.004

Zhang M, **Shi W**, **Chen S**, **Zhan Z and Shi Z** (2021) Deep multiple instance learning for landslide mapping. *IEEE Geoscience and Remote Sensing Letters 18*(10), 1711–1715. https://doi.org/10.1109/lgrs.2020.3007183

Cite this article: Early J, Deweese Y.-J.C, Evers C and Ramchurn S (2023). Extending scene-to-patch models: Multi-resolution multiple instance learning for Earth observation. *Environmental Data Science*, 2: e42. doi:10.1017/eds.2023.30