IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 December 2025, accepted 21 January 2026, date of publication 30 January 2026, date of current version 10 February 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3659734

RESEARCH ARTICLE

Core-Set Selection for Data-Efficient Land Cover Segmentation

KEILLER NOGUEIRA[1], AKRAM ZAYTAR[2], WANLI MA[3,4], (Member, IEEE), RIBANA ROSCHER[5], RONNY HÄNSCH[6], (Senior Member, IEEE), CALEB ROBINSON[2], ANTHONY ORTIZ[2], SIMONE FOBI[2], RAHUL DODHIA[2], JUAN M. LAVISTA FERRES[2], OKTAY KARAKUŞ[3], (Member, IEEE), AND PAUL L. ROSIN[3]

[1]University of Liverpool, Liverpool, L69 7ZX England, U.K.
[2]Microsoft AI for Good Research Lab, Redmond, WA 98052, USA
[3]School of Computer Science and Informatics, Cardiff University, CF24 4AG Cardiff, U.K.
[4]University of Cambridge, Cambridge, CB3 0FA England, U.K.
[5]University of Bonn, 53113 Bonn, Germany
[6]Department SAR Technology, German Aerospace Center (DLR), 82234 Wessling, Germany

Corresponding author: Keiller Nogueira (keiller.nogueira@liverpool.ac.uk)

This work was supported in part by Deutsche Forschungsgemeinschaft (DFG), German Research Foundation under Germany's Excellence Strategy under Grant EXC-2070-390732324-PhenoRob.

ABSTRACT The increasing accessibility of remotely sensed data and their potential to support large-scale decision-making have driven the development of deep learning models for many Earth Observation tasks. Traditionally, such models rely on large datasets. However, the common assumption that larger training datasets lead to better performance tends to overlook issues related to data redundancy, noise, and the computational cost of processing massive datasets. Effective solutions must therefore consider not only the quantity but also the quality of data. Towards this, in this paper, we introduce six basic core-set selection approaches – that rely on imagery only, labels only, or a combination of both – and investigate whether they can identify high-quality subsets of data capable of maintaining – or even surpassing – the performance achieved when using full datasets for remote sensing semantic segmentation. We benchmark such approaches against two traditional baselines on three widely used land-cover classification datasets (DFC2022, Vaihingen, and Potsdam) using two different architectures (SegFormer and U-Net), thus establishing a general baseline for future works. Our experiments show that all proposed methods consistently outperform the baselines across multiple subset sizes, with some approaches even selecting core sets that surpass training on all available data. Notably, on DFC2022, a selected subset comprising only 25% of the training data yields slightly higher SegFormer performance than training with the entire dataset. This result shows the importance and potential of data-centric learning for the remote sensing domain. The code is available at https://github.com/keillernogueira/data-centric-rs-classification/

INDEX TERMS Core-set selection, data-centric machine learning, land-cover classification, semantic segmentation.

I. INTRODUCTION

The rapid advancements in satellite technologies have significantly enhanced accessibility to Earth observation data, opening new opportunities for a better understanding of the Earth's surface [1]. This accessibility has naturally led to the development and training of several deep learning methods using increasingly larger labeled datasets [2], often emphasizing label quantity over quality. However, indiscriminately increasing dataset size does not necessarily translate into improved model performance. This is because larger datasets require significant human effort or the integration

The associate editor coordinating the review of this manuscript and approving it for publication was Tai Fei.

of weak labels that, in turn, can lead to the introduction of noise, bias, and inaccurate annotations [3]. In general, the current assumption that more data inherently leads to better outcomes tends to overlook the complexities of data distribution [3], the potential for introducing biases and noise, spurious correlations, the energy consumption [4], and the computational resources required for processing, labeling, and storing vast datasets. Therefore, effective solutions should consider not only the quantity but also the quality of data.

One promising data-centric strategy that addresses these challenges is core-set selection, which aims to identify a small but informative subset of training examples that preserves the essential characteristics of the full dataset, while maintaining or even enhancing overall model performance. Such a paradigm can assist in several aspects, such as improved computational efficiency, cost-effective data handling, enhanced model performance, effective use of labeled data, or efficient labeling of unlabeled data [5], [6], [7], [8].

Existing work has explored core-set selection across various domains, employing distinct approaches, such as clustering algorithms and gradient approximation [9], [10], [11]. Some works perform core-set selection before training the final machine-learning model [9], [10], making them agnostic to the downstream model but limited to a single selection stage. In contrast, others integrate core-set selection into the training process by updating the selected subset at each epoch [11], [12], [13], yielding model-specific optimization at the cost of increased computational overhead. Importantly, while some of these works perform core-set selection for image classification, to the best of our knowledge, there have been no initiatives exploiting this paradigm for remote sensing image segmentation, which presents unique and complex challenges.

To address this gap, in this paper, we propose to establish a general and comprehensive baseline for core-set selection in remote sensing image segmentation. Towards this, we introduce and benchmark six basic **model-agnostic** core-set selection approaches for remote sensing image segmentation based on several distinct premises, that rely on imagery only, labels only, and a combination of both, and that can be readily adapted to different scenarios, applications, and modern architectures, including vision-language and foundation models. Specifically, given the training instances (i.e., images and their corresponding segmentation masks), the proposed approaches rank the examples from most to least valuable for training, based on their representativeness according to specific criteria. We then leverage such rankings to select the most representative examples (core-set) and train the models accordingly, reducing the training time and improving overall effectiveness by filtering out non-representative and/or noisy examples. The main contributions of this paper are the following:

- Introduction of six core-set selection approaches for remote sensing image segmentation, each based on different premises; and
- A full set of experiments comparing these approaches against two common baselines on three widely used datasets using two different architectures, thus establishing a benchmark for future research in core-set selection.

Overall, this work, an outcome of the Data-Centric Land Cover Classification Challenge of the Workshop on Machine Vision for Earth Observation and Environment Monitoring (MVEO) 2023, fills a critical gap in the literature and demonstrates the potential of core-set selection in advancing remote sensing image segmentation as well as data curation and labeling.

The remainder of this paper is organized as follows: Section II reviews related work, Section III introduces the six core-set selection methods, Section IV describes the experimental setup, Section V presents results and discussion, and Section VI concludes the paper.

II. RELATED WORK

One of the major reasons to work with a core-set instead of the full data set is an improvement in computational efficiency [9], [13]. Reducing the size of the dataset allows for quicker processing and experimentation. This makes it possible to use complex machine-learning models without immense computational costs. For this, oftentimes data-only techniques such as naïve random sampling are applied before model training, i.e., these methods do not need access to labels and are independent of the learning objective and application. An example is the identification of a subset that approximates the loss function of the whole dataset [9], [11], [13]. Furthermore, in addition to computational efficiency, smaller datasets reduce storage and maintenance costs, which is crucial when managing vast amounts of data from Earth observation systems.

Another reason is to improve model performance by enhancing the learning process and reducing overfitting through filtering out noisy or incorrect data points, thus creating cleaner datasets [14]. In remote sensing, predominantly prior knowledge, label information, or an existing model is used to identify a clean core-set [15], [16]. Santos et al. [10], for example, use clustering for satellite time series to identify instances that are mislabeled or have low accuracy, with the goal of removing them from the training set to avoid a decrease in model performance. Moreover, many methods have been developed for data with known sources of uncertainty, such as clouds [17], [18], [19]. Furthermore, this reason is directly related to the enhancement of the accuracy and robustness of machine learning models by removing low-quality or redundant examples. Such models can perform as well as, or even better than, those trained on the full dataset (e.g., [20], [21], [22]).

Another reason for core-set selection is to support clear and non-misleading explanations of a model and the data. Generally, in the field of explainable machine learning, new methods are introduced to calculate importance scores, sensitivities, or contributions of features and interpret them as relevance [23]. However, redundancies and correlations

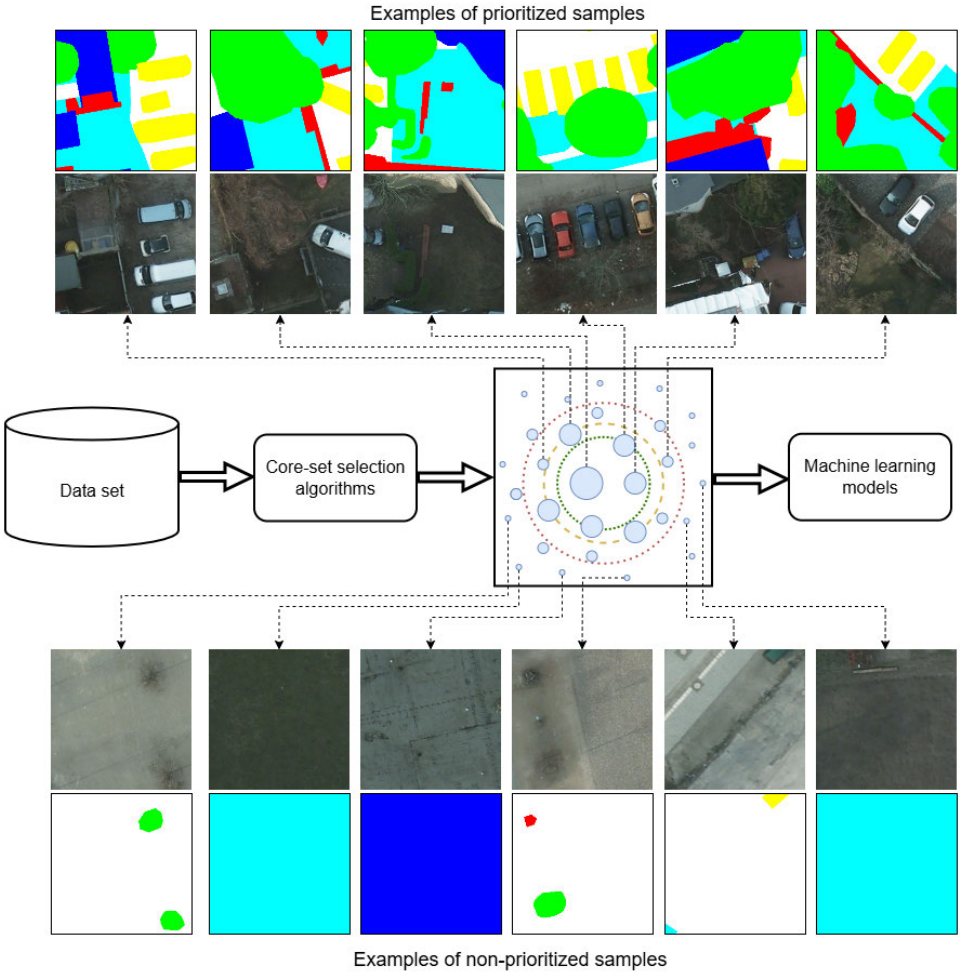


FIGURE 1. General overview of core-set selection. An input data set is first processed by a core-set selection algorithm that, based on some criteria, prioritizes certain examples over others (represented by the size of the blue circles). Based on this, it is possible to select the core-set data depending on the amount of data one would like to retain (illustrated by red, orange, and green dotted circles). Finally, the selected core-set is used to train a machine learning model, thus reducing the training time while maintaining, or even improving, task performance.

distort the derived insights, therefore they should be removed before interpreting and explaining the results. With the goal of analyzing geospatial air quality estimations and the relevance of specific measurement locations, Stadtler et al. [24] demonstrated that removing redundant examples only slightly decreases test accuracy, as these are not relevant for training.

In general, the principles of core-set selection are closely related to areas like active learning [12], [25], [26] and self-training [27]. For unlabeled datasets, core-set selection can guide efficient labeling by identifying the most representative examples. This optimizes resource allocation for manual annotation - a common objective in active learning scenarios [12], [25], [26]. In case the data is already labeled, core-set selection can help prioritize the most informative examples. This maximizes the use of labeled data and may reduce the need for further labeling efforts.

Overall, although most of the aforementioned works perform core-set selection for image classification, to the best of our knowledge, our work is the first research to design and benchmark core-set selection techniques specifically for remote sensing image segmentation.

III. CORE-SET SELECTION METHODS

Given a set of satellite images $\mathbf{X}$ and their corresponding label masks $\mathbf{M}$ (also known as segmentation maps), we introduce six basic core-set selection methods that assign an importance score, s_i, to the i-th instance (X_i, M_i). These scores range from 0 (least valuable for training) to 1 (most valuable for training). The aim of these methods is to rank the examples based on their informativeness, allowing us to select a subset of the data — a *core-set* — that can achieve good model performance with reduced training time and data size.

The proposed methods are categorized into three types: *label-based* methods, which rely solely on the label masks $\mathbf{M}$; *image-based* methods, which use information only from the input images $\mathbf{X}$; and *combined* methods that integrate both sources. A full summary of all proposed methods can be seen in Table 1.

TABLE 1. Overview of the six proposed core-set selection methods for remote sensing image segmentation.

Method	Category	Input	Key Technique	Selection Criterion
LC	Label-only	Masks ($\mathbf{M}$)	Entropy calculation	High entropy in class distribution indicates complex, informative masks
FD	Image-only	Images ($\mathbf{X}$)	K-Means clustering	Diverse feature representations across clusters
LC/FD	Hybrid	Both ($\mathbf{M}, \mathbf{X}$)	Combined ranking	Diversity for small sets, complexity for larger sets (cutoff at $m = 770$)
FA	Image-only	Images ($\mathbf{X}$)	Feature activation statistics	High mean and standard deviation in ResNet-18 embeddings
CB	Label-only	Masks ($\mathbf{M}$)	Iterative entropy maximization	Balanced class distribution across selected samples
FA/CB	Hybrid	Both ($\mathbf{M}, \mathbf{X}$)	Weighted ensemble ($\lambda = 0.5$)	Combined feature activation and class balance scores

For a given training budget b, the core-set consists of the top-b examples ranked by their scores. Next, we describe the methods in detail.

A. LABEL COMPLEXITY (LC)

LC is a *label-based* method that scores an data instance based on the **complexity** of its label mask M_i. The underlying assumption is that examples with high-complexity label masks are more informative for training segmentation models. It is important to emphasize that this approach does not explicitly guarantee representativeness but instead prioritizes complex label masks to potentially generate more informative training signals, which may lead to improved model performance.

The complexity of a label mask is quantified using the entropy of the class distribution in the label mask – high entropy class distributions will include more and mixed classes, while low or zero entropy class distributions will be dominated by a single class. Precisely, for each instance i, we compute the score s_i^{LC} based on the entropy $H(M_i)$ as:

$$s_i^{\text{LC}} = H(M_i) = -\sum_{c=1}^{C} p_{i,c} \log_C(p_{i,c}), \tag{1}$$

where C is the number of classes and $p_{i,c}$ is the proportion of pixels belonging to class c in M_i. Classes that are labeled as ''unknown'' or ''ignored'' (in DFC2022) are excluded from this computation.

Higher scores correspond to examples that have a more uniform distribution of class labels with potentially more informative label masks, while low scores correspond to examples that are dominated by a single class.

B. FEATURE SPACE DIVERSITY (FD)

The *FD* method is an *image-based* approach that aims to select a diverse core-set of examples from $\mathbf{X}$ by leveraging feature embeddings extracted from a pre-trained deep learning model.

First, we embed each image, X_i, using a ResNet-18 model [28] pre-trained on ImageNet [29]. We use the final feature representation layer (i.e. after spatial pooling) from the model, which results in a feature vector, $F_i \in \mathbb{R}^{512}$, that encodes higher-level semantic information per image.

Next, we group the feature embeddings into K clusters using the K-Means algorithm. We search for a value of K that minimizes the average Vendi score [30] across clusters. The Vendi score is a measure of diversity over a set of vectors – by minimizing the average within-cluster diversity we ensure that an instance from that cluster is representative of the others. Starting with $K = 2$, we cluster all image embeddings with K-Means, measure the average per-cluster Vendi score, then increment K, and repeat until the change in the average Vendi score falls below a threshold of δ (we choose $\delta = 0.5\%$) for three iterations.

Given a clustering of the examples, we sequentially choose one instance randomly from within each cluster in a round-robin fashion (the first cluster is randomly selected from the set of K clusters), resulting in ordered K-sized segments of cross-cluster examples. The first example has the highest importance score $s_i^{\text{FD}} = 1$, while the last selected example receives the lowest score $s_i^{\text{FD}} = 0$.

It is important to highlight here that this random selection step does not inherently produce a fixed ordering of scores. However, this does not impact the primary objective of this method, which is to ensure diversity among the selected examples.

C. COMPLEXITY/DIVERSITY HYBRID (LC/FD)

The *LC/FD* method is a *hybrid* method that combines the most important examples from the *LC* and *FD* methods, following an assumption that diversity is important when working with small datasets, while label complexity becomes increasingly important for medium to large datasets.

Specifically, the *LC/FD* method is a *hybrid* approach that uses the ranked lists of examples from *LC* and *FD*, denoted as $\mathcal{R}_{\text{LC}}$ and $\mathcal{R}_{\text{FD}}$, respectively. A cutoff point m is defined, and the hybrid ranking is constructed by taking the top m examples from $\mathcal{R}_{\text{FD}}$, followed by all examples from $\mathcal{R}_{\text{LC}}$ that are not already included. This ensures that the selected core-set includes a mix of label-complex examples and feature-diverse instances.

D. FEATURE ACTIVATION (FA)

The *FA* method is an *image-based* approach that uses statistics from image embeddings created by a pre-trained neural network to rank examples.

First, a ResNet-18 [28], pre-trained on the ImageNet dataset, is used to extract the image embeddings (after the final spatial pooling layer), resulting in a feature vector, $F_i \in \mathbb{R}^{512}$, that encodes higher-level semantic information per image.

By construction, all values in F_i are non-negative due to the application of a ReLU activation within the network.

We assume that examples with high activation magnitudes (large mean) and significant variations across different dimensions (high standard deviation) in feature space are likely to carry more relevant information. Accordingly, we compute the score s_i^{FA} by combining the mean μ_i and the standard deviation σ_i of the embedding vector F_i, with both quantities scaled to the interval $(0, 1]$, as follows:

$$s_i^{\text{FA}} = 1 - \left[\frac{\gamma_i - \min_{F_j \in \mathbf{F}}[\gamma_j]}{\max_{F_j \in \mathbf{F}}[\gamma_j] - \min_{F_j \in \mathbf{F}}[\gamma_j]} \right], \tag{2}$$

where $\gamma_i = -(1 - \mu_i) \cdot log(\sigma_i)$.

According to the aforementioned assumption, examples with low scores are likely to have lower diversity, contain more noise, etc., and are therefore likely to be less important for training.

E. CLASS BALANCE (CB)

Similar to the LC, the *CB* method is a *label-based* technique that aims to select a subset of examples with a uniform class distribution by using a time-efficient strategy that preprocesses and computes the class distribution of each image for subset selection.

The method consists of N steps, where N refers to the number of examples in the dataset. In each step, the most suitable instance is selected from the dataset to ensure that the overall class distribution of the selected examples approaches a uniform distribution. Specifically, the most suitable instance is the one that maximizes the entropy of the class distribution obtained by taking the union of the current core-set and the candidate example.

The order in which the examples are selected determines their importance score: the first instance is ranked as the most important and the last example is ranked as least important. Formally, let r_i be the rank of instance i, then:

$$s_i^{CB} = 1 - \frac{r_i}{N}. \tag{3}$$

F. FEATURE ACTIVATION/CLASS BALANCE HYBRID (FA/CB)

The *FA/CB* method is a *hybrid* method that uses a weighted ensemble of importance scores calculated by the previous two methods, that is:

$$s_i^{\text{FA/CB}} = \lambda \cdot s_i^{\text{FA}} + (1 - \lambda) \cdot s_i^{\text{CB}}, \tag{4}$$

where λ is a trade-off weight between the two scores.

IV. EXPERIMENTAL SETUP

A. DATASETS

We test our approaches using three high-resolution datasets for semantic segmentation with remotely-sensed imagery as described below.

1) IEEE GRSS DATA FUSION CONTEST 2022 (DFC2022) DATASET

The DFC2022 dataset [31] was released as part of the annual IEEE GRSS Data Fusion Contest. This dataset consists of images gathered in and around 19 urban areas from different regions in France. Each instance of this dataset contains a high-resolution RGB aerial image and its corresponding segmentation mask, both having approximately 2000 × 2000 pixels and a spatial resolution of 50 cm per pixel, along with a Digital Elevation Model (DEM) with 1000 × 1000 pixels at a spatial resolution of 100cm/pixel. For our experiments, we resample the DEM data to match the dimensions of the image and mask. The masks in this dataset contain 12 classes: ''urban fabric'', ''industrial, commercial, public, military, private and transport units'', ''mine, dump and construction sites'', ''artificial non-agricultural vegetated areas'', ''arable land'', ''permanent crops'', ''pastures'', ''forests'', ''herbaceous vegetation associations'', ''open spaces with little or no vegetation'', ''wetlands'', and ''water''. Examples of this dataset are shown in Figure 2.

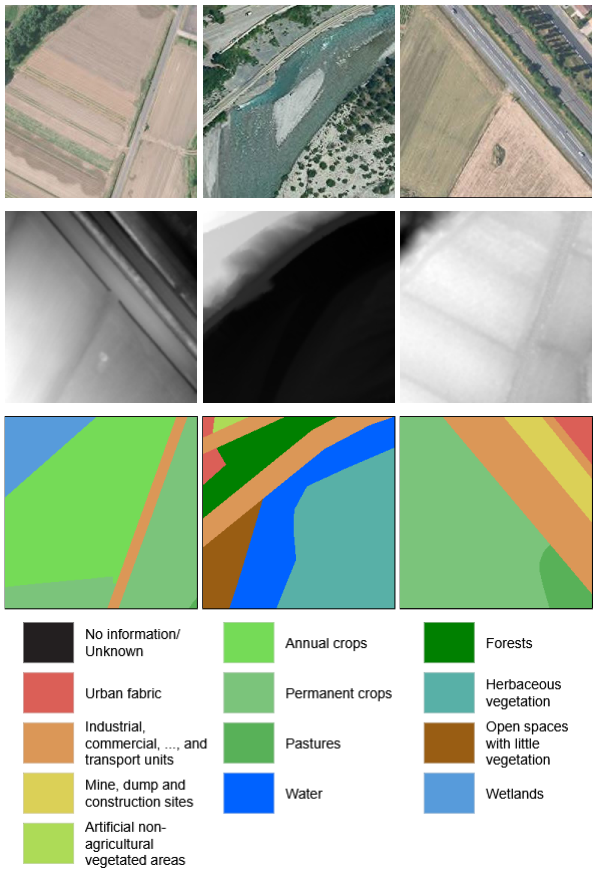


FIGURE 2. Example images (first row) of the DFC2022 dataset [31], their DEM data (second row), and the respective reference data (third row).

2) ISPRS VAIHINGEN AND POTSDAM DATASETS

The Vaihingen and Potsdam datasets [32] were released for the 2D semantic labeling contest of the International Society for Photogrammetry and Remote Sensing (ISPRS). Both

datasets consist of aerial imagery, Digital Surface Model (DSM) data, and label masks, as shown in Figure 3. The Vaihingen dataset contains 33 patches, with an average size of 2494 × 2064 pixels. The aerial images have three bands (near-infrared, red, and green) with a spatial resolution of 9 cm per pixel. The Postdam dataset contains 38 tiles of 6000 × 6000 pixels. The imagery consists of four bands (red, green, blue, and near-infrared) with a spatial resolution of 5 cm per pixel. The label masks in both datasets contain six classes: ''impervious surfaces'', ''building'', ''low vegetation'', ''tree'', ''car'', and ''clutter/background''.

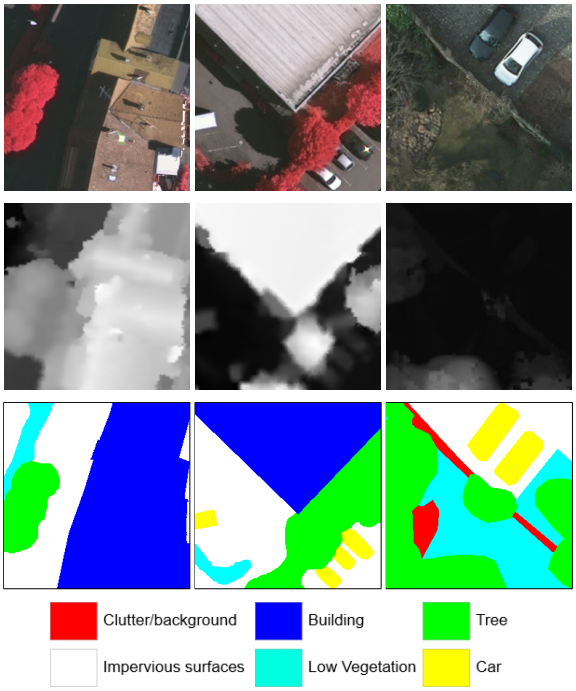


FIGURE 3. Example images (first row) of the Vaihingen and Potsdam datasets [32], their DSM data (second row), and the respective reference data (third row).

B. IMPLEMENTATION DETAILS

We preprocess the aforementioned datasets by tiling them into non-overlapping 256 × 256 patches that are used in all subsequent steps.

We evaluate the effectiveness of the core-set selection methods using two popular segmentation networks: U-Net [33] (with a ResNet-18 [28] backbone), and SegFormer [34]. The former is a convolutional segmentation model, while the latter is based on the transformer architecture. For each experimental dataset, both models are trained exclusively on the selected core-set and evaluated on a held-out test set. Importantly, our training and testing routine is fixed over the experiments, the only difference is in the subset (and size of subset) used to train the segmentation model.

All proposed methods[1] are implemented using Pytorch. During training, we use the following hyper-parameters: 100 training epochs, AdamW [35] as optimizer, learning rate of 0.001, and batch size of 64.

For the *LC/FD* method, we let $m = 770$ based on preliminary experiments with the DFC2022 dataset (we observed that *FD* outperformed *LC* for small subsets, while *LC* added value for larger datasets).

For the *FA/CB* method, we let $\lambda = 0.5$ to equally weight the importance from the *FA* and *FB* methods.

C. BASELINES

Given our general aim, we focus on model-agnostic core-set selection baseline methods, motivated by their high adaptability and usability, allowing them to be easily customized and applied across different downstream models and scenarios. For all datasets, we compare the proposed techniques with two **traditional model-agnostic** baseline models: (i) **Random**, which selects the core set uniformly at random, and (ii) **CoreSet** [12], [21], [22], which selects samples to optimally cover the embedding space. Precisely, this approach iteratively expands the core-set by adding the data point that is farthest from its nearest neighbor in the current (core) set. In this case, we used the Euclidean distance in the activations of the last spatial pooling layer of the ResNet-18 [28], similar to the approach used in the FA and FD methods.

D. EXPERIMENTAL PROTOCOL

For the DFC2022 dataset, 90% of the data originally released for the data fusion contest is made available to be ranked by the proposed core-set algorithms, while the remaining 10% is used for validation.

For the Vaihingen and Potsdam datasets, we follow the standard protocol commonly exploited in the literature [36]. Specifically, for the Vaihingen dataset, 11 images originally released for the contest are made available for the proposed core-set techniques, and 5 images (with IDs 11, 15, 28, 30, 34) are employed for validation. For the Potsdam dataset, 18 images released for the contest are made available for the proposed techniques, whereas 6 images (with IDs 02_12, 03_12, 04_12, 05_12, 06_12, 07_12) are used for validation.

For all datasets, the validation is only employed to assess the training of the U-Net models, after the selection of the core-set. The final evaluation of the trained U-Net models uses the original test set of each dataset. The overall performance of each method is measured by the mean Intersection over Union (mIoU) across all segmentation classes and averaged over three different model training runs.

V. EXPERIMENTS AND DISCUSSION

A. QUANTITATIVE RESULTS

To compare the performance of the introduced methods, we train and test both a U-Net and a SegFormer model (following the configuration described in Section IV-B) on the top 1%, 5%, 10%, 25%, 50%, and 75% ranked patches from each method. The same experimental protocol is applied to the baseline approaches, with the additional inclusion

[1]The code is made publicly available at https://github.com/keillernogueira/data-centric-rs-classification/

TABLE 2. Results (% mIoU) of the U-Net model [33] trained on subsets of varying sizes (1%, 5%, 10%, 25%, 50%, 75%, and 100%) for the DFC2022, Vaihingen, and Potsdam datasets. Underlined values indicate the results that outperformed the corresponding baselines per training percentage (statistically significant paired t-test at $\alpha = 0.05$). Bold values represent the best results overall for the dataset.

Methods	DFC 2022							
	Category	1%	5%	10%	25%	50%	75%	100%
Random	-	10.42 ± 0.46	10.84 ± 0.15	11.55 ± 0.32	11.30 ± 0.66	12.41 ± 0.30	12.40 ± 0.36	12.71 ± 0.53
CoreSet [12], [21], [22]	-	10.73 ± 0.23	11.47 ± 0.34	11.84 ± 0.42	11.47 ± 0.82	12.33 ± 0.27	12.23 ± 0.43	-
Label Complexity (LC)	Label-only	10.42 ± 0.76	12.44 ± 0.52	12.83 ± 0.22	12.68 ± 0.29	**13.41 ± 0.61**	12.64 ± 0.35	-
Feature Diversity (FD)	Image-only	9.93 ± 0.24	11.87 ± 0.34	11.31 ± 0.29	11.61 ± 0.21	12.23 ± 0.08	12.00 ± 0.44	-
LC/FD Hybrid	Both	10.55 ± 0.16	11.98 ± 0.27	12.84 ± 0.33	13.23 ± 0.12	12.62 ± 0.16	12.39 ± 0.30	-
Feature Activation (FA)	Image-only	8.67 ± 0.07	11.23 ± 0.65	11.14 ± 0.34	11.80 ± 0.42	12.66 ± 0.15	12.57 ± 0.11	-
Class Balance (CB)	Label-only	9.93 ± 1.36	10.38 ± 0.81	10.79 ± 0.20	11.52 ± 0.42	11.71 ± 0.21	12.27 ± 0.45	-
FA/CB Hybrid	Both	8.84 ± 0.53	10.44 ± 0.57	10.73 ± 0.50	11.90 ± 0.85	11.27 ± 0.32	12.22 ± 0.39	-

Methods	Vaihingen							
	Category	1%	5%	10%	25%	50%	75%	100%
Random	-	31.59 ± 1.59	39.89 ± 1.29	43.02 ± 1.48	53.72 ± 1.60	51.23 ± 1.89	57.86 ± 0.68	58.87 ± 0.57
CoreSet [12], [21], [22]	-	29.71 ± 2.40	37.96 ± 2.30	41.39 ± 2.56	54.13 ± 1.68	55.00 ± 1.87	57.38 ± 0.96	-
Label Complexity (LC)	Label-only	34.98 ± 1.67	42.58 ± 1.25	42.24 ± 1.39	55.14 ± 0.88	52.89 ± 3.85	58.35 ± 3.35	-
Feature Diversity (FD)	Image-only	31.40 ± 0.48	38.84 ± 1.81	42.99 ± 2.94	52.18 ± 3.12	53.19 ± 2.59	59.14 ± 1.67	-
LC/FD Hybrid	Both	23.52 ± 2.22	37.91 ± 1.30	41.22 ± 3.77	52.98 ± 2.72	50.44 ± 4.08	58.45 ± 0.42	-
Feature Activation (FA)	Image-only	27.53 ± 0.40	37.94 ± 4.58	43.59 ± 1.26	52.61 ± 0.11	56.34 ± 0.46	57.96 ± 0.19	-
Class Balance (CB)	Label-only	28.16 ± 5.08	32.22 ± 2.24	45.65 ± 1.08	55.48 ± 1.61	54.64 ± 3.98	60.54 ± 1.33	-
FA/CB Hybrid	Both	30.48 ± 5.86	40.69 ± 0.42	45.33 ± 1.90	56.77 ± 1.42	54.79 ± 2.97	**61.27 ± 0.47**	-

Methods	Potsdam							
	Category	1%	5%	10%	25%	50%	75%	100%
Random	-	48.40 ± 1.97	68.42 ± 2.38	55.96 ± 2.81	72.04 ± 1.77	69.93 ± 1.28	78.75 ± 1.42	78.67 ± 1.40
CoreSet [12], [21], [22]	-	52.84 ± 2.04	66.68 ± 3.55	67.13 ± 7.55	72.26 ± 2.04	68.74 ± 4.50	78.79 ± 1.40	-
Label Complexity (LC)	Label-only	56.45 ± 1.85	65.03 ± 1.28	66.48 ± 4.34	71.62 ± 4.31	76.90 ± 3.48	80.39 ± 0.29	-
Feature Diversity (FD)	Image-only	49.06 ± 3.25	68.54 ± 0.95	66.08 ± 4.49	74.67 ± 0.81	68.58 ± 2.03	79.54 ± 0.26	-
LC/FD Hybrid	Both	51.12 ± 3.52	69.39 ± 0.33	68.14 ± 6.37	72.96 ± 2.40	73.60 ± 5.87	80.98 ± 3.75	-
Feature Activation (FA)	Image-only	48.17 ± 1.38	67.30 ± 0.59	55.12 ± 1.66	70.37 ± 7.33	68.98 ± 0.99	80.04 ± 0.11	-
Class Balance (CB)	Label-only	48.36 ± 4.20	57.03 ± 5.13	64.35 ± 0.62	72.83 ± 0.07	73.89 ± 4.62	74.86 ± 5.36	-
FA/CB Hybrid	Both	52.51 ± 3.22	63.68 ± 0.85	62.64 ± 1.02	73.76 ± 0.23	76.96 ± 1.21	**81.28 ± 0.06**	-

of models trained on 100% of the available training data, which serves as a robust reference baseline. To account for potential variability due to randomness, three models are trained for each approach and subset size, using the same selected examples (per subset) and hyperparameters. Finally, we used a paired t-test with $\alpha = 0.05$ to evaluate statistically significant differences in results across methods.

Tables 2 and 3 show the results for each method across both evaluated network architectures, the three datasets, and the different core-set sizes. Overall, the proposed approaches consistently outperform the baselines across all datasets and architectures, including the setting where models are trained on 100% of the available data. Moreover, in most cases, the proposed approaches outperform both baselines even when using substantially fewer training examples. For instance, on the Potsdam dataset, the U-Net model [33] trained using only 25% of the top-ranked samples selected by the FA/CB Hybrid approach achieves a higher mIoU score than both baselines trained on the top-ranked 50% of the data. These results demonstrate that the proposed methods effectively identify representative and informative training samples, enabling more data-efficient and performant segmentation models.

Furthermore, although the methods show varying degrees of effectiveness, the *label-based* techniques outperformed the baselines at least once on all datasets across both networks. Overall, these methods achieved the best overall performance (outperforming the baseline trained on 100% of available data) in 3 dataset–model combinations, highlighting the importance of label diversity for effective core-set selection. In comparison, the *image-based* approaches yielded more moderate gains. Although they outperformed the baselines in fewer cases, they still achieved the best performance in one dataset–model combination, specifically for the SegFormer model trained on the Vaihingen dataset, indicating that image-level representativeness alone can be beneficial in certain settings. Finally, *combined* approaches outperformed the baselines at least once across all datasets and both architectures, achieving the best overall results in 2 dataset–model combinations, showing that jointly exploiting label and image representativeness can be particularly beneficial in certain scenarios.

TABLE 3. Results (% mIoU) of the SegFormer model [34] trained on subsets of varying sizes (1%, 5%, 10%, 25%, 50%, 75%, and 100%) for the DFC2022, Vaihingen, and Potsdam datasets. Underlined values indicate the results that outperformed the corresponding baselines per training percentage (statistically significant paired t-test at $\alpha = 0.05$). Bold values represent the best results overall for the dataset.

Methods	DFC 2022							
	Category	1%	5%	10%	25%	50%	75%	100%
Random	-	11.55 ± 0.26	12.48 ± 0.17	11.93 ± 0.37	12.27 ± 0.06	12.69 ± 0.50	11.87 ± 0.70	12.00 ± 0.45
CoreSet [12], [21], [22]	-	11.84 ± 0.46	12.29 ± 0.55	12.02 ± 0.22	12.29 ± 0.37	12.23 ± 0.75	12.35 ± 0.24	-
Label Complexity (LC)	Label-only	10.82 ± 0.53	12.21 ± 0.34	12.60 ± 0.55	**13.01 ± 0.32**	12.64 ± 0.18	12.13 ± 0.11	-
Feature Diversity (FD)	Image-only	9.95 ± 0.58	11.44 ± 0.42	12.43 ± 0.01	11.68 ± 0.54	11.62 ± 0.15	12.42 ± 0.25	-
LC/FD Hybrid	Both	11.03 ± 0.86	11.91 ± 1.07	12.86 ± 0.06	11.97 ± 0.27	12.23 ± 0.23	11.82 ± 0.44	-
Feature Activation (FA)	Image-only	10.11 ± 1.14	10.90 ± 0.49	11.85 ± 0.57	12.59 ± 0.20	12.48 ± 0.06	12.86 ± 0.26	-
Class Balance (CB)	Label-only	9.82 ± 0.42	10.16 ± 0.73	11.29 ± 0.45	11.39 ± 0.58	11.78 ± 0.76	12.33 ± 0.76	-
FA/CB Hybrid	Both	10.85 ± 1.03	10.35 ± 1.35	11.39 ± 0.87	11.08 ± 1.03	11.72 ± 0.65	12.36 ± 0.24	-

Methods	Vaihingen							
	Category	1%	5%	10%	25%	50%	75%	100%
Random	-	12.02 ± 0.01	34.44 ± 8.05	51.96 ± 1.02	56.19 ± 2.25	56.01 ± 0.62	58.67 ± 1.01	57.71 ± 1.28
CoreSet [12], [21], [22]	-	28.88 ± 2.58	42.34 ± 2.38	50.82 ± 2.38	57.77 ± 1.13	56.48 ± 1.55	60.09 ± 0.06	-
Label Complexity (LC)	Label-only	38.81 ± 0.05	48.27 ± 0.30	51.62 ± 0.13	54.92 ± 2.07	58.54 ± 0.16	58.02 ± 0.38	-
Feature Diversity (FD)	Image-only	29.36 ± 0.32	40.74 ± 2.53	50.81 ± 1.82	56.41 ± 1.57	56.72 ± 2.16	56.99 ± 2.81	-
LC/FD Hybrid	Both	30.65 ± 2.25	40.32 ± 6.68	51.30 ± 1.81	55.26 ± 0.78	55.36 ± 3.86	58.85 ± 1.22	-
Feature Activation (FA)	Image-only	34.12 ± 5.74	46.18 ± 2.41	51.41 ± 0.89	54.02 ± 0.77	55.60 ± 1.82	**60.97 ± 0.46**	-
Class Balance (CB)	Label-only	37.87 ± 1.98	41.63 ± 2.87	46.40 ± 4.24	54.31 ± 0.92	58.60 ± 1.91	58.72 ± 2.74	-
FA/CB Hybrid	Both	34.40 ± 1.49	48.72 ± 3.26	50.12 ± 3.61	58.34 ± 0.25	57.12 ± 3.32	60.39 ± 0.85	-

Methods	Potsdam							
	Category	1%	5%	10%	25%	50%	75%	100%
Random	-	60.57 ± 0.05	70.62 ± 0.33	71.29 ± 3.63	68.88 ± 1.29	78.75 ± 0.14	76.62 ± 5.11	80.54 ± 2.07
CoreSet [12], [21], [22]	-	60.56 ± 3.73	70.18 ± 0.35	69.55 ± 4.45	72.91 ± 3.81	78.94 ± 0.28	80.79 ± 2.25	-
Feature Diversity (FD)	Image-only	56.07 ± 6.49	69.98 ± 0.30	69.91 ± 2.93	67.18 ± 1.72	76.82 ± 3.43	78.65 ± 7.72	-
Label Complexity (LC)	Label-only	60.75 ± 0.86	58.92 ± 2.19	71.87 ± 0.16	62.82 ± 1.42	78.13 ± 3.36	83.78 ± 0.05	-
LC/FD Hybrid	Both	54.03 ± 0.87	71.58 ± 0.64	67.48 ± 3.73	65.87 ± 1.58	78.43 ± 3.04	83.83 ± 0.08	-
Feature Activation (FA)	Image-only	56.91 ± 0.43	60.72 ± 6.54	67.76 ± 2.49	63.71 ± 2.55	78.37 ± 1.31	82.91 ± 0.02	-
Class Balance (CB)	Label-only	53.29 ± 0.70	55.33 ± 3.51	63.33 ± 2.63	67.71 ± 1.73	77.20 ± 3.87	**83.96 ± 0.03**	-
FA/CB Hybrid	Both	59.27 ± 1.44	58.89 ± 4.63	65.98 ± 4.41	67.73 ± 3.01	76.56 ± 3.59	83.91 ± 0.08	-

We also observe that each dataset has a different IoU convergence rate. DFC2022 has the most label noise (as evidenced by performance degradation of models in later splits), and the best-performing methods reached near-peak performance on both architectures by utilizing only 25-50% of the data, noticeably outperforming training on 100% of the data. This rapid convergence suggests that, for datasets with specific characteristics (e.g., redundancy, noise), a relatively small core-set can be as effective, or even more effective, than the full dataset. In contrast, on datasets with less label noise, such as Vaihingen and Potsdam, the performance continues to gradually improve on both architectures as the number of training examples increases. However, even in these cases, the introduced methods are able to outperform the baseline trained on the full training set, demonstrating the importance of selecting a core set to deal with relevant issues such as noise, representativeness, and so on.

In addition to performance gains, Table 4 reports the training time per epoch (in seconds) of the U-Net model [33]. All experiments were conducted on a machine equipped with an Intel Xeon E5-2695 v4 (Broadwell) CPU, 128GB of RAM, and an NVIDIA P100 (Pascal) GPU with 16GB of memory, running CUDA 11.2 on Red Hat Enterprise Linux 8.2. Training time is reported exclusively for the U-Net architecture as a representative model. This is motivated by the fact that the relative reduction in computational cost is primarily driven by the size of the selected core-set and exhibits similar scaling behaviour across the different architectures. Consequently, the observed trends generalize to the other model considered in this work. Since all models are trained for a fixed number of 100 epochs, the reported per-epoch times allow us to directly quantify the computational savings enabled by core-set selection. For example, when training the U-Net model [33] on the DFC2022 dataset, the best-performing model (using only 50% of the data) completed training more than 24 hours faster than the baseline trained on 100% of the data, while achieving a superior test-set performance. It is essential to observe that the computational overhead of the core-set selection process itself is minimal, amounting to at most the cost of a single training epoch using the full dataset, and is therefore **negligible** compared to the total training time. In general,

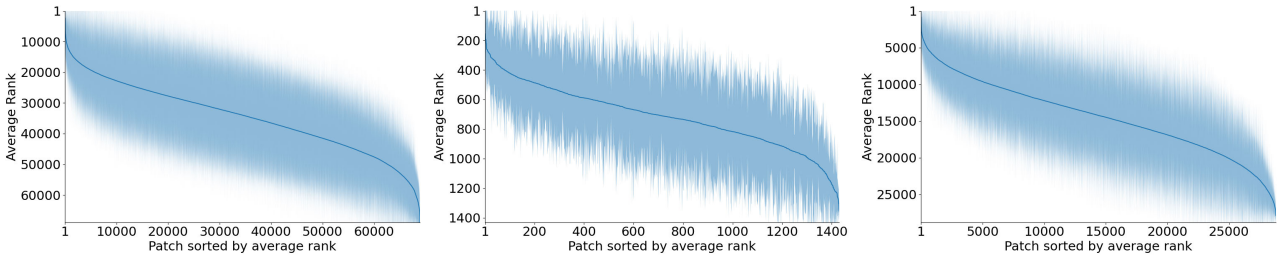


FIGURE 4. Visualizations of the proposed methods' rankings. The line represents the average rank position for each patch across all proposed approaches, while the shaded area represents one standard deviation.

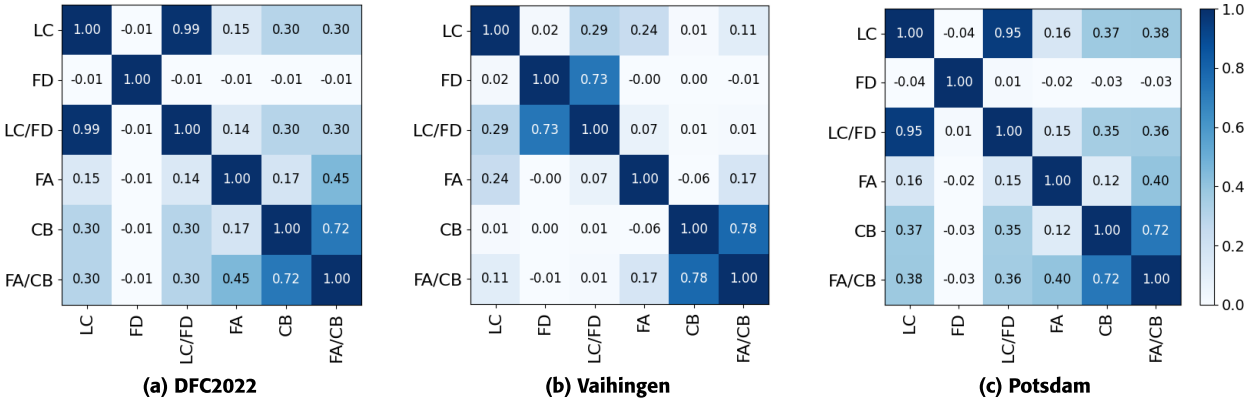


FIGURE 5. Correlation of methods according to Kendall Tau coefficient. A high correlation value means that the methods produce similar rankings.

beyond reducing time, faster training enables the use of more training epochs and/or more extensive hyperparameter tuning, further supporting improved model optimization.

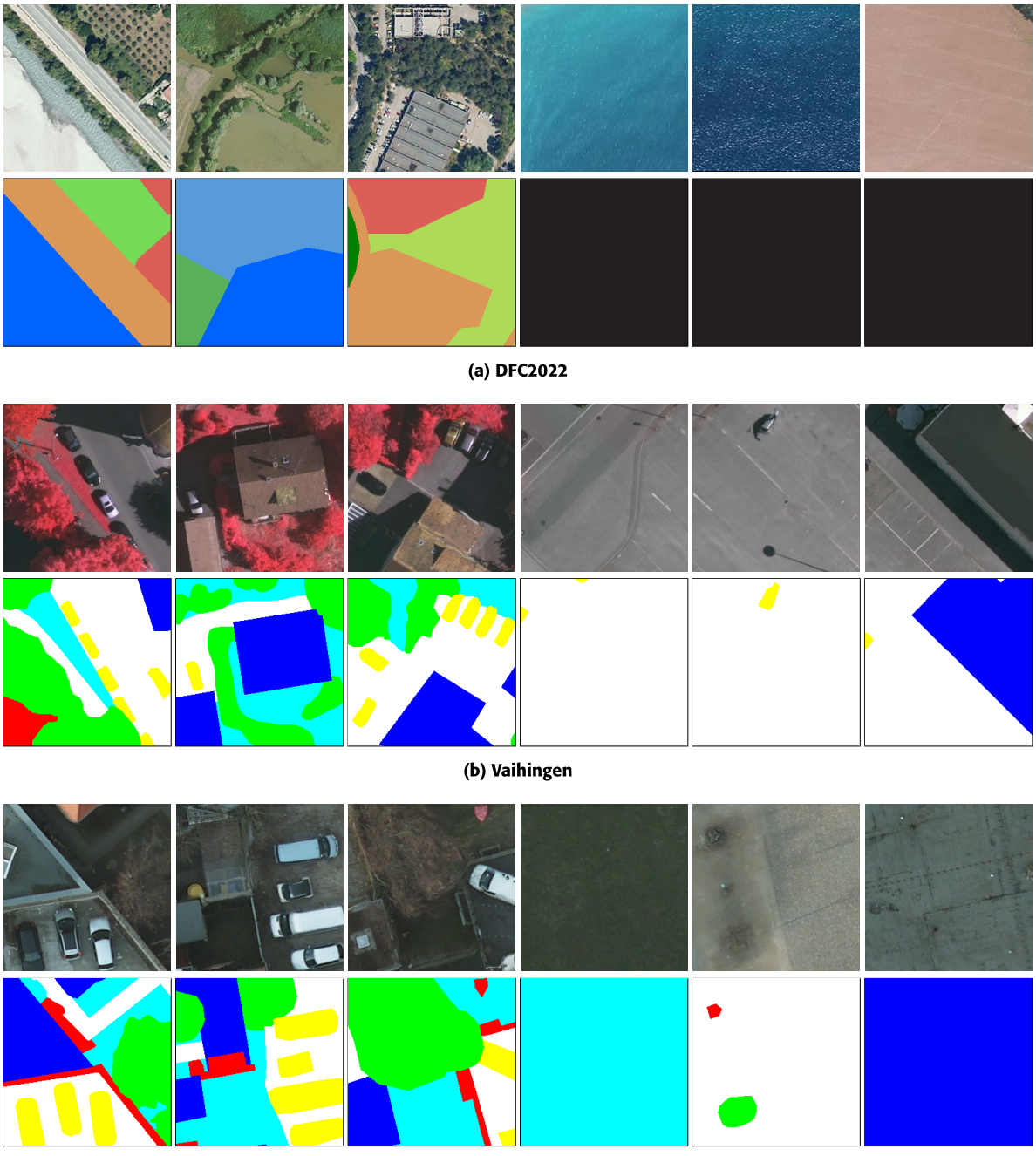
TABLE 4. Training time per epoch (in seconds) of the U-Net model [33] for different core-set sizes. The per-epoch training time is independent of the specific core-set selection method and depends solely on the size of the selected subset, given that the computational overhead of the core-set selection process itself is negligible compared to the full training procedure.

Dataset	1%	5%	10%	25%	50%	75%	100%
DFC2022	40.65	107.34	194.62	428.14	810.97	1191.38	1981.11
Vaihingen	4.10	7.92	8.70	11.91	20.11	28.94	38.47
Potsdam	6.55	10.20	14.57	28.04	48.51	71.26	93.79

B. QUALITATIVE RESULTS

To facilitate the analysis and comparison of the proposed methods' outputs, we include visualizations of the average generated rankings, as can be seen in Figure 4. To generate these visualizations, the rankings produced by the different introduced methods are first averaged by patch position and then sorted, making the highest-ranking patches across the methods appear at the top. Additionally, the standard deviation is calculated to capture the variability of the assigned ranks and provide insight into the approaches' consistency. In addition to these visualizations, to allow for a better analysis, we also report the Kendall Tau correlations between each pair of methods in Figure 5 and provide examples of the most and least frequently selected instances across all introduced methods in Figure 6.

For all datasets, it is possible to observe that the approaches exhibit notable consistency, frequently assigning more importance to a specific (core) set of high-complexity examples that are consistently selected across the proposed methods. Similarly, such approaches also tend to agree on the least important examples, assigning lower scores to low-complexity patches, indicating that the explored datasets contain a subset of non-representative or noisy instances that either contribute minimally to the overall performance or, in some cases, may even degrade it. Furthermore, this level of agreement between the proposed techniques can be further observed in the correlation plots, wherein several methods show substantial correlation (particularly for the DFC2022 and Potsdam datasets), suggesting that they can capture underlying dataset patterns (such as the core sets). Overall, the ability to select the core set, along with the identification of less valuable examples, highlights the robustness and efficiency of the proposed techniques in distinguishing between high- and low-quality data, thereby resulting in better performance and training time. This is also qualitatively demonstrated in Figure 6: examples that are consistently highly ranked by the proposed methods show clear imagery with artifacts that are of a certain visual and semantic complexity (as illustrated by the corresponding label maps). A few of these images allow learning the appearance and spatial relation of several classes at once.



(a) DFC2022

(b) Vaihingen

(c) Potsdam

FIGURE 6. Examples of the highest-ranked (first three columns) and lowest-ranked instances (last three columns) considering the average ranking of all proposed methods.

On the other hand, images that are consistently rejected show very homogeneous scenes (such as large water bodies or parking lots) with neither much visual variation nor complex semantic content.

VI. CONCLUSION

In this paper, we introduce and systematically benchmark six basic core-set selection approaches for remote sensing image segmentation based on distinct premises - which rely on imagery only, labels only, or a combination of both - thereby establishing a general and comprehensive baseline for future works. The proposed methods are able to consistently and effectively select the most important subset of examples (i.e., core-set), filtering out non-representative and noisy samples while preserving (or even improving) segmentation performance.

Extensive experiments are conducted using two different architectures (U-Net [33] and SegFormer [34]) across three high-resolution remote sensing datasets with very distinct properties: (i) IEEE GRSS Data Fusion Contest

2022 (DFC2022) dataset [31], consisting of very high-resolution visible spectrum images and Digital Elevation Model imagery, and (ii) Vaihingen and Potsdam datasets [32], both composed of high-resolution multispectral images and normalized Digital Surface Model data. This diverse experimental setup enables a robust assessment of the proposed methods across distinct settings.

Experimental results demonstrate the effectiveness and computational efficiency of the proposed core-set selection strategies, which consistently outperform traditional baselines. Notably, on the DFC2022 dataset, the proposed approaches outperform the baseline trained on 100% of the data while using only 25-50% of the available examples. Similarly, on the Vaihingen and Potsdam datasets, the same superior performance is achieved using just 75% of the data. These findings highlight that carefully selected core-sets can not only improve model performance but also substantially reduce training time and computational costs.

In summary, this work addresses a crucial gap in the literature and demonstrates the potential of core-set selection in advancing remote sensing image segmentation, as well as data creation and labeling. The presented conclusions open opportunities towards: (i) the integration of core-set selection with other advanced techniques, such as self-supervised learning and foundation models, and (ii) a more efficient and effective exploitation of both existing and new datasets for a better understanding of the Earth's surface, an essential characteristic for most applications.

ACKNOWLEDGMENT

(Keiller Nogueira, Akram Zaytar, and Wanli Ma contributed equally to this work.)

REFERENCES

[1] Y. Ban, P. Gong, and C. Giri, "Global land cover mapping using Earth observation satellite data: Recent progresses and challenges," *ISPRS J. Photogramm. Remote Sens.*, vol. 103, pp. 1–6, May 2015.

[2] M. Schmitt, S. A. Ahmadi, Y. Xu, G. Taşkin, U. Verma, F. Sica, and R. Hänsch, "There are no data like more data: Datasets for deep learning in Earth observation," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 3, pp. 63–97, Sep. 2023.

[3] R. Roscher, M. Russwurm, C. Gevaert, M. Kampffmeyer, J. A. Dos Santos, M. Vakalopoulou, R. Hänsch, S. Hansen, K. Nogueira, J. Prexl, and D. Tuia, "Better, not just more: Data-centric machine learning for earth observation," *IEEE Geosci. Remote Sens. Mag.*, vol. 12, no. 4, pp. 335–355, Apr. 2024.

[4] L. Lannelongue, J. Grealey, and M. Inouye, "Green algorithms: Quantifying the carbon footprint of computation," *Adv. Sci.*, vol. 8, no. 12, Jun. 2021, Art. no. 2100707.

[5] J. M. Phillips, "Coresets and sketches," in *Handbook of Discrete and Computational Geometry*. London, U.K.: Chapman & Hall, 2017, pp. 1269–1288.

[6] A. Ng, "Unbiggen AI," *IEEE Spectr.*, vol. 9, Feb. 2022.

[7] M. H. Jarrahi, A. Memariani, and S. Guha, "The principles of data-centric AI," *Commun. ACM*, vol. 66, no. 8, pp. 84–92, Aug. 2023.

[8] L. Aroyo, M. Lease, P. Paritosh, and M. Schaekermann, "Data excellence for AI: Why should you care?" *Interactions*, vol. 29, no. 2, pp. 66–69, Mar. 2022.

[9] C. Chai, J. Wang, N. Tang, Y. Yuan, J. Liu, Y. Deng, and G. Wang, "Efficient coreset selection with cluster-based methods," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2023, pp. 167–178.

[10] L. A. Santos, K. R. Ferreira, G. Camara, M. C. A. Picoli, and R. E. Simoes, "Quality control and class noise reduction of satellite image time series," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 75–88, Jul. 2021.

[11] O. Pooladzandi, D. Davini, and B. Mirzasoleiman, "Adaptive second order coresets for data-efficient machine learning," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 17848–17869.

[12] O. Şener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[13] B. Mirzasoleiman, J. Bilmes, and J. Leskovec, "Coresets for data-efficient training of machine learning models," in *Proc. Int. Conf. Mach. Learn.*, vol. 1, 2019, pp. 6950–6960.

[14] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "CleanML: A study for evaluating the impact of data cleaning on ML classification tasks," in *Proc. IEEE 37th Int. Conf. Data Eng. (ICDE)*, Apr. 2021, pp. 13–24.

[15] I. F. Ilyas and T. Rekatsinas, "Machine learning and data cleaning: Which serves the other?" *J. Data Inf. Qual.*, vol. 14, no. 3, pp. 1–11, Sep. 2022.

[16] F. Neutatz, B. Chen, Z. Abedjan, and E. Wu, "From cleaning before ML to cleaning for ML," *IEEE Data Eng. Bull.*, vol. 44, no. 1, pp. 24–41, Jan. 2021.

[17] Y. Zhang, F. Wen, Z. Gao, and X. Ling, "A coarse-to-fine framework for cloud removal in remote sensing image sequence," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5963–5974, Aug. 2019.

[18] J. Li, Z. Wu, Z. Hu, J. Zhang, M. Li, L. Mo, and M. Molinier, "Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 373–389, Aug. 2020.

[19] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022.

[20] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *J. Artif. Intell. Res.*, vol. 70, pp. 1373–1411, Apr. 2021.

[21] G. Citovsky, G. DeSalvo, C. Gentile, L. Karydas, A. Rajagopalan, A. Rostamizadeh, and S. Kumar, "Batch active learning at scale," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 11933–11944.

[22] D. Bahri, H. Jiang, T. Schuster, and A. Rostamizadeh, "Is margin all you need? An extensive empirical study of active learning on tabular data," 2022, *arXiv:2210.03822*.

[23] R. Roscher, B. Bohn, M. Duarte, and J. Garcke, "Explain it to me–facing remote sensing challenges in the bio-and geosciences with explainable machine learning," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 817–824, Jan. 2020.

[24] S. Stadtler, C. Betancourt, and R. Roscher, "Explainable machine learning reveals capabilities, redundancy, and limitations of a geospatial air quality benchmark dataset," *Mach. Learn. Knowl. Extraction*, vol. 4, no. 1, pp. 150–171, Feb. 2022.

[25] Y. Kim and B. Shin, "In defense of core-set: A density-aware core-set selection for active learning," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 804–812.

[26] W. Zhang, Z. Guo, R. Zhi, and B. Wang, "Deep active learning for human pose estimation via consistency weighted core-set approach," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 909–913.

[27] B. Wei, C. Yi, Q. Zhang, H. Zhu, J. Zhu, and F. Jiang, "ActiveSelfHAR: Incorporating self-training into active learning to improve cross-subject human activity recognition," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 6833–6847, Feb. 2024.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[30] D. Friedman and A. Bousso Dieng, "The vendi score: A diversity evaluation metric for machine learning," 2022, *arXiv:2210.02410*.

[31] R. Hänsch, C. Persello, G. Vivone, J. C. Navarro, A. Boulch, S. Lefevre, and B. Saux, "The 2022 IEEE GRSS data fusion contest: Semisupervised learning [technical committees]," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 334–337, Mar. 2022.

[32] *International Society for Photogrammetry and Remote Sensing (ISPRS)*. Accessed: Aug. 1, 2024. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/semantic-labeling.aspx

[33] O. Ronneberger, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, 2024, pp. 234–241.

[34] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.
[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[36] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7503–7520, Oct. 2019.

KEILLER NOGUEIRA received the B.Sc. degree in computer science from the Universidade Federal de Viçosa, Brazil, in 2012, and the M.Sc. and Ph.D. degrees in computer science from the Universidade Federal de Minas Gerais, Brazil, in 2015 and 2019, respectively. Since 2024, he has been a Lecturer with the School of Computer Science and Informatics with the University of Liverpool, U.K. Before that, from 2019 to 2024, he was a Lecturer with the data science, University of Stirling, U.K. He has published several high-quality articles in leading journals and conferences. His research interests include deep and machine learning, pattern recognition, image processing, computer vision, and remote sensing.

AKRAM ZAYTAR is currently a Senior Applied Research Scientist with the Microsoft AI for Good Research Laboratory. His work leverages satellite imagery, computer vision, and data science to address some of the most pressing environmental challenges of our time, including climate change and food security.

Before joining Microsoft, he completed his Postdoctoral Research with IBM Research, where he focused on geospatial machine learning and applied research. During his time there, he worked on projects including crop mapping at national scales by fusing multiple satellite bands, quantifying the role of extreme weather events on agricultural yield variation, and post-processing ECMWF forecasts using probabilistic ML models. He has a strong interest in GeoSpatial machine learning, deep neural networks, foundation models, and self/weakly-supervised learning.

WANLI MA (Member, IEEE) received the M.S. degree in image and video communications and signal processing from the University of Bristol, Bristol, U.K., in 2020, and the Ph.D. degree in computer science and informatics from Cardiff University, Cardiff, U.K., in 2025. He is currently a Postdoctoral Researcher with the Department of Engineering, University of Cambridge, Cambridge, U.K. His research interests include computer vision, minimal supervision, and remote sensing image analysis.

RIBANA ROSCHER received the Dipl.-Ing. and Ph.D. degrees in geodesy from the University of Bonn, Bonn, Germany, in 2008 and 2012, respectively. From 2015 to 2022, she was a Junior Professor of remote sensing with the University of Bonn. From 2022 to 2025, she was a Professor of Data Science for Crop Systems with the University of Bonn and led the Data Science for Crop Systems Group with the Institute of Bio- and Geosciences (IBG)-2, Forschungszentrum Jülich, Jülich, Germany. Since 2025, she has been a Professor of machine learning in agriculture with the Institute of Geodesy and Geoinformation, University of Bonn.

RONNY HÄNSCH (Senior Member, IEEE) received the bachelor's, M.Sc., and Ph.D. degrees from TU Berlin, Berlin, Germany, in 2005, 2007, and 2014, respectively. He is currently a Postdoctoral Research Fellow, Department SAR Technology German Aerospace Center (DLR), Oberpfaffenhofen, Germany. Since 2019, he has been a Postdoctoral Research Fellow with the Department SAR Technology, German Aerospace Center (DLR), Oberpfaffenhofen. He also serves as the lead of the "Machine Learning" Team with the Department of SAR Technology, German Aerospace Center, Cologne, Germany. His research interests include advanced machine learning techniques, including deep learning and random forests, applied to remote sensing with a particular emphasis on polarimetric synthetic aperture radar imagery, and innovative domains, such as self-supervised learning and cross-modal learning.

CALEB ROBINSON received the Ph.D. degree from Georgia Institute of Technology, Atlanta, GA, USA, in 2020. He is currently a Principal Research Science Manager with the Microsoft AI for Good Research Laboratory. His work focuses on tackling large-scale problems at the intersection of remote sensing and machine learning/computer vision. At the AI for Good Laboratory, he co-leads the Geospatial ML research group and is the Lead Researcher on the Global Renewables Watch, rapid damage assessment, and global building density estimation teams.

ANTHONY ORTIZ received the Bachelor of Science degree in telematics engineering (electrical and computing engineering equivalent) from the Pontificia Universidad Catolica Madre y Maestra (PUCMM), Dominican Republic with Summa Cum Laude distinction, and the Ph.D. degree in computer science with The University of Texas at El Paso (UTEP), where he was advised by Dr. Olac Fuentes and Dr. Christopher Kiekintveld.

He is currently a Principal Research Science Manager with the Microsoft AI for Good Research Laboratory led by Juan Lavista Ferres. His work focuses on tackling large scale problems in medical imaging, remote sensing, machine learning, and computer vision in support of Microsoft's AI for Earth and AI for Health initiatives. During his Ph.D., he spent several months at Mila in Montreal working under the guidance of Prof. Yoshua Bengio. He also spent time at Microsoft Research as an AI Research Intern working with Nebojsa Jojic and Dan Morris, Orbital Insight, and USC Institute for Creative Technologies. His research interests include intersection between machine learning, computer vision, and remote sensing with particular interest in the application of artificial intelligence to solve problems affecting society. He is also interested in self-supervised learning, conditional computation, out of distribution (OOD) generalization, and anything related to deep neural networks.

SIMONE FOBI received the M.S. degree from Stanford University and the Ph.D. degree in mechanical engineering from Columbia University. She is a Senior Applied Research Scientist at the Microsoft AI for Good Research Laboratory, where she develops geospatial and computer vision methods to tackle challenges in climate change, food security, energy, and humanitarian resilience. Her research interests span geospatial foundation models, self-supervised learning, remote sensing, and energy analytics.

RAHUL DODHIA received the Ph.D. degree in cognitive psychology from Columbia University, New York, NY, USA. As a Deputy Director of Microsoft's AI for Good Laboratory, he leads the development of AI solutions for ecological conservation, medical imaging, and geospatial applications in disaster response, agriculture, and deforestation. He is passionate about addressing social, environmental, and infrastructure challenges in the Global South, ensuring equitable access to AI-driven innovation. His experience includes roles at NASA, Amazon, and several technology startups.

JUAN M. LAVISTA FERRES received the degree in data mining and machine learning from Johns Hopkins University, Baltimore, MD, USA, and the dual degrees in computer science from the Catholic University of Uruguay, Montevideo, Uruguay, and the Ph.D. degree in AI for healthcare from Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. He is currently the Corporate Vice President and the Chief Data Scientist with Microsoft, where he leads the AI for Good Laboratory—a global initiative leveraging artificial intelligence to address some of the world's most pressing challenges. His work spans critical domains including sustainability, healthcare, humanitarian response, accessibility, and digital equity. In 2024, he was named one of the 100 most innovative people in Latin America by Bloomberg. He began his Microsoft career in 2009, working on the Microsoft Experimentation Platform (EXP) and later leading data science efforts for the Bing Data Mining team. He also initiated Microsoft's research into Sudden Infant Death Syndrome (SIDS), with results published in Pediatrics. In addition to his corporate role, he is editor of the Microsoft Journal of Applied Research (MSJAR), a Faculty Member with Singularity University, and a frequent speaker at forums, such as TEDx, IEEE, and institutions including Cornell and UC Berkeley. He resides in Kirkland, Washington. In 2024, he co-authored the book *AI for Good: Applications in Sustainability, Humanitarian Action, and Health*, which explores real-world applications of AI in solving global challenges.

OKTAY KARAKUŞ (Member, IEEE) received the B.Sc. degree (Hons.) in electronics engineering from Istanbul Kültür University, Istanbul, Türkiye, in 2009, and the M.Sc. and Ph.D. degrees in electronics and communication engineering from İzmir Institute of Technology (IZTECH), Urla, Türkiye, in 2012 and 2018, respectively. From 2009 to 2018, he held a research assistant positions with several universities, in Türkiye. In 2017, he was a Visiting Scholar with the Institute of Information Science and Technologies (ISTI-CNR), Pisa, Italy. From 2018 to 2021, he was a Research Associate of image processing with the Visual Information Laboratory, University of Bristol, U.K. He is currently a Lecturer with the School of Computer Science and Informatics, Cardiff University, where he is the Deputy Director of the Cardiff Data Science Academy and leads the Remote Sensing Image and Data Analysis (ReSIDA) Research Group. His research interests include remote sensing image analysis, environmental data science for marine and coastal sciences, ecology and wildlife, machine learning and AI, and football analytics. He is on the editorial board of multiple reputable journals and has published in high-impact venues in remote sensing, environmental monitoring, and artificial intelligence.

PAUL L. ROSIN is currently a Professor with the School of Computer Science and Informatics, Cardiff University. He has worked on many aspects of computer vision over the last 40 years, covering both fundamental algorithms in areas, such as low level image processing, performance evaluation, shape analysis, facial analysis, medical image analysis, surveillance, 3-D mesh processing, cellular automata, and non-photorealistic rendering, as well as multidisciplinary collaborations, such as: determining the effectiveness of surgery from facial morphology and temporal dynamics, the perception of trustworthiness from smiles, segmentation of 3-D OCT scans of retinas, interpreting lava flows, identification of desmids and otoliths, analyzing the effects of alcohol on crowd dynamics and violence, and digitally unrolling of fragile parchments from 3-D X-ray scans.