

Article

Empirical Evidence Regarding Few-Shot Learning for Scene Classification in Remote Sensing Images

Valdivino Alexandre de Santiago Júnior 

Laboratório de Inteligência Artificial para Aplicações AeroEspaciais e Ambientais (LIAREA), Coordenação de Pesquisa Aplicada e Desenvolvimento Tecnológico (COPDT), Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, São Paulo 12227-010, Brazil; valdivino.santiago@inpe.br

Abstract: Few-shot learning (FSL) is a learning paradigm which aims to address the issue of machine/deep learning techniques which traditionally need huge amounts of labelled data to work out. The remote sensing (RS) community has explored this paradigm with numerous published studies to date. Nevertheless, there is still a need for clear pieces of evidence on FSL-related issues in the RS context, such as which of the inference approaches is more suitable: inductive or transductive? Moreover, how does the number of epochs used during training, based on the meta-training (base) dataset, relate to the number of unseen classes during inference? This study aims to address these and other relevant questions in the context of FSL for scene classification in RS images. A comprehensive evaluation was conducted considering eight FSL approaches (three inductive and five transductive) and six scene classification databases. Some conclusions of this research are as follows: (1) transductive approaches are better than inductive ones. In particular, the transductive technique Transductive Information Maximisation (TIM) presented the best overall performance, where in 20 cases it got the first place; (2) a larger number of training epochs is more beneficial when there are more unseen classes during the inference phase. The most impressive gains occurred particularly considering the AID (6-way) and RESISC-45 (9-way) datasets. Notably, in the AID dataset, a remarkable 58.412% improvement was achieved in 1-shot tasks going from 10 to 200 epochs; (3) using five samples in the support set is statistically significantly better than using only one; and (4) a higher similarity between unseen classes (during inference) and some of the training classes does not lead to an improved performance. These findings can guide RS researchers and practitioners in selecting optimal solutions/strategies for developing their applications demanding few labelled samples.



Citation: Santiago Júnior, V.A.d. Empirical Evidence Regarding Few-Shot Learning for Scene Classification in Remote Sensing Images. *Appl. Sci.* **2024**, *14*, 10776. <https://doi.org/10.3390/app142310776>

Academic Editor: Douglas O'Shaughnessy

Received: 13 October 2024

Revised: 7 November 2024

Accepted: 13 November 2024

Published: 21 November 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: few-shot learning; remote sensing; image scene classification; deep learning; statistical tests

1. Introduction

Some learning paradigms aim to avoid relying on a significant amount of labelled samples. Among them, weakly supervised learning [1,2], semi-supervised learning [3,4], and self-supervised learning [5–7] have more recently attracted the attention of the artificial intelligence (AI) community in the context of deep learning (DL). However, in practice, the traditional supervised learning [8,9] can still be considered the most popular paradigm of all as it has existed since the early days of machine learning (ML). However, the issue of supervised learning is precisely the usual need for a huge amount of labelled samples so that models can produce a satisfactory performance.

Another attempt to address this necessity of a huge amount of labelled data is few-shot learning (FSL) [10,11]. FSL leverages prior knowledge to quickly generalise to new tasks that have only a few samples with supervised information. There is a huge interest in FSL since few-shot tasks can be defined with a few labelled examples per class (typically one or five samples), the *support set* S , and a *query set* Q with unlabelled examples. Thus, sampled few-shot tasks are derived based on the test dataset (depending on the situation, it can also

be from the validation and training sets) where each K -way N_s -shot task involves sampling N_s labelled samples from each of the K different classes, also chosen at random.

Regarding FSL, two points should be stressed. Firstly, there are basically two types of training: episodic and classical. In the episodic case, training is structured in a series of learning problems (or episodes), and hence the model is trained on such a series of episodes where each episode simulates the FSL scenario with corresponding support S and query Q sets [12]. This closely mimics the inference phase within the FSL context, where the model will encounter few-shot tasks with limited supervised samples. Classical training is the standard supervised learning approach where models are trained on labelled datasets to learn fixed parameters for classifying data. There are no episodes or few-shot tasks during the training phase in this type of training.

Secondly, the type of inference can be inductive or transductive. In inductive inference within FSL, models are trained on the base set with a large number of examples, and then adaptation happens to new tasks with only a few labelled samples. Meta-learning and metric-based learning are generally considered types of inductive inference approaches within FSL. In the transductive case, at the inference time, models predict the class label jointly for all the unlabelled query samples, rather than for one sample/episode at a time [13]. In other words, in this situation, all unlabelled query examples of a few-shot task are classified simultaneously, unlike inductive methods that process samples individually.

The remote sensing (RS) community has also paid attention to FSL [14–26]. Handling synthetic aperture radar (SAR) [16,17,25] and hyperspectral [21,22] images as well as addressing the computer vision tasks of scene classification [15,19,20], object detection [23,24], and semantic segmentation [14] are just a few directions where FSL has been used by such a community. A brief search conducted in August 2024 on the LENS.ORG portal [27] using the search string ‘few-shot learning’ AND ‘remote sensing’, considering the period of August 2015 to August 2024, yielded approximately 450 publications. Venues more related to the RS field with a few of them being part of the general field of AI were taken into account. This shows the growth of FSL within the RS community.

Despite these numerous studies, there is still a lack of evidence supporting the suitability of FSL techniques within DL applied for RS. Given the increasing significance of AI/DL across various fields such as medicine [28], agriculture [29], financial applications [30], autonomous driving [31], and naturally RS, several important questions remain unanswered, particularly for scene classification in RS images. Some of these research questions (RQs) are as follows:

1. RQ_1—Which of the inference approaches is more suitable: inductive or transductive?
2. RQ_2—Considering classical training, how does the number of epochs used during training, based on the meta-training (base) dataset, relate to the number of unseen classes during inference?
3. RQ_3—Is relying on 5-shot tasks statistically significantly better than 1-shot ones?
4. RQ_4—Would a higher similarity between unseen classes (during inference) and some of the existing classes in the base training dataset improve technique performance?

To support the arguments above, some recent publications proposing FSL approaches for scene classification in RS images [15,19,20,32–38] are presented in Section 2.1. As detailed in this section of related work, these approaches focus on proposing new FSL techniques, and the articles present comparisons to various other studies. However, there is no emphasis on addressing whether, in general, inductive or transductive techniques are better. The main point is, of course, to demonstrate that the proposed technique performs better compared to several others. To this end, RQ_1 was proposed.

Furthermore, the number of interactions (epochs/episodes) is fixed during the training phase. Unlike what is proposed in RQ_2, there are no variations in the number of epochs during training to observe how this relates to the number of unseen classes during the inference phase. The number of unseen classes during inference is fixed at five ($K = 5$). That is, in all the articles, the tasks are one of these two types: 5-way 1-shot or 5-way 5-shot.

This differs from what is proposed in this research, where K varies from smaller values (2-way) to larger ones (9-way).

Although comparisons are made to various other FSL approaches, which is quite interesting, the values presented are usually accuracies without the application of statistical tests to provide a more solid foundation for the superiority of the proposed techniques. This explains the motivation for RQ_3. Finally, none of these recent studies investigated whether a greater similarity between unseen classes (during inference) and some of the existing classes in the training base set actually improves the performance of the approaches, as asked in question RQ_4. To the best of our knowledge, even older FSL approaches for RS image scene classification did not investigate this aspect either.

By properly answering these questions, a valuable contribution can be given to the RS community where some pieces of evidence point to the most suitable solutions to be used in practice. This is precisely the objective of this study, where we aim to address some relevant questions in the context of FSL for RS image scene classification. A comprehensive evaluation was conducted considering eight FSL techniques where three are inductive, Prototypical Networks (PNs) [39], SimpleShot (SS) [40], and Fine-Tuning (FT) [41], and five are transductive, Bias Diminishing Cosine Similarity-based Prototypical Network (BD-CSPN) [42], LaplacianShot (LS) [43], Power Transform Maximum A Posteriori (PT-MAP) [44], Transductive Information Maximisation (TIM) [45], and Transductive Fine-Tuning (TFT) [46]. Moreover, six RS image scene classification datasets were considered in different unseen class configurations (K -way). EuroSAT [47] and XAI4SAR [48] were used with a lower number of unseen classes (2-way), UC Merced [49] and WHU-RS19 [50] were used with a medium number of unseen classes (4-way), and a higher number of unseen classes were taken into account for AID (6-way) [51] and RESISC-45 (9-way) [52]. With these different settings, we were able to achieve the goal of conducting a comprehensive evaluation to address the proposed research questions. Such a comprehensive evaluation is therefore the main contribution of this research, which sought to obtain detailed answers to each of the questions.

This article is structured as follows. Background and related work are in Section 2. Section 3 presents the materials and methods, including the research questions. Section 4 shows the results and some discussions based on the findings are presented in Section 5. Section 6 concludes the article and presents future directions.

2. Background

FSL can be formally described as follows [45]. Let $X_{base} = \{x_i, y_i\}$ be a labelled training set where x_i denotes the raw features of sample i ; y_i is its corresponding label, $i = 1 \dots N_{base}$; and N_{base} is the number of classes in this training set. This dataset X_{base} is usually referred to as the *meta-training* or *base* dataset. Let Y_{base} be the set of classes for X_{base} . In the FSL context, there is a test set $X_{test} = \{x_i, y_i\}$ where $i = 1 \dots N_{test}$ and N_{test} is the number of classes in Y_{test} , i.e., the set of classes for X_{test} . Hence, the following situation is valid: $Y_{base} \cap Y_{test} = \emptyset$. In other words, the classes in the test dataset presented during the inference phase are completely unseen (new) to the model trained on the base dataset. Thus, randomly sampled few-shot tasks are created based on the test set where each K -way N_S -shot task involves sampling N_S -labelled samples from each of the K different classes, also chosen at random. Set S is the support set with these labelled samples with size $|S| = N_S \times K$. Moreover, a few-shot task has also a query set Q with $|Q| = N_Q \times K$ unlabelled (unseen) examples from each of the K classes. FSL techniques leverage base set-trained models and labelled support sets to quickly adapt to new tasks. Their performance is assessed on unlabelled query sets.

Figure 1 shows an example of an arrangement for a 2-way 1-shot few-shot task based on images of the RESISC-45 set [52]. The classes *forest* and *beach* are unknown to the model at inference time, meaning the model had no prior knowledge of samples from these classes during the training phase. In this case, there are two unseen classes (2-way) and only one labelled sample per class (1-shot) in the support set. The query set consists of three samples

randomly selected from each unseen class. The expectation is that an FSL approach, which has not seen any samples from the classes *forest* and *beach* during training, can correctly classify them during the inference phase.

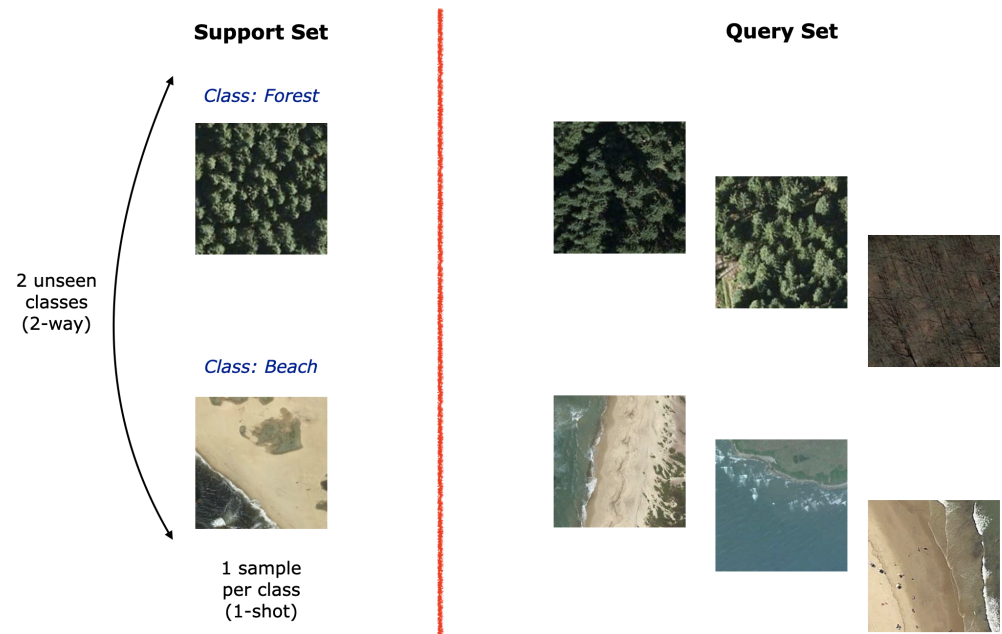


Figure 1. An arrangement for a 2-way 1-shot few-shot task.

2.1. Related Work

This section presents some recent studies related to this research. In [15], the authors proposed a ranking-preserving Knowledge Distillation (KD) approach that trains a student network to replicate the ranking of support images produced by teacher networks. By combining multi-teacher KD with learning-to-rank, they developed a new distillation loss based on Plackett–Luce distributions, resulting in a novel few-shot classification model termed the ranking network. As for the experimental part, the datasets used were RESISC-45 [52] and AID [51]. They relied on episodic training while our study chose classical training. Moreover, tasks are of the traditional types: 5-way 1-shot or 5-way 5-shot. To answer the proposed questions, our approach considered a wide variation of unseen classes during the inference phase: from smaller values (2-way) to larger ones (9-way).

In [19], a Multi-Grained Global-Local Semantic Feature Fusion (MGGL-SFF) method for FSL remote sensing scene classification, addressing the challenge of complex scenes with hierarchical and coupled spatial relations (e.g., internal and external spatial contexts) that impede feature extraction, was proposed. The MGGL-SFF method effectively combines global discriminative spatial semantic features with local transferable fragment features, establishing a robust prototype representation for FSL. The RESISC-45 was the only dataset used. It is important to note that while the authors selected 15 novel classes for the test set, the experimental setup was limited to evaluating performance on only five of these classes. Thus, tasks are again 5-way 1-shot or 5-way 5-shot. The authors did not apply statistical tests to have conclusions based on more solid foundations.

To create the Scene Graph Matching Network for Few-Shot Remote Sensing Scene Classification (SGMNet) [20], the authors took into account FSL natural image classification methods. However, they argued that such methods overlook two distinctive features of RS images: (1) object co-occurrence, where multiple objects frequently appear together within a scene, and (2) object spatial correlation, where these co-occurring objects follow specific spatial patterns. To leverage these features, the SGMNet approach was then proposed. This framework includes a scene graph construction module to represent each

test image or scene class as a scene graph, with nodes capturing object co-occurrences and edges representing spatial correlations. A scene graph matching module then evaluates the similarity between each test image and each scene class. Regarding the experiments, RESISC-45, AID, UC Merced [49], and WHU-RS19 [50] were the selected datasets. In our study, in addition to these four sets, EuroSAT [47] and XAI4SAR [48] were also used to have a more significant number of sets of RS images. In the meta-training phase, they considered episodic training rather than the classical one. Few-shot tasks are once more of the types 5-way 1-shot or 5-way 5-shot, and no statistical tests were conducted.

Another FSL approach for RS image scene classification was proposed in [32]. It is a class distribution learning method using support samples via an intra-class feature aggregation (IFA) module, intra-class feature homogenisation (IFH) module, and inter-class cross-calibration (ICC) module. The IFA and IFH modules promote the aggregation and even distribution of intra-class features around model-generated prototypes, reducing intra-class distances and preventing extreme samples near decision boundaries. The selected datasets were RESISC-45, UC Merced, and WHU-RS19. The same previous remarks as before regarding the tasks (5-way 1-shot or 5-way 5-shot) and no statistical tests were conducted.

High interclass similarity in RS images can lead to classification confusion. To address this issue, the authors of [33] created a double discriminative constraint-based affine non-negative representation for FSL RS scene classification. Specifically, a novel representation-based classifier with two discriminative constraint terms and affine non-negative constraints was introduced to reduce interclass correlation and enhance class-specific parameter learning. Episodic training was the option for this study and, as for the experiments, RESISC-45 and RSD46-WHU [53] were the sets of images. The same types of few-shot tasks were considered, as in the previous studies, and no sign of statistical tests appear in the article.

Other recent FSL approaches published in the literature for scene classification include an Iterative Distribution Learning Network (IDLN) [34], Manifold Augmentation-based Self-supervised Contrastive Learning under the meta-learning framework [35], a framework based on Metric Learning and Local Descriptors (MLLD) [36], a Multi-Scale Interaction Prototypical Network [37], and a Data Augmentation Technique based on Distortion Magnitude Optimisation [38]. The RS datasets used are some of those previously mentioned in this section, and the same observations which have already been stated hold here.

To reiterate the points made in Section 1, despite these interesting solutions, the following differences exist between our study and theirs: (1) these approaches introduce new FSL techniques and compare them to various strategies. However, they do not focus on determining whether inductive or transductive techniques are generally superior; (2) the number of interactions (epochs/episodes) during training is not varied as is proposed in this study, aiming to observe how this relates to the number of unseen classes during the inference phase; (3) in all these articles, the number of unseen classes during inference is fixed at five ($K = 5$). Few-shot tasks are one of these two types: 5-way 1-shot or 5-way 5-shot; (4) no statistical tests were applied; and (5) none of the studies have explored the impact of similarity between unseen classes during inference and some training set classes on model performance.

3. Materials and Methods

Figure 2 presents the workflow (activity diagram) of the method proposed for this study. The first step is to define the research questions in a way that analyses topics not yet addressed in the context of FSL for scene classification in RS images. The selection of proper RS datasets followed by the choice of some suitable and recent FSL approaches are the next activities. The experiment design with all the established options, as well as the selection of metrics (it can be a single metric too), are the next steps. The analysis of the results of the executed experiment, the answers to all research questions, and the conduct of an additional discussion conclude the method proposed in this work. The next sections will describe all these activities in detail.

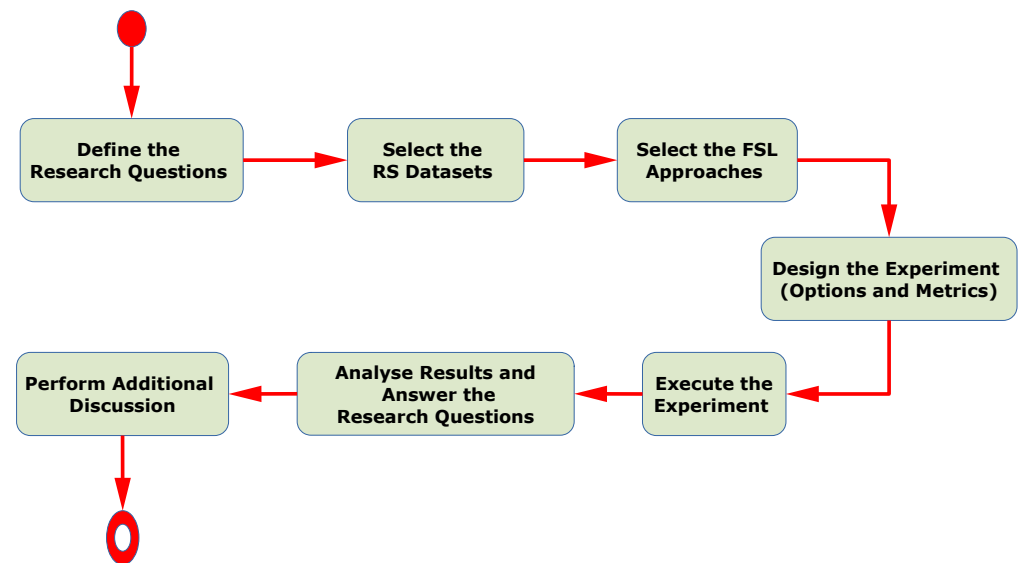


Figure 2. The workflow of the method related to this study.

3.1. Research Questions

The research questions (RQs) associated with this study were already mentioned in Section 1. They are repeated below for completeness:

1. RQ_1—Which of the inference approaches is more suitable: inductive or transductive?
2. RQ_2—Considering classical training, how does the number of epochs used during training, based on the meta-training (base) dataset, relate to the number of unseen classes during inference?
3. RQ_3—Is relying on 5-shot tasks statistically significantly better than 1-shot ones?
4. RQ_4—Would a higher similarity between unseen classes (during inference) and some of the existing classes in the base training dataset improve technique performance?

In Sections 1 and 2.1, the motivations for the RQs and how this study differs from others in the literature were already presented. However, the general reasoning behind the creation of these RQs is emphasised here. With RQ_1, the goal is to understand which of the two inference approaches, inductive or transductive, is the best so that the answer to this question can motivate the development/use of solutions of a certain type in the context of the RS community. Note that several previous studies claim that transductive approaches are in general better than inductive ones [43,45], but the datasets used do not contain RS images. Thus, the idea is to see if this conclusion generalises to the RS domain. RQ_2 can be seen as a question related to the “reduction of the reduction” of computational demand. As discussed in Section 1, while FSL minimises the need for a large number of labelled samples, there is still a training phase (whether episodic or classical training). The aim of this question is to explore whether the training process can be made even less demanding and how this reduction relates to the number of unseen classes during the inference phase.

It seems natural to expect a better performance from a technique with 5-shot tasks than with 1-shot ones. But is this improved performance statistically significant? RQ_3 addresses this perspective. Let us assume that the set of unseen classes during the inference phase in scenario A has a greater similarity to some classes from the base dataset (training phase) than in scenario B. Does this mean that techniques will perform better in scenario A? This is the focus of RQ_4.

3.2. Datasets

This section briefly describes the six RS scene classification datasets selected for this research. The number of unseen classes (K -way) during the inference phase was varied.

EuroSAT [47] and XAI4SAR [48] were used with a 2-way setting, which is a low number of unseen classes during the inference phase. Sentinel-2 images were used to create EuroSAT, a dataset encompassing 13 spectral bands and featuring 10 classes, with a total of 27,000 labelled and geo-referenced images. The resolution of the images (patches) is 64×64 pixels. XAI4SAR consists of Sentinel-1 SAR image patches, including both single look complex (SLC) and ground range detected (GRD) products. Seven sea-ice types are annotated, with GRD image patches having a resolution of 256×256 pixels. Additionally, a Gaofen-3 SAR scene covering a large urban area was used, with seven land use and land cover (LULC) classes annotated. Figure 3 presents samples of some classes from the EuroSAT and XAI4SAR datasets.

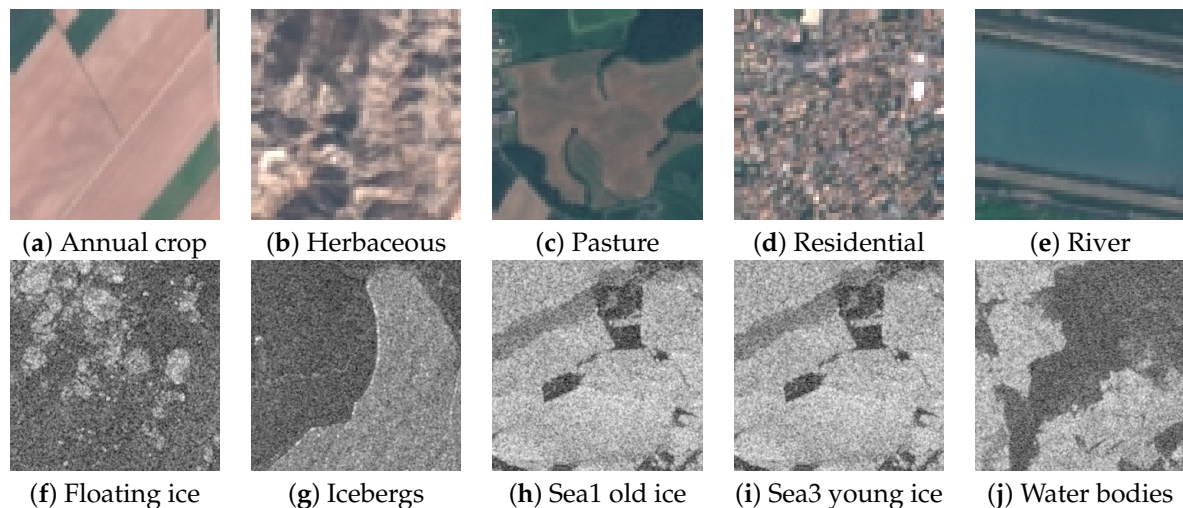


Figure 3. Samples from EuroSAT (top row: (a–e)) and XAI4SAR (bottom row: (f–j)) datasets.

In terms of a medium number of unseen classes, the UC Merced [49] and WHU-RS19 [50] datasets were considered with a 4-way configuration. UC Merced is a set with 21 classes where there are 100 images for each of the 21 classes. Each image (patch) has a resolution of 256×256 pixels. The images were manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the USA. WHU-RS19 is a collection of satellite images sourced from Google Earth, providing high-resolution imagery with a spatial resolution of up to 0.5 m. The dataset encompasses 19 distinct scene classes including *airport*, *beach*, *bridge*, *commercial*, *desert*, *farmland*, *football field*, *forest*, and others. For each class, there are about 50 samples. The dataset includes images from various regions and resolutions, leading to potential differences in scale, orientation, and illumination within each class. Figure 4 presents samples of some classes from the UC Merced and WHU-RS19 datasets.

With AID [51] and RESISC-45 [52], the option was to have a higher number of unseen classes: 6-way and 9-way, respectively. AID is a large-scale dataset designed to push forward image scene classification within RS. AID consists of over 10,000 aerial scene images that were carefully collected and annotated. Sample images were collected from Google Earth imagery, and there are 30 aerial scene types: *airport*, *bare land*, *baseball field*, *beach*, *bridge*, *centre*, *church*, *commercial*, *dense residential*, *desert*, *farmland*, *forest*, *industrial*, *meadow*, *medium residential*, *mountain*, *park*, *parking*, *playground*, *pond*, *port*, *railway station*, *resort*, *river*, *school*, *sparse residential*, *square*, *stadium*, *storage tanks*, and *viaduct*. Image resolution is 600×600 pixels.

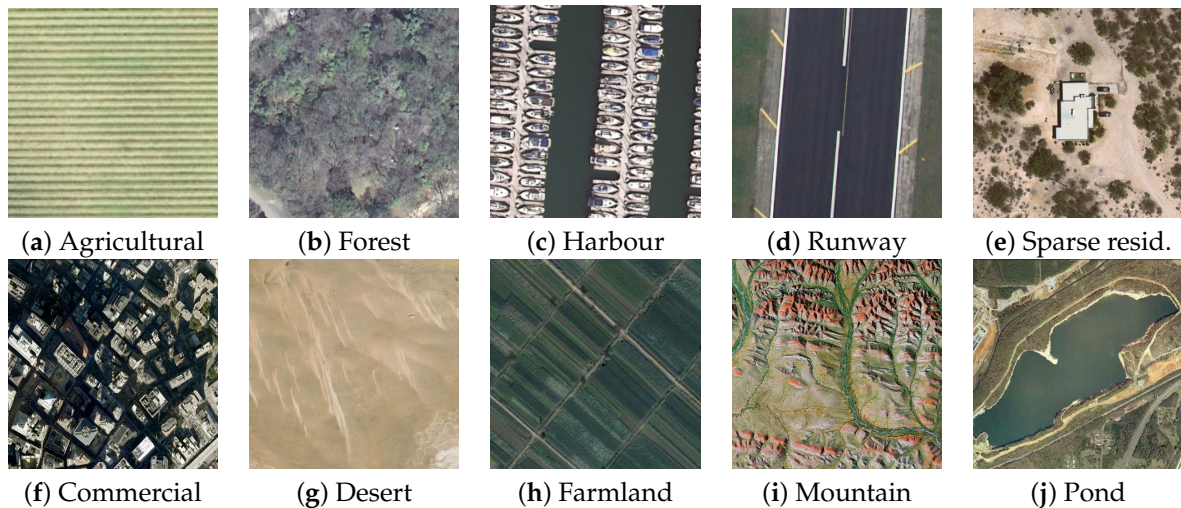


Figure 4. Samples from UC Merced (top row: (a–e)) and WHU-RS19 (bottom row: (f–j)) datasets. Caption: resid. = residential.

RESISC-45 is another large-scale dataset for RS image scene classification. Such a dataset includes 31,500 images spanning 45 scene classes, with 700 images per class. RESISC-45 is characterised by its large scale, both in terms of scene classes and total image count, and significant variations in factors such as translation, spatial resolution, viewpoint, object pose, illumination, background, and occlusion. Image resolution (dimension) is 256×256 pixels. Figure 5 presents samples of some classes from the AID and RESISC-45 datasets.

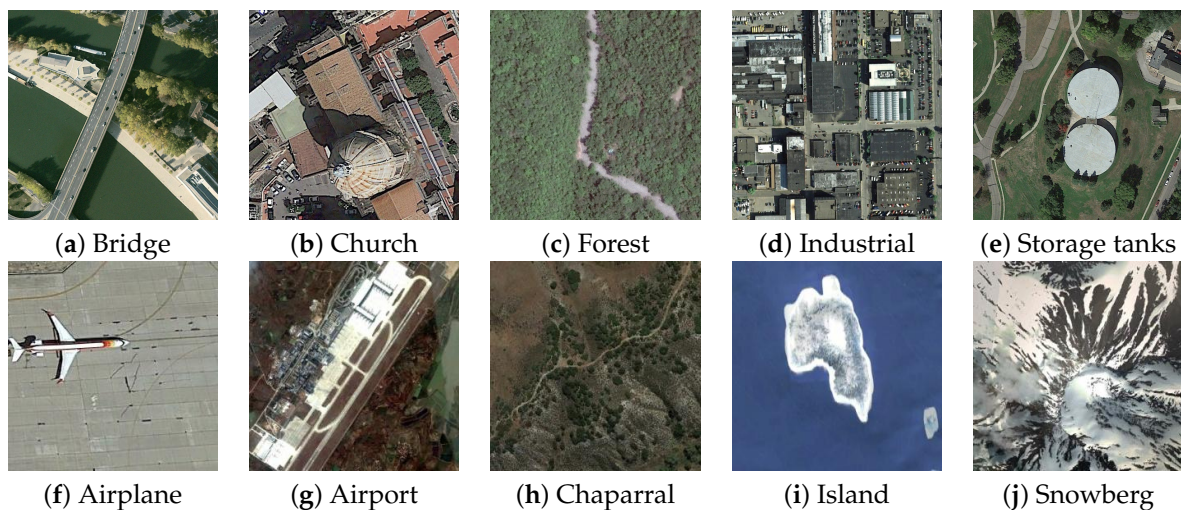


Figure 5. Samples from AID (top row: (a–e)) and RESISC-45 (bottom row: (f–j)) datasets.

As remarked in Section 2.1, the RESISC-45, AID, WHU-RS19, and UC Merced datasets are very traditional and they have been used in many studies. This explains the reason behind their selection. EuroSAT was chosen due to the lower resolution (dimension) of its images, and XAI4SAR was aimed at considering a different type of image, i.e., SAR images.

3.3. FSL Approaches

This section presents the selected FSL strategies for this study. It is important to state that some of these techniques are more traditional while others were developed more recently.

3.3.1. Inductive Approaches

The first of the three selected inductive techniques is Prototypical Networks (PNs), which learn a metric space where classification is performed by calculating the distances to prototype representations of each class [39]. Compared to other FSL techniques, PNs incorporate a simpler inductive bias, which can be advantageous in scenarios with limited data. PNs are thus based on the idea that there exists an embedding in which points cluster around a single prototype representation for each class.

In SimpleShot (SS), the authors challenge a common view to mitigate overfitting where FSL approaches extract image features using a convolutional neural network (CNN) and use a combination of meta-learning and nearest-neighbour classification to perform the recognition [40]. They thus showed that nearest-neighbour classifiers can achieve very good results without meta-learning. According to the authors, applying basic feature transformations to the features prior to nearest-neighbour classification results in a highly competitive performance.

In [41], the authors presented a thorough comparative analysis of many FSL approaches with results showing that deeper backbones significantly decrease the performance differences among techniques on datasets with limited domain differences. They also showed a modified baseline strategy that achieves a competitive performance when compared to other approaches on both non-RS datasets mini-ImageNet [54] and CUB-200-2011 [55]. Their baseline model follows the standard transfer learning procedure of network pre-training and fine-tuning. This approach is denoted here as Fine-Tuning (FT).

3.3.2. Transductive Approaches

Regarding the five transductive approaches, Bias Diminishing Cosine Similarity-based Prototypical Network (BD-CSPN) [42] was the first selected. The motivation for their research was the fact that PNs trained on a narrow-size distribution of scarce data usually tend to obtain biased prototypes. They then identified two major factors influencing such a process: intra-class bias and cross-class bias. To address these, a straightforward yet effective technique for prototype rectification in a transductive setting was proposed. This approach leverages label propagation to reduce intra-class bias and feature shifting to mitigate cross-class bias. Their goal was to find the expected prototypes that maximise cosine similarity with all data points within the same class. To achieve this, CSPN was proposed, which is designed to extract discriminative features and compute fundamental prototypes from limited samples.

In [43], the authors presented a transductive Laplacian-regularised inference for few-shot tasks. Given any feature embedding learned from the base classes (classes of the base dataset), they minimise a quadratic binary-assignment function containing two terms: (1) a unary term assigning query samples to the nearest class prototype, and (2) a pairwise Laplacian term encouraging nearby query samples to have consistent label assignments. Their transductive inference does not re-train the base model and can be seen as a graph clustering of Q , subject to supervision constraints from S . This technique is called here as LaplacianShot (LS).

With Power Transform Maximum A Posteriori (PT-MAP) [44], the authors addressed the challenge of transfer-based FSL with a twofold approach: (1) preprocessing the data extracted from the backbone to align them with a specific distribution (e.g., Gaussian-like), and (2) exploiting this distribution using a carefully designed algorithm based on maximum a posteriori estimation and optimal transport, applicable specifically in the transductive few-shot setting.

Transductive Information Maximisation (TIM) [45] maximises the mutual information between the query features and their label predictions for a certain few-shot task, in conjunction with a supervision loss based on S . Moreover, the authors also derived a new alternating-direction solver for the mutual-information loss, which substantially accelerates transductive inference convergence over gradient-based optimisation while

yielding a similar performance. TIM inference is modular and can be applied on top of any base-trained feature extractor.

To propose Transductive Fine-Tuning (TFT) [46], the authors claimed that fine-tuning a deep neural network trained using standard cross-entropy loss serves as a strong baseline for FSL. When fine-tuned in a transductive manner, it surpasses up-to-date FSL approaches on non-RS datasets such as mini-ImageNet [54], tiered-ImageNet [56], and others, all while using the same hyperparameters.

It is important to emphasise that all such solutions, whether inductive or transductive, have been proposed and evaluated on diverse datasets but not with RS images/patches. Hence, one of the relevant points of this research is precisely to realise their performances considering the characteristics of several RS datasets.

3.4. Experiment Design Options and Metric

This section describes all the design options defined to answer the proposed RQs, as well as the metric chosen to measure the performance of the FSL techniques. In fact, a series of experiments/evaluations were carried out, but it was decided to call them an “experiment” for the sake of simplicity. Firstly, classical training was preferred over episodic training. Although there are some arguments against classical training saying it is less effective for FSL because the model is optimised to perform well on tasks with a lot of training data, and may struggle when faced with new tasks with only a few samples, some authors also state some issues of episodic training. For instance, in [12], the researchers investigated the usefulness of episodic training/learning in methods which use non-parametric approaches, such as nearest neighbours, and concluded that the constraints imposed by episodic learning are not necessary but that they in fact lead to a data-inefficient way of exploiting training batches. Moreover, using classical training is more attractive to practitioners from the RS community since this is based on standard supervised learning which, as mentioned before, can still be considered the most popular paradigm in practice.

To make the tasks more challenging for the FSL approaches, we dynamically cropped the images of all datasets to a resolution of 128×128 pixels. The only exception refers to EuroSAT since the resolution of the images/patches is 64×64 pixels, and thus we took this into account. Furthermore, as for the low (2-way; EuroSAT and XAI4SAR) and medium (4-way; UC Merced and WHU-RS19) number of unseen classes, 100 images per class were selected. But WHU-RS19 has less than 100 images per class. Thus, horizontal flipping was used and the set was then augmented so that it reached such a quantity of images per class. As for the high number of unseen classes (6-way for AID and 9-way for RESISC-45), 200 images per class were selected. Again, the idea is to make the models’ task more difficult, for example, by providing few total samples per class and thus making learning difficult in the training stage.

Each dataset was split into the three traditional sets: training, validation, and test. We used approximately 60% of the classes to compose the training set, 20% of the classes for the validation set, and 20% to form the test set. Since it is classical training, the training set was not sampled in the shape of few-shot tasks, but the validation and test datasets were. The number of few-shot tasks was 50 considering the validation set and 1000 for the test dataset. In terms of the number of samples in the support set, the options were $N_S = 1$ (1-shot) and $N_S = 5$ (5-shot), and recall that the cardinality of such a set is $|S| = N_S \times K$. There were 10 query images for each few-shot task. Table 1 summarises the information about datasets and few-shot tasks.

To answer RQ_2, the number of epochs during training was varied as follows: 10, 50, 100, and 200 epochs. Thus, we can grasp an idea about the relationship between the number of epochs used during training and the number of unseen classes during inference. As for RQ_3, we obtained the values of the metric for all 1-shot tasks, considering all datasets and the number of epochs, and compared them to the respective values obtained for the 5-shot tasks. Then, we applied the Shapiro–Wilk test with a significance level $\alpha = 0.05$ to check data normality. A quantile–quantile (Q–Q) plot was used to confirm the results due to the

Shapiro–Wilk test. Since in all cases the data happen to be not normal (see Section 4), the non-parametric Wilcoxon test, with the alternative *greater* and significance level $\alpha = 0.05$, was applied.

Table 1. Information about datasets and few-shot tasks. Caption: Cl = classes; Val = validation.

Dataset	#Cl Training	#Cl Val	#Cl Test	#Tasks Val	#Tasks Test
EuroSAT	6	2	2	50	1000
XAI4SAR	4	2	2	50	1000
UC Merced	13	4	4	50	1000
WHU-RS19	11	4	4	50	1000
AID	18	6	6	50	1000
RESISC-45	27	9	9	50	1000

To answer RQ 4, we saw in which dataset the approaches presented the worst performances. RESISC-45 was the most challenging, and this is somehow expected since here there is the highest number of unseen classes (nine). Thus, we obtained nine classes in the original training set, identified this subset *Or9*, and compared it to all other individual classes. Based on the structural similarity index measure (SSIM) [57], it was possible to realise which of the remaining classes are more similar to *Or9*. Thus, in the *More* scenario, the original test set was changed so that now it has the nine most similar classes to *Or9*. Naturally, some classes had to be moved from the original test set into the updated training or validation datasets. Conversely, in the *Less* scenario, the opposite happened: the nine less similar classes to *Or9* were identified, and now there is another different test dataset. Moreover, all FSL approaches with 10, 50, and 100 epochs using the RESISC-45 dataset were executed again, considering 30 few-shot validation and 200 few-shot test tasks in this case.

We used the implementation of the approaches as shown in the repository Easy Few-Shot Learning [58] with some adaptations. The backbone for all cases was ResNet-12 [59], and the optimiser was stochastic gradient descent (SGD) with learning rate = 0.1, momentum = 0.9, and weight decay = 5×10^{-4} . MultiStepLR was used to decay the learning rate of each parameter group by gamma = 0.1 once the number of epochs reaches one of the milestones: 150 and 180 epochs. Runnings were performed using the SDumont supercomputer [60]. Finally, the average accuracy across all few-shot test tasks was the metric used to determine the performance of the FSL solutions.

4. Results

This section presents the results that address each of the proposed RQs.

4.1. RQ_1

The average accuracies for training with 10 epochs are shown in Table 2 (the EuroSAT, XAI4SAR, and UC Merced datasets) and Table 3 (the WHU-RS19, AID, and RESISC-45 datasets). In **bold**, we see the best performance for each shot configuration of a task (1-shot, 5-shot) considering each dataset. For instance, SS was the best approach not only for 1-shot (0.512) but also for 5-shot (0.527) tasks considering the EuroSAT set. TIM was the best for both configurations (1-shot: 0.428; 5-shot: 0.543) taking into account the AID dataset. Moreover, in **bold** and **blue** we present the best result for all datasets. Thus, TIM obtained the highest average accuracy for the 1-shot tasks considering all datasets with 10 epochs: 0.694 with the XAI4SAR dataset. As for the 5-shot case, BS-CSPN and TIM tied with the highest value (0.787) with the WHU-RS19 dataset. It is clear that the transductive approaches presented better performances than the inductive counterparts. Only for the EuroSAT dataset did an inductive technique (SS) exhibit a superior performance. For all other cases, transductive approaches were superior, with TIM and PT-MAP standing out.

Table 2. Average accuracy, 10 epochs: EuroSAT, XAI4SAR, and UC Merced. As for 10 epochs and considering each dataset, the best performance for each shot configuration of a task is in **bold**. In **bold** and **blue**, the best result for all datasets is presented (1-shot tasks).

Approach	EuroSAT		XAI4SAR		UC Merced	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
PNs	0.509	0.526	0.618	0.674	0.571	0.687
SS	0.512	0.527	0.650	0.712	0.552	0.666
FT	0.510	0.518	0.670	0.767	0.548	0.659
BD-CSPN	0.501	0.520	0.656	0.705	0.557	0.663
LS	0.505	0.523	0.600	0.630	0.570	0.685
PT-MAP	0.492	0.507	0.654	0.748	0.655	0.742
TIM	0.510	0.524	0.694	0.769	0.557	0.674
TFT	0.509	0.526	0.618	0.679	0.571	0.687

Table 3. Average accuracy, 10 epochs: WHU-RS19, AID, and RESISC-45. As for 10 epochs and considering each dataset, the best performance for each shot configuration of a task is in **bold**. In **bold** and **blue**, the best result for all datasets is presented (5-shot tasks).

Approach	WHU-RS19		AID		RESISC-45	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
PNs	0.649	0.775	0.405	0.528	0.368	0.474
SS	0.674	0.782	0.427	0.537	0.389	0.480
FT	0.668	0.785	0.424	0.538	0.387	0.487
BD-CSPN	0.682	0.787	0.422	0.518	0.386	0.472
LS	0.670	0.785	0.415	0.506	0.381	0.465
PT-MAP	0.679	0.785	0.416	0.505	0.399	0.502
TIM	0.679	0.787	0.428	0.543	0.393	0.485
TFT	0.649	0.776	0.405	0.528	0.367	0.475

Tables 4 and 5 show the results for 50 epochs. PNs, an inductive approach, and TFT, a transductive one, were the best considering 5-shot tasks and the WHU-RS19 dataset. And their result (0.886) was the highest taking into account all datasets and 50 epochs. As for the 1-shot case, the highest of all values was due to LS (0.881) evaluated on the WHU-RS19. Again, transductive solutions were quite superior where, once again, TIM exhibited the best results in several cases, but BD-CSPN also performed significantly well in this scenario.

Table 4. Average accuracy, 50 epochs: EuroSAT, XAI4SAR, and UC Merced. As for 50 epochs and considering each dataset, the best performance for each shot configuration of a task is in **bold**.

Approach	EuroSAT		XAI4SAR		UC Merced	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
PNs	0.525	0.558	0.535	0.573	0.623	0.744
SS	0.527	0.563	0.577	0.662	0.659	0.760
FT	0.526	0.566	0.568	0.654	0.657	0.760
BD-CSPN	0.521	0.555	0.559	0.633	0.661	0.735
LS	0.513	0.540	0.520	0.543	0.652	0.722
PT-MAP	0.525	0.553	0.522	0.547	0.756	0.829
TIM	0.529	0.569	0.584	0.686	0.664	0.772
TFT	0.525	0.559	0.534	0.574	0.623	0.744

By increasing the number of epochs, the transductive techniques reaffirm their best performances. In the case of 100 epochs, see Tables 6 and 7, most of the best results were obtained by transductive approaches, where the best performance across all datasets was achieved by LS, both for 1-shot tasks (0.834) and for 5-shot tasks (0.896). The exception

was FT (inductive), tied with TIM, in the 5-shot tasks and the RESISC-45 dataset. LS and, once again, TIM were the highlights here. Furthermore, a pattern begins to emerge from this point. The WHU-RS19 dataset appears to be the “easiest” of all, where, proportionally and when analysing the shot task type separately, the best results are obtained. This can be explained by the low variability of images within the classes, as data augmentation was necessary in this case. On the other hand, the RESISC-45 dataset stands out as the “most difficult”, likely due to the larger number of unknown classes during the inference phase.

Table 5. Average accuracy, 50 epochs: WHU-RS19, AID, and RESISC-45. As for 50 epochs and considering each dataset, the best performance for each shot configuration of a task is in **bold**. In **bold** and **blue**, the best result for all datasets is presented (1-shot and 5-shot tasks).

Approach	WHU-RS19		AID		RESISC-45	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
PNs	0.774	0.886	0.570	0.722	0.422	0.594
SS	0.773	0.866	0.634	0.714	0.482	0.635
FT	0.775	0.870	0.634	0.775	0.481	0.638
BD-CSPN	0.807	0.871	0.658	0.779	0.490	0.635
LS	0.811	0.879	0.656	0.772	0.463	0.600
PT-MAP	0.749	0.833	0.481	0.631	0.391	0.515
TIM	0.782	0.874	0.640	0.779	0.487	0.639
TFT	0.774	0.886	0.602	0.762	0.422	0.594

Table 6. Average accuracy, 100 epochs: EuroSAT, XAI4SAR, and UC Merced. As for 100 epochs and considering each dataset, the best performance for each shot configuration of a task is in **bold**.

Approach	EuroSAT		XAI4SAR		UC Merced	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
PNs	0.538	0.590	0.633	0.722	0.669	0.816
SS	0.541	0.589	0.699	0.806	0.721	0.849
FT	0.540	0.595	0.680	0.805	0.721	0.852
BD-CSPN	0.536	0.581	0.688	0.803	0.734	0.847
LS	0.529	0.576	0.639	0.719	0.723	0.828
PT-MAP	0.552	0.606	0.620	0.707	0.777	0.854
TIM	0.545	0.595	0.719	0.815	0.728	0.858
TFT	0.538	0.590	0.632	0.724	0.669	0.816

Table 7. Average accuracy, 100 epochs: WHU-RS19, AID, and RESISC-45. As for 100 epochs and considering each dataset, the best performance for each shot configuration of a task is in **bold**. In **bold** and **blue**, the best result for all datasets is presented (1-shot and 5-shot tasks).

Approach	WHU-RS19		AID		RESISC-45	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
PNs	0.763	0.892	0.578	0.771	0.418	0.594
SS	0.783	0.882	0.599	0.770	0.467	0.619
FT	0.786	0.889	0.597	0.765	0.469	0.624
BD-CSPN	0.832	0.893	0.623	0.779	0.480	0.616
LS	0.834	0.896	0.640	0.784	0.463	0.592
PT-MAP	0.746	0.813	0.482	0.633	0.425	0.554
TIM	0.794	0.889	0.607	0.775	0.473	0.624
TFT	0.763	0.892	0.577	0.771	0.417	0.594

As for 200 epochs, see Tables 8 and 9, the same trend of performances is shown. The transductive approaches presented better results overall, with only in two cases an inductive technique, FT, being the best: 5-shot tasks with WHU-RS19 tied with BD-CSPN and TIM,

and 5-shot tasks with RESISC-45. TIM once again had some better performances, but the highlight here was BD-CSPN, which achieved the highest number of first-place rankings.

Table 9 shows the extreme results of the evaluation. The cells with green background mean that these are the best results considering all the number of epochs. Thus, with 200 epochs, the best values of average accuracies overall for both tasks occurred with the WHU-RS19 dataset. But, the cells with a yellow background and values in red show that PT-MAP with RESISC-45 presented the worst results for both tasks and with 200 epochs. Table 10 summarises the number of times a certain approach achieved the best result, as well as when it was the second best. While BD-CSPN obtained the highest number of first positions for 1-shot tasks, TIM was the most outstanding considering 5-shot tasks and obtained the highest number of second-best places for 1-shot ones. Altogether, we can state that TIM was the best of all approaches, followed by BD-CSPN, PT-MAP (but see the discussion later), and LS. All of these techniques are transductive. The first inductive technique appears in fifth place (FT). This conclusion confirms the results presented for other domains: transductive approaches outperform inductive ones in the context of RS image scene classification.

Table 8. Average accuracy, 200 epochs: EuroSAT, XAI4SAR, and UC Merced. As for 200 epochs and considering each dataset, the best performance for each shot configuration of a task is in **bold**.

Approach	EuroSAT		XAI4SAR		UC Merced	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
PNs	0.548	0.615	0.673	0.782	0.717	0.864
SS	0.554	0.621	0.704	0.803	0.754	0.869
FT	0.554	0.623	0.702	0.806	0.753	0.869
BD-CSPN	0.558	0.615	0.717	0.807	0.774	0.867
LS	0.551	0.603	0.694	0.777	0.774	0.869
PT-MAP	0.560	0.627	0.647	0.739	0.670	0.816
TIM	0.558	0.627	0.714	0.810	0.759	0.874
TFT	0.548	0.615	0.673	0.783	0.717	0.864

Table 9. Average accuracy, 200 epochs: WHU-RS19, AID, and RESISC-45. As for 200 epochs and considering each dataset, the best performance for each shot configuration of a task is in **bold**. In **bold** and blue, the best result for all datasets is presented. Moreover, the cells with green background mean that these are the best results considering all the number of epochs. The cells with a yellow background and values in red mean these were the worst results considering all the number of epochs.

Approach	WHU-RS19		AID		RESISC-45	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
PNs	0.783	0.908	0.660	0.827	0.467	0.660
SS	0.831	0.916	0.703	0.837	0.503	0.676
FT	0.835	0.919	0.704	0.838	0.503	0.679
BD-CSPN	0.868	0.919	0.750	0.847	0.526	0.674
LS	0.856	0.912	0.746	0.842	0.519	0.663
PT-MAP	0.739	0.806	0.366	0.534	0.235	0.390
TIM	0.839	0.919	0.709	0.840	0.507	0.679
TFT	0.783	0.908	0.660	0.827	0.467	0.660

Table 10. Number of best and second-best rankings. The values in **bold** emphasise the FSL approach that achieved the highest number of first places (columns 1st) and second places (columns 2nd) for each task type (Per Task) and overall (Total).

Approach	Per Task				Total	
	1st (1-Shot)	2nd (1-Shot)	1st (5-Shot)	2nd (5-Shot)	1st	2nd
PNs	0	1	1	2	1	3
SS	1	4	1	6	2	10
FT	0	2	3	10	3	12
BD-CSPN	9	5	4	3	13	8
LS	4	4	2	4	6	8
PT-MAP	6	1	5	2	11	3
TIM	5	10	15	2	20	12
TFT	0	1	1	2	1	3

4.2. RQ_2

To reason about the relationship between the number of epochs used during training and the quantity of unseen classes during inference, we calculated the improvement in the average accuracies when training with a higher number of epochs compared to the base case with 10 epochs. Tables 11 and 12 present the average of the results with all techniques per dataset and few-shot tasks. In the column Comparison, the meaning is this: a higher number of epochs compared to (→) the lower (base) number of epochs.

Table 11. Average improvement in the average accuracies: EuroSAT, XAI4SAR, and UC Merced. The values in **red** indicate a decrease in performance.

Comparison	EuroSAT		XAI4SAR		UC Merced	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
50 → 10	3.549%	7.013%	−14.729%	−14.226%	15.630%	11.075%
100 → 10	6.733%	13.200%	2.938%	7.554%	25.561%	23.168%
200 → 10	9.480%	18.582%	7.187%	11.384%	29.820%	26.347%

Table 12. Average improvement in the average accuracies: WHU-RS19, AID, and RESISC-45.

Comparison	WHU-RS19		AID		RESISC-45	
	1-Shot	5-Shot	1-Shot	5-Shot	1-Shot	5-Shot
50 → 10	16.749%	11.242%	45.791%	41.197%	18.496%	26.429%
100 → 10	17.814%	12.530%	40.692%	43.873%	17.648%	25.600%
200 → 10	22.140%	15.087%	58.412%	51.887%	21.717%	32.704%

Tables 11 and 12 show that the improvement when increasing the number of epochs is not very high for the settings with low-number (2-way) unseen classes during the inference phase, i.e., the datasets EuroSAT and XAI4SAR. The maximum improvement is 18.512% in 5-shot tasks when the number of epochs goes from 10 to 200, considering the EuroSAT set. The decrease in the performance in the XAI4SAR case (the **red** values in Table 11) will be discussed later.

As for the medium number of unseen classes (4-way), i.e., the UC Merced and WHU-RS19 datasets, the increase is better than in the previous comparison but not as much better. The highest improvement is 29.820% in 1-shot tasks with the UC Merced set going from 10 to 200 epochs. The most impressive gains happened with the highest number of unseen classes, i.e., AID (6-way) and RESISC-45 (9-way). Particularly, improvements related to the AID dataset are remarkable reaching 58.412% in 1-shot tasks (from 10 to 200 epochs). The results considering RESISC-45 are comparable to UC Merced but better

where a maximum gain of 32.704% was perceived in the 5-shot tasks (from 10 to 200 epochs). Thus, we conclude that, within classical training, a larger number of training epochs is more beneficial when there are more unseen classes during the inference phase.

4.3. RQ_3

As mentioned in Section 3.4, the values of the average accuracy for all 1-shot tasks, considering all datasets and the number of epochs, were compared to the respective values obtained for 5-shot tasks. The Shapiro–Wilk test for the 1-shot set resulted in a statistic of 0.982878 and p -value of 1.914196×10^{-2} . Even though the test statistic suggests the data are nearly normal, the low p -value indicates that this difference from normality is statistically significant. As for the 5-shot set, the values are as follows: statistic = 0.956710 and p -value = 1.325552×10^{-5} . The same reasoning as in the previous case applies here. Thus, we reject the null hypothesis that the sets were drawn from a normal distribution.

Figure 6 shows the Q-Q plots for both sets to confirm this conclusion. Recall that a Q-Q plot is a graphical tool used to compare the distributions of two datasets or to check if a dataset follows a specific distribution, such as the normal distribution. The 45-degree line (the red one in the plots) represents the expected quantiles if the data follow a normal distribution. If the data points (in blue) closely align with the 45-degree line, this suggests that the data are normally distributed. It is clear by looking at Figure 6 that these are not the cases and this supports the rejection of the null hypothesis.

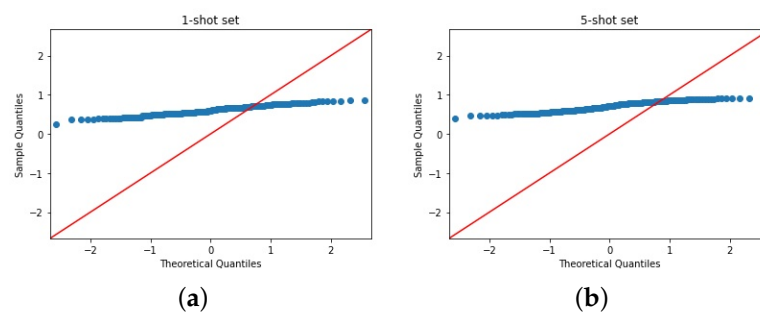


Figure 6. Q-Q plots for the 1-shot and 5-shot sets. (a) 1-shot set. (b) 5-shot set.

The application of the non-parametric Wilcoxon test, with the alternative *greater* and paired samples $d = 5\text{-shot} - 1\text{-shot}$, gave the following results: statistic = 18,528 and p -value = 1.466873×10^{-33} . The mean value of the 5-shot set is 0.703370 while for the 1-shot set is 0.602152. Thus, we conclude that there is strong statistical evidence that the values in the 5-shot set are consistently and significantly greater than those in the 1-shot set. Figure 7 presents the average accuracies for all samples from the 5-shot and 1-shot sets.

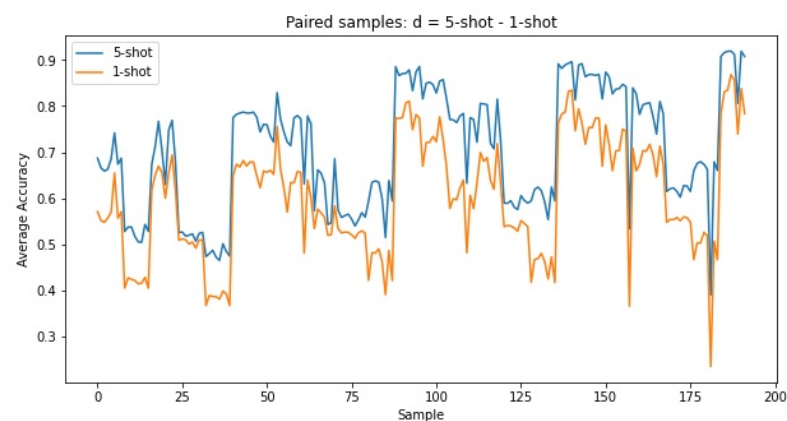


Figure 7. Average accuracies: 5-shot and 1-shot sets.

4.4. RQ_4

RESISC-45 was the dataset where the FSL techniques faced the greatest difficulty. The *Or9* subset was defined with the following classes of the training set: *Or9* = {airport, basketball court, bridge, church, commercial area, golf course, harbour, intersection, island}. By calculating the structural similarity (SSIM) with all other RESISC-45 classes, the scenario (the new test dataset) *More*, containing the most similar classes, resulted in the following: *More* = {meadow, rectangular farmland, runway, circular farmland, baseball diamond, lake, ship, river, cloud}. On the other hand, the scenario (the new test dataset) with the less similar classes is *Less* = {snowberg, dense residential, chaparral, mobile home park, palace, railway station, storage tank, industrial area, medium residential}.

Considering both tasks (1-shot and 5-shot) and all the number of epochs, the three sets do not follow a normal distribution as before, as shown by the application of the Shapiro–Wilk test and confirmation by Q-Q plots. The results of the non-parametric Wilcoxon test, with the alternative *greater* and paired samples $d = Or9 - More$, were as follows: statistic = 1094.5 and $p\text{-value} = 1.018490 \times 10^{-17}$. The mean value of the *Or9* set is 0.473309 while for the *More* set is 0.438200. Thus, there is strong statistical evidence that the values in the *Or9* set are consistently and significantly greater than those in the *More* set. In other words, if the test set has classes that are more similar to some classes in the training set, this not only fails to improve the performance of the techniques but, in fact, there is evidence that randomly selecting the classes in the test set is much better.

As for the *Less* scenario, the outcomes of the non-parametric Wilcoxon test, with the alternative *greater* and paired samples $d = Or9 - Less$, were as follows: statistic = 1176 and $p\text{-value} = 8.115114 \times 10^{-10}$. The mean values are as follows: *Or9* = 0.473309 and *Less* = 0.426466. Again, there is strong statistical evidence that the values in the *Or9* set are consistently better than the ones in the *Less* set.

The results per type of shot task were obtained too. Again, not only for the 1-shot but also for the 5-shot tasks, all sets do not follow the normal distribution. As for the 1-shot tasks, these were the outcomes of the Wilcoxon test for the *More* scenario, with the alternative *greater* and paired samples $d = Or9 - More$: statistic = 300, $p\text{-value} = 5.960465 \times 10^{-8}$, mean *Or9* = 0.408720, and mean *More* = 0.364363. As for the *Less* scenario, paired samples $d = Or9 - Less$: statistic = 300, $p\text{-value} = 5.960465 \times 10^{-8}$, mean *Or9* = 0.408719, and mean *Less* = 0.348023.

Finally, the results of the Wilcoxon test for the 5-shot tasks are presented. As for the *More* scenario, they were as follows: statistic = 229, $p\text{-value} = 1.146602 \times 10^{-2}$, mean *Or9* = 0.537898, and mean *More* = 0.512037. As for the *Less* scenario, they were as follows: statistic = 300, $p\text{-value} = 5.960465 \times 10^{-8}$, mean *Or9* = 0.537898, and mean *Less* = 0.504910. Once again, we observe the same previous conclusions: the *Or9* set presents values that are statistically significantly better than the respective values in the *More* and *Less* sets, both for the 1-shot and 5-shot tasks.

The *More* and *Less* sets were also compared directly. When considering all tasks (both 1-shot and 5-shot) and 1-shot tasks alone, the *More* set performed significantly better than the *Less* set from a statistical standpoint. However, for 5-shot tasks alone, there was no statistically significant difference between the two sets. Consequently, it is not definitively possible to conclude that, when comparing the most extreme cases, a scenario with highly similar images consistently outperforms one with less similar images.

Therefore, the answer to this question is as follows: the existence of a set of images in the test set that are more similar to some images in the training set does not necessarily mean that the FSL strategies will exhibit a better performance. Conversely, images that are less similar in the test set do not necessarily mean that the performance of the approaches will be worse. In a direct comparison between the randomly selected test set (*Or9*) and the others, the *Or9* set performed significantly better.

5. Discussion

In Section 4, it was occasionally observed that the differences between the average accuracies were quite small, where the decision between the first- and second-best techniques was determined by the third decimal place of the value of the metric. When comparing approaches in the context of DL, it is always advisable to take scalability into account. With 1000 images, a 0.1 percent difference is not particularly significant, but 0.1% of 10,000,000 images is. Therefore, even though some differences were very small, this was taken into account to decide which approaches were the best.

In the same Section 4, transductive approaches presented better results compared to the inductive ones, with TIM emerging as the most effective technique. It is very likely that TIM's superior performance is largely due to how its loss function was designed. In TIM, the authors incorporated supervision from the support set S by combining a standard cross-entropy loss with empirical weighted mutual information between the query samples and their latent labels. This mutual information consists of two components. The first is a Monte Carlo estimate of the conditional entropy of the labels given the raw query features, aimed at minimising the uncertainty of the posteriors at unlabelled query samples and encouraging confident model outputs. The second component, a label-marginal entropy regulariser, promotes a uniform marginal distribution of labels, helping to avoid degenerate solutions that could arise from minimising only the conditional entropy.

By analysing the results to answer RQ_2, a particular approach showed some unstable behaviour. With the XAI4SAR dataset, PT-MAP decreased its performance in all situations when the number of epochs was increased, considering both types of tasks. The most significant decreases here occurred when the number of epochs increased from 10 to 50, with a drop of 20.229% in 1-shot tasks and 26.811% in 5-shot tasks. There was also a performance drop with the RESISC-45 dataset with 1-shot tasks when the number of epochs increased to 50, as well as in the case of the AID dataset considering 1-shot tasks, when the number of epochs reached 200. However, the worst performance was certainly when the number of epochs varied from 10 to 200 in the RESISC-45 dataset. In this case, for 1-shot tasks, the average accuracy dropped by 41.059% and 22.198% for 5-shot tasks. As can be seen in Table 9, the average accuracies of the PT-MAP approach were the worst considering all datasets, epochs, and techniques.

One possible explanation for this unstable behaviour of PT-MAP is the assumption of data distribution. The approach relies on preprocessing feature vectors to fit a Gaussian-like distribution. This assumption may not hold true for all datasets or tasks, potentially limiting the technique's applicability in real-world scenarios where data distributions can be complex and varied. Also, the results presented in their article [44] show some sensitivity to hyperparameter tuning, which may have affected its performance. Finally, it is important to mention that the decision was to go up to the end of the number of epochs with no early stopping defined. Thus, it is possible that validation overfitting happened when the best model was saved multiple times during the training step and, thus, the model might begin to learn patterns specific to the validation set that do not generalise well to the test dataset. Even if no hyperparameter tuning was accomplished, this possibility might not be discarded.

Table 11 exhibits a decrease in the average accuracy for the XAI4SAR dataset when the number of epochs increased to 50, considering both 1-shot and 5-shot tasks. In this case, not only the PT-MAP technique but all approaches presented this decrease in both tasks. By increasing to 100 and 200 epochs, there was a positive but not very significant gain. SAR images are quite different from optical imagery. SAR signals contain inherent speckle noise and complex texture patterns, which can make it harder for DL approaches to properly converge. The variations in performance could be the models struggling to learn the underlying patterns in the SAR data while also handling noise. Thus, it is reasonable to say that with less epochs (10), the models might have learned some general patterns but not yet captured the complexity or fine details of the SAR images. As training progresses, the models may start to overfit to noisy features or irrelevant

patterns (leading to a decreased performance at 50 epochs). However, with more epochs (100, 200), the techniques might learn to discard these noisy or redundant features, resulting in a performance recovery.

As previously mentioned, we expected a better performance from a technique with 5-shot tasks than with 1-shot ones. But it is important to realise whether this difference is statistically significant. If the difference is not statistically significant, practitioners may choose a less costly scenario with only one image per class in the support set. However, the answer to RQ_3 showed strong statistical evidence that the 5-shot option is significantly better than the 1-shot solution. This highlights the importance of having more images per class in the support set.

There is no evidence that having images in the test set that are more similar to some images in the training set improves the performance of the approaches. Nor did we find that having less similar images worsens the situation. Perhaps a different choice of similarity metric, instead of the SSIM, could yield different results, but this metric was selected due to its widespread use. Randomly “selecting” classes for the test set proved to be the best option.

6. Conclusions

The RS community has been exploiting FSL, and there are several studies in this path. In spite of that, independent evaluations aimed at providing pieces of evidence that certain solutions are more valuable to the community than others are fundamentally important. This is because the RS community, in general, is more of a consumer of AI to solve its problems, which is perfectly appropriate, and such hints are relevant for the development of research.

This is precisely the motivation for this research. We selected several FSL approaches, some inductive and others transductive, which had been evaluated with image datasets that did not contain the characteristics of RS images. A comprehensive evaluation with six datasets created for scene classification was conducted: EuroSAT, XAI4SAR, UC Merced, WHU-RS19, AID, and RESISC-45. TIM, a transductive approach, proved to be the best solution, while PT-MAP, also transductive, exhibited unstable behaviour.

The responses of the four research questions suggested in this study are as follows: (1) transductive approaches are better than inductive ones; (2) a larger number of training epochs is more beneficial when there are more unseen classes during the inference phase; (3) using five samples in the support set is statistically significantly better than using only one; and (4) a higher similarity between unseen classes (during inference) and some of the training classes does not lead to an improved performance. This research then advocates for further development of transductive approaches in the RS community given their superior performance. Additionally, to address the ongoing challenge of limited labelled data, our findings suggest a need for novel FSL techniques capable of handling a higher number of new classes at inference time. In practical terms, the results suggest for practitioners not to stipulate a large number of epochs in the context of FSL classical training. Additionally, no type of preprocessing activity is necessary regarding the similarity between images of the test set and images of the training set in order to obtain a better performance. These findings can guide researchers and professionals in selecting optimal solutions/strategies for developing their applications demanding few labelled samples.

Threats to external validity weaken confidence in generalising experiment results across individuals, settings, and time. Population threats are examples of threats to external validity, and they are related to the representativeness of the experiment samples. In this study, six RS datasets were considered. While six databases have a significant number of samples with characteristics that can represent the scene classification context well, the quantity of selected FSL techniques (eight) cannot be considered totally representative. To combat this limitation, it is necessary to choose more techniques, both inductive and transductive, and perform a new evaluation to see whether the results obtained by this research are still valid.

Therefore, future directions include precisely increasing the number of FSL approaches, as just pointed out. It is also possible to augment the number of RS image scene classification datasets, although the current selected six sets represent this task well. Thus, a new robust evaluation should be carried out, and hence it is possible to check whether the results and conclusions of this study still hold. Additionally, other computer vision tasks, such as object detection and semantic segmentation, will be addressed.

Funding: This research was supported by the *Agência Espacial Brasileira* (AEB—Brazilian Space Agency).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this research are already publicly available. The reference list of this manuscript includes all articles that provide descriptions of the datasets used.

Acknowledgments: This research was developed within the project *Classificação de imagens e dados via redes neurais profundas para múltiplos domínios* (Image and data classification via Deep neural networks for multiple domainS—IDeepS). The IDeepS (available online: <https://github.com/vsantjr/IDeepS> (accessed on 4 November 2024)) project is supported by the *Laboratório Nacional de Computação Científica* (LNCC—National Laboratory for Scientific Computing, MCTI, Brazil) via resources of the SDumont supercomputer. This research was also supported by the *Agência Espacial Brasileira* (AEB—Brazilian Space Agency).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhou, Z.H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **2017**, *5*, 44–53. [CrossRef]
2. Li, Y.F.; Guo, L.Z.; Zhou, Z.H. Towards Safe Weakly Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 334–346. [CrossRef] [PubMed]
3. van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
4. Chen, Y.; Tan, X.; Zhao, B.; Chen, Z.; Song, R.; Liang, J.; Lu, X. Boosting Semi-Supervised Learning by Exploiting All Unlabeled Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7548–7557.
5. Rani, V.; Nabi, S.T.; Kumar, M.; Mittal, A.; Kumar, K. Self-supervised Learning: A Succinct Review. *Arch. Comput. Methods Eng.* **2023**, *30*, 2761–2775. [CrossRef] [PubMed]
6. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap Your Own Latent—A New Approach to Self-Supervised Learning. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Virtual, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Volume 33, pp. 21271–21284.
7. Zhu, W.; Liu, J.; Huang, Y. HNSSL: Hard Negative-Based Self-Supervised Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Vancouver, BC, Canada, 18–22 June 2023; pp. 4778–4787.
8. Toche Tchio, G.M.; Kenfack, J.; Kassegne, D.; Menga, F.D.; Ouro-Djobo, S.S. A Comprehensive Review of Supervised Learning Algorithms for the Diagnosis of Photovoltaic Systems, Proposing a New Approach Using an Ensemble Learning Algorithm. *Appl. Sci.* **2024**, *14*, 2072. [CrossRef]
9. Aljuaid, A.; Anwar, M. Survey of Supervised Learning for Medical Image Processing. *SN Comput. Sci.* **2022**, *3*, 292. [CrossRef]
10. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Comput. Surv.* **2020**, *53*, 1–34. [CrossRef]
11. Song, Y.; Wang, T.; Cai, P.; Mondal, S.K.; Sahoo, J.P. A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities. *ACM Comput. Surv.* **2023**, *55*, 1–40. [CrossRef]
12. Laenen, S.; Bertinetto, L. On episodes, prototypical networks, and few-shot learning. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Virtual, 6–14 December 2021.
13. Zhu, H.; Koniusz, P. Transductive Few-Shot Learning with Prototype-Based Label Propagation by Iterative Graph Refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 23996–24006.
14. Sun, Q.; Chao, J.; Lin, W.; Xu, Z.; Chen, W.; He, N. Learn to Few-Shot Segment Remote Sensing Images from Irrelevant Data. *Remote Sens.* **2023**, *15*, 4937. [CrossRef]
15. Liu, Y.; Zhang, L.; Han, Z.; Chen, C. Integrating Knowledge Distillation with Learning to Rank for Few-Shot Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]

16. Tang, J.; Zhang, F.; Zhou, Y.; Yin, Q.; Hu, W. A Fast Inference Networks for SAR Target Few-Shot Learning Based on Improved Siamese Networks. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1212–1215. [\[CrossRef\]](#)
17. Wang, L.; Yang, X.; Tan, H.; Bai, X.; Zhou, F. Few-Shot Class-Incremental SAR Target Recognition Based on Hierarchical Embedding and Incremental Evolutionary Network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–11. [\[CrossRef\]](#)
18. Li, Y.; Bian, C. Few-Shot Fine-Grained Ship Classification with a Foreground-Aware Feature Map Reconstruction Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [\[CrossRef\]](#)
19. Liu, Y.; Zhang, T.; Zhuang, Y.; Wang, G.; Chen, H. Multi-Grained Global-Local Semantic Feature Fusion for Few Shot Remote Sensing Scene Classification. In Proceedings of the IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; pp. 6235–6238. [\[CrossRef\]](#)
20. Zhang, B.; Feng, S.; Li, X.; Ye, Y.; Ye, R.; Luo, C.; Jiang, H. SGMNet: Scene Graph Matching Network for Few-Shot Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [\[CrossRef\]](#)
21. Liu, Q.; Peng, J.; Ning, Y.; Chen, N.; Sun, W.; Du, Q.; Zhou, Y. Refined Prototypical Contrastive Learning for Few-Shot Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [\[CrossRef\]](#)
22. Zhao, S.; Bai, Y.; Shao, S.; Liu, W.; Ge, X.; Li, Y.; Liu, B. SELM: Self-Motivated Ensemble Learning Model for Cross-Domain Few-Shot Classification in Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 1–5. [\[CrossRef\]](#)
23. Yang, Z.; Zhang, Y.; Zheng, J.; Yu, Z.; Zheng, B. Scale Information Enhancement for Few-Shot Object Detection on Remote Sensing Images. *Remote Sens.* **2023**, *15*, 5372. [\[CrossRef\]](#)
24. Huang, X.; He, B.; Tong, M.; Wang, D.; He, C. Few-Shot Object Detection on Remote Sensing Images via Shared Attention Module and Balanced Fine-Tuning Strategy. *Remote Sens.* **2021**, *13*, 3816. [\[CrossRef\]](#)
25. Wang, L.; Bai, X.; Gong, C.; Zhou, F. Hybrid Inference Network for Few-Shot SAR Automatic Target Recognition. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9257–9269. [\[CrossRef\]](#)
26. Pan, C.; Huang, J.; Gong, J.; Hao, J. Few-shot learning with hierarchical pooling induction network. *Multimed. Tools Appl.* **2022**, *81*, 32937–32952. [\[CrossRef\]](#)
27. LENS.ORG. LENS.ORG: Explore Global Science and Technology Knowledge. Available online: <https://www.lens.org/> (accessed on 4 November 2024).
28. Piccialli, F.; Somma, V.D.; Giampaolo, F.; Cuomo, S.; Fortino, G. A survey on deep learning in medicine: Why, how and when? *Information Fusion* **2021**, *66*, 111–137. [\[CrossRef\]](#)
29. Albahar, M. A Survey on Deep Learning and Its Impact on Agriculture: Challenges and Opportunities. *Agriculture* **2023**, *13*, 540. [\[CrossRef\]](#)
30. Ozbayoglu, A.M.; Gudelek, M.U.; Sezer, O.B. Deep learning for financial applications: A survey. *Appl. Soft Comput.* **2020**, *93*, 106384. [\[CrossRef\]](#)
31. Elallid, B.B.; Benamar, N.; Hafid, A.S.; Rachidi, T.; Mrani, N. A Comprehensive Survey on the Application of Deep and Reinforcement Learning Approaches in Autonomous Driving. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 7366–7390. [\[CrossRef\]](#)
32. Zhao, M.; Liu, Y. A class distribution learning method for few-shot remote sensing scene classification. *Remote Sens. Lett.* **2024**, *15*, 558–569. [\[CrossRef\]](#)
33. Yuan, T.; Liu, W.; Liu, B. Double Discriminative Constraint-Based Affine Nonnegative Representation for Few-Shot Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [\[CrossRef\]](#)
34. Zeng, Q.; Geng, J.; Jiang, W.; Huang, K.; Wang, Z. IDLN: Iterative Distribution Learning Network for Few-Shot Remote Sensing Image Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [\[CrossRef\]](#)
35. Sheng, Y.; Xiao, L. Manifold Augmentation Based Self-Supervised Contrastive Learning for Few-Shot Remote Sensing Scene Classification. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 2239–2242. [\[CrossRef\]](#)
36. Yuan, Z.; Tang, C.; Yang, A.; Huang, W.; Chen, W. Few-Shot Remote Sensing Image Scene Classification Based on Metric Learning and Local Descriptors. *Remote Sens.* **2023**, *15*, 831. [\[CrossRef\]](#)
37. Pei, S.; Wang, Y.; Ma, J.; Tang, X.; Yang, Y. Multi-Scale Interaction Prototypical Network For Few-Shot Remote Sensing Scene Classification. In Proceedings of the IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; pp. 6231–6234. [\[CrossRef\]](#)
38. Dong, Z.; Lin, B.; Xie, F. Optimizing Few-Shot Remote Sensing Scene Classification Based on an Improved Data Augmentation Approach. *Remote Sens.* **2024**, *16*, 525. [\[CrossRef\]](#)
39. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4080–4090.
40. Wang, Y.; Chao, W.L.; Weinberger, K.Q.; van der Maaten, L. SimpleShot: Revisiting Nearest-Neighbor Classification for Few-Shot Learning. *arXiv* **2019**, arXiv:1911.04623.
41. Chen, W.; Liu, Y.; Kira, Z.; Wang, Y.F.; Huang, J. A Closer Look at Few-shot Classification. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
42. Liu, J.; Song, L.; Qin, Y. Prototype Rectification for Few-Shot Learning. In *Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Elsevier: Cham, Switzerland, 2020; pp. 741–756.

43. Ziko, I.; Dolz, J.; Granger, E.; Ayed, I.B. Laplacian Regularized Few-Shot Learning. In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020; Volume 119, pp. 11660–11670.
44. Hu, Y.; Gripon, V.; Pateux, S. Leveraging the Feature Distribution in Transfer-Based Few-Shot Learning. In *Proceedings of the Artificial Neural Networks and Machine Learning—ICANN 2021, Online, 14–17 September 2021*; Farkaš, I., Masulli, P., Otte, S., Wermter, S., Eds.; Elsevier: Cham, Switzerland, 2021; pp. 487–499.
45. Boudiaf, M.; Masud, Z.I.; Rony, J.; Dolz, J.; Piantanida, P.; Ayed, I.B. Transductive information maximization for few-shot learning. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Virtual, 6–12 December 2020; pp. 2445–2457.
46. Dhillon, G.S.; Chaudhari, P.; Ravichandran, A.; Soatto, S. A Baseline for Few-Shot Image Classification. In Proceedings of the Eight International Conference on Learning Representations, Virtual, 26 April–1 May 2020.
47. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226. [\[CrossRef\]](#)
48. Huang, Z.; Yao, X.; Liu, Y.; Dumitru, C.O.; Datcu, M.; Han, J. Physically explainable CNN for SAR image classification. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 25–37. [\[CrossRef\]](#)
49. Yang, Y.; Newsam, S. Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279. [\[CrossRef\]](#)
50. Xia, G.S.; Yang, W.; Delon, J.; Gousseau, Y.; Sun, H.; Maitre, H. Structural High-resolution Satellite Image Indexing. In Proceedings of the ISPRS TC VII Symposium—100 Years ISPRS, Vienna, Austria, 5–7 July 2010; pp. 298–303.
51. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [\[CrossRef\]](#)
52. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [\[CrossRef\]](#)
53. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [\[CrossRef\]](#)
54. Ravi, S.; Larochelle, H. Optimization as a Model for Few-Shot Learning. In Proceedings of the Fifth International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
55. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *Caltech-UCSD Birds-200-2011 (CUB-200-2011) Dataset*; Technical Report CNS-TR-2011-001; California Institute of Technology: Pasadena, CA, USA, 2011.
56. Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J.B.; Larochelle, H.; Zemel, R.S. Meta-Learning for Semi-Supervised Few-Shot Classification. In Proceedings of the Sixth International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
57. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [\[CrossRef\]](#)
58. Sicara. Easy Few-Shot Learning. Available online: <https://github.com/sicara/easy-few-shot-learning> (accessed on 4 November 2024).
59. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [\[CrossRef\]](#)
60. Laboratório Nacional de Computação Científica (LNCC). SDumont: Sistema de Computação Petaflopica do SINAPAD. Available online: <https://sdumont.lncc.br/> (accessed on 4 November 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.