International Journal of Computer Vision
https://doi.org/10.1007/s11263-025-02518-z

Analysing Satellite Imagery Classification under Spatial Domain Shift across Geographic Regions

Sara A. Al-Emadi[1,2] · **Yin Yang**[2] · **Ferda Ofli**[1]

Received: 29 September 2024 / Accepted: 25 June 2025

Abstract

Deep learning models are designed based on the i.i.d. assumption; consequently, they experience a significant performance drop due to the distribution shifts when deployed in real environments. Domain Generalisation (DG) aims to bridge the distribution shift between the source and target domains by improving the generalisability of the model to Out-Of-Distribution (OOD) data. This challenge is prominent in satellite imagery classification due to the scarcity of data from underrepresented regions such as Africa and Oceania. In this paper, we address the limitations of existing datasets in capturing distribution shifts caused by geospatial differences between geographic regions by constructing a new, large-scale dataset called Domain Shift across Geographic Regions (DSGR). This dataset aims to help researchers better understand the impact of distribution shifts on satellite imagery classification. Furthermore, we perform rigorous experiments on DSGR to investigate and benchmark the robustness of existing DG techniques under single- and multi-source domain settings and the role of foundation models in enhancing the DG techniques. Our evaluations reveal that recent DG techniques have a comparable, yet weak, performance on DSGR. However, when combined with a foundation model like CLIP, ERM (introduced in 1999) achieves highly competitive results, surpassing even recent state-of-the-art DG solutions in enhancing the generalisability of deep learning models across different geographic regions. Our dataset and code are available at https://github.com/RWGAI/DSGR.

Keywords Domain Generalisation · Distribution shift · Out-of-Distribution Generalisation · Domain Shift · Land Use Classification · Remote Sensing

1 Introduction

Deep learning (DL) models are extensively used in remote sensing applications such as object detection in satellite imagery (Xu et al., 2022b; Peng et al., 2022), land cover (Kalita et al., 2021; Luo & Ji, 2022) and land-use classification (Voreiter et al., 2020; Zheng et al., 2020; Xu et al., 2022a), road extraction (Lu et al., 2022), flood mapping (Drakonakis et al., 2022; Sadiq et al., 2022), semantic segmentation (Tasar et al., 2021; Wu et al., 2020), and others (Suel et al., 2021; Li et al., 2023a; Nguyen et al., 2024; Pott et al., 2021). One issue that is often neglected is that DL models are designed based on the i.i.d. assumption. Consequently, they tend to fail in bridging the *domain shift* gap experienced when the samples at test time come from a *target* domain that has a different underlying distribution than that of the *source* domain(s) seen during training. This limitation deteriorates their ability to generalise from the *In-Distribution* (ID) data to new and unseen *Out-Of-Distribution* (OOD) data, leading to significant performance degradation whenever the model is faced with OOD data.

To address this issue, *Domain Generalisation* (DG) techniques[1] aim to bridge the gap by improving the generalisability of modern DL models to OOD data while being strictly limited to the source domain data during training (Zhou et al., 2022a; Wang et al., 2022; Gulrajani & Lopez-Paz, 2021;

Communicated by Yongchan Kwon.

✉ Sara A. Al-Emadi
salemadi@hbku.edu.qa

Yin Yang
yyang@hbku.edu.qa

Ferda Ofli
fofli@hbku.edu.qa

[1] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

[2] College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

[1] More details about DG techniques are discussed in Section 2.1.

Published online: 08 August 2025

Robey et al., 2021; Ding et al., 2022; Arjovsky et al., 2019; Rame et al., 2022; Harary et al., 2022; Sun & Saenko, 2016; Lin et al., 2021; Shi et al., 2022; Eastwood et al., 2022; Li et al., 2023b; Arpit et al., 2022). Their efficacy is then typically evaluated on a number of standard DG benchmark datasets such as RotatedMNIST (Ghifary et al., 2015), PACS (Li et al., 2017), DomainNet (Peng et al., 2019), VLCS (Fang et al., 2013), and OfficeHome (Venkateswara et al., 2017), among others. However, these DG datasets do not reflect the challenging and complex environments in which DL models are deployed for earth observation tasks such as land-use classification (Voreiter et al., 2020; Zheng et al., 2020; Xu et al., 2022a).

While there are a few recent remote sensing datasets that attempt to reflect the domain shift gap, these datasets have narrow application scopes (e.g., Auto Arborist (Beery et al., 2022) and PovertyMap-WILDS (Koh et al., 2021)) and limited geographical coverage (e.g., GeoNet (Kalluri et al., 2023)). In particular, when it comes to the task of land-use classification using satellite imagery, WILDS (Koh et al., 2021) took a step in this direction by introducing FMoW-WILDS, a DG dataset designed to study *temporal* domain shifts in this task, where the domains were defined in terms of years. While understanding the effects of the temporal domain shift is vital in such a task, less attention has been devoted to introducing a DG dataset that contributes towards investigating the performance of DL models in handling yet another crucial domain shift, known as the *spatial* domain shift (i.e., covariate shift) at a global scale. Such a shift arises from differences in the appearance of built structures and land cover due to factors like natural landscape, architectural design, financial and economic development, social and cultural characteristics, human settlement patterns and demographics, etc. Huang et al. (2020); Ma et al. (2024).

One could argue that the temporal domain shift follows a natural order, that is the time distance between domains, would provide insight on the effect of the shift in the data distribution. For example: considering that Years are domains, the closer the years between two domains representing a building, the chances that the two images would be similar is higher. Whereas, a higher shift would be observed for the images of two buildings 30 years apart. This is due to the continuous space nature of the time based shift. However, this is not true for spatial domain shifts as there is no order between the data (similar to the concept of categorical split) which makes it a more challenging problem to tackle. The difference between temporal and spatial domain shift in satellite imagery, represented through samples of the same class, is illustrated in Figure 1 below.

To address the shortcomings of the aforementioned datasets, we propose **D**omain **S**hift across **G**eographic **R**egions (**DSGR**), a novel DG dataset to explicitly portray the effects of spatial domain shift (i.e., covariate shift) on satellite imagery classification at a global scale and measures the performance of modern DG algorithms on data gathered from different regions. With the goal of creating a DG dataset that reflects different real-life scenarios, we follow the geographic region categorisation defined by the Population Division of the Department of Economic and Social Affairs in the United Nations (Nations, 2022). Therefore, DSGR consists of six domains: Asia, Africa, Oceania, Europe, Latin America and the Caribbean, and Northern America.

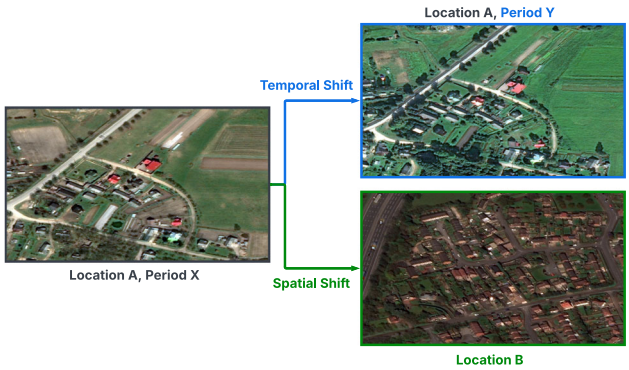


Fig. 1 Comparison between spatial and temporal domain shifts

A robust land-use classification model under domain shift is crucial in real-life applications, where a model is trained on data from a specific geographic region and is then deployed and expected to maintain its high performance on a new geographic region. Such applications range from assessing the environmental impacts of land-use and the socioeconomic development of a region to land resource management, etc. Xu et al. (2020). Figure 2 illustrates the phenomenon of spatial domain shift across geographic regions in the land-use classification task. Since each geographic region is characterised by unique geospatial features, a model trained on the source domains, e.g., Europe and Latin America and the Caribbean, to classify whether a building belongs to *Single-Unit Residential* or *Oil and Gas Facility* class will experience a significant degradation in performance when presented with satellite images from the target domain, e.g., Asia. This is due to the domain shift gap introduced by the uniqueness of features of the geographic regions used as source domains versus the features of the target geographic region. More specifically, the land cover types, which can be seen in the samples of both classes, the architectural differences when it comes to the *Oil and Gas Facility* class or the density of units in the *Single-Unit Residential* class. This issue of spatial domain shift across geographic regions is crucial as it impedes the large-scale deployment of DL models in practice.

In this work, we designed DSGR to address the underexplored state of DG benchmark datasets with spatial domain shift across geographies worldwide for land-use classification, providing a solid benchmark dataset to aid the research

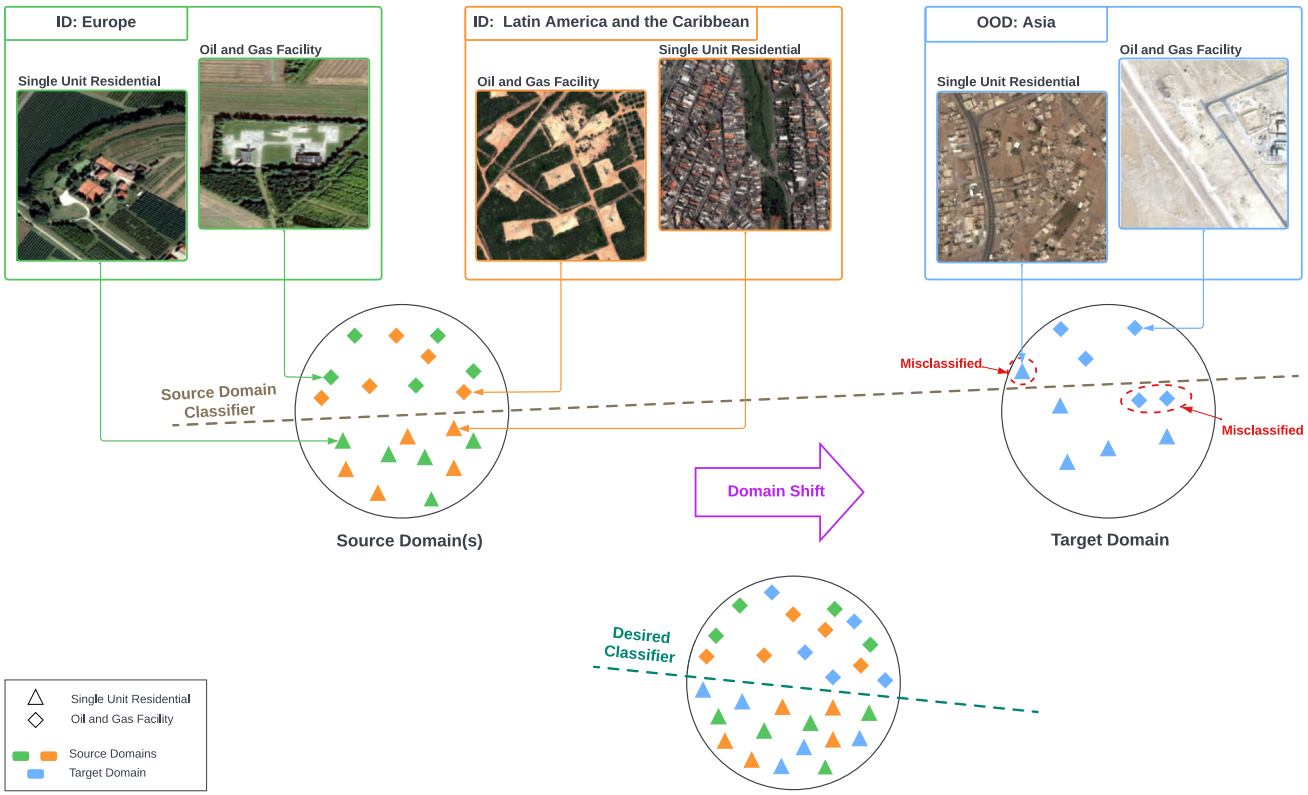


Fig. 2 Domain shift in satellite imagery with domains defined as geographic regions: (left) Europe, (centre) Latin America and the Caribbean and (right) Asia. It can be observed that the buildings and the geographical landscapes are different across the different domains. A classifier trained on source domains under the i.i.d. assumption (top-left) will likely misclassify samples from the OOD target domain (top-right). The aim of DG is to produce a classifier that generalises to OOD data while maintaining its performance on ID data (bottom-centre)

community in assessing the DG techniques designed for such a task and underscoring the need for advanced datasets to evaluate the performance of DL-based satellite imagery classification under such a domain shift. Furthermore, using DSGR, we investigate and provide new insights into the OOD performance of several modern DG algorithms under different source–target domain settings. Moreover, with the rise and success of foundation models in many applications, we show that their usage as backbone models in SOTA DG techniques can further enhance the generalisability of the models to OOD data.

Our contributions are summarised as follows:

- We introduce DSGR, a novel, realistic and challenging DG dataset for land-use classification in satellite imagery (Section 3).
- We provide insights into the role of single-source versus multi-source training in addressing DG for spatial domains in satellite imagery, where we deduce that multi-source training results in a higher generalisability of DL models (Sections 5.3 and 5.4).
- We facilitate the community with an in-depth analysis of the performance of the SOTA DG algorithms on DSGR. We observe that their performance varies slightly with the experimental setup. However, these variations are insignificant (Sections 5.3 and 5.4).
- We examine the influence of foundation models on the overall performance of the SOTA DG algorithms, using two versions of CLIP, each with a different backbone size. Our experiments reveal that the classical Empirical Risk Minimisation (ERM) method, when paired with CLIP, outperforms other SOTA DG methods (Section 5.5).
- We investigate the impact and limitations of a popular OOD-aware training scheme, using a left-out source domain as an OOD validation set during training, on the SOTA DG models. We conclude that such scheme falls short of improving the generalisability of DL models under spatial domain shift (Section 5.6).

The rest of the paper is organised as follows: Section 2 expands the discussion to provide an in-depth literature review and highlight the most relevant works to this paper. Section 3 presents DSGR, the satellite imagery DG dataset

for land-use classification introduced in this paper. Then, we lay the foundation to our experiments in Section 4 and present our experimental results, analyses and performance evaluations in Section 5. In Section 6, we provide a thorough discussion, highlighting the implications of DSGR in real-life applications, limitations of this work and our future directions. Finally, Section 7 provides the conclusion to this paper.

2 Related Work

This work is related to the existing work in domain generalisation in terms of benchmarks and analyses of domain generalisation techniques (Section 2.1) and datasets explicitly designed to analyse the performance of those techniques under domain shifts, including the recent DG remote sensing datasets (Section 2.2).

2.1 Domain Generalisation Techniques

The last few years have witnessed a surge of interest in developing DG techniques to address the performance hit experienced by DL models when faced with OOD data. One of the most intuitive and preferred techniques is increasing the size and the diversity of the source dataset with the aim of covering a larger distribution that might be close to that of the target in the latent space. This is typically achieved using a range of *data augmentation* techniques (Volpi & Murino, 2019; Robey et al., 2021; Zhang et al., 2018). However, a major shortcoming of data augmentation is that in real-world applications, the target data and its distribution are unknown during the training phase. Therefore, adding data from additional source domains does not guarantee that such an addition will reflect positively, if any, on the generalisability of DL models.

In a recent and popular study among the DG community, known as DomainBed (Gulrajani & Lopez-Paz, 2021), the authors shifted their focus towards *optimisation of classical algorithms* for DG. Their work suggests that using the decades-old Empirical Risk Minimisation (ERM) algorithm (Vapnik, 1999), where the model is trained to minimise the average training loss, with rigorous tuning is sufficient to boost the model's performance such that it outperforms the SOTA DG algorithms when tested on OOD datasets.

Another vein of studies focuses on designing *learning strategies* explicitly for DG such as *ensemble techniques* (Li et al., 2023b; Arpit et al., 2022) and *regularisation-based methods* (Shi et al., 2022; Rame et al., 2022; Eastwood et al., 2022; Arjovsky et al., 2019; Sagawa et al., 2019). As an example of a regularisation-based method, the Invariant Risk Minimisation (IRM) technique (Arjovsky et al., 2019) extends on ERM by adding a penalty to its objective function in order to enforce the notion of invariance across different source domains. This is achieved through penalising the model on feature distributions that have a different optimal linear classifier for each domain. This has been shown to increase the overall generalisability of the model on various DG datasets.

While not designed explicitly to measure the domain shift between different spatial domains, group Distributionally Robust Optimisation (group DRO) (Sagawa et al., 2019) aims to improve the generalisability of the predictive models by optimising the worst predictive loss of predefined distinct groups through distributionally robust optimisation. Due to the high-level of similarity between the notion of groups and domains, it has been used in recent literature as a DG technique.

Feature disentanglement (Lin et al., 2021; Bui et al., 2021) and *learning domain-invariant representations* (Ding et al., 2022; Harary et al., 2022; Sun & Saenko, 2016) are DG techniques that gained attention in the community recently. For instance, the authors introduced LRDG (Ding et al., 2022) to eliminate domain-specific features where they trained classifiers to learn features from different domains, then used an encoder-decoder model to remove these features. Similarly, building on the success of ERM, Deep Correlation Alignment (Deep CORAL) (Sun & Saenko, 2016) adds a penalty to its objective function, defined by differences in the means and covariances of feature distributions across domains, aiming to align these distributions. Although Deep CORAL has been originally used for Domain Adaptation (DA), its utilization has been extended to train DG models as well.

With the rise of *foundation models* and their effectiveness in generalisation, recent studies (Shu et al., 2023; Singha et al., 2024) attempt to address the domain shift gap by incorporating foundation models such as CLIP (Radford et al., 2021) as backbone architectures to SOTA DG algorithms. For example, CLIPood (Shu et al., 2023) describes an architecture coupled with CLIP through a new training objective and optimisation strategy. Their proposed technique is designed to maintain the strength of the pretrained parameters and leverage the new information obtained through fine-tuning the model on a new task. The authors have shown that CLIPood outperforms ERM and other SOTA DG algorithms on the traditional DG benchmark datasets such as those discussed in Section 2.2. We refer the reader to recent surveys (Zhou et al., 2022a; Wang et al., 2022; Shen et al., 2021) for further introduction and in-depth discussion about the SOTA DG methods.

In this work, we investigate the performance of five SOTA DG techniques, namely ERM (Vapnik, 1999), IRM (Arjovsky et al., 2019), Deep CORAL (Sun & Saenko, 2016), group DRO (Sagawa et al., 2019), and CLIPood (Shu et al., 2023) on our dataset, DSGR. We provide more details on the selected algorithms in Section 4.2.

2.2 Domain Generalisation Datasets

2.2.1 Traditional DG Benchmark Datasets

Some of the most widely used benchmark datasets for DG are relatively simple datasets such as PACS (Li et al., 2017), OfficeHome (Venkateswara et al., 2017), DomainNet (Peng et al., 2019), VLCS (Fang et al., 2013) or synthetic variations of classical datasets such as RotatedMNIST (Ghifary et al., 2015) and ColoredMNIST (Arjovsky et al., 2019). While these datasets are successful in demonstrating the phenomena of domain shift, they are considered not realistic enough (e.g., in Koh et al. (2021); Beery et al. (2022)) to represent complex, real-world scenarios (Zhang et al., 2023). This is due to the fact that they contain homogeneous domain shifts, where the shifts between source-to-source and source-to-target domains are highly correlated and predictable (Zhou et al., 2022a). For example, in PACS the shift is due to the style changes. Whereas, in a real-world scenario, the source-to-target shift is said to be unpredictable, which is also known as a heterogeneous domain shift (Zhou et al., 2022a). Furthermore, unlike the aforementioned datasets, an inherent difficulty of a real-world application such as land-use classification using satellite imagery is that the context are densely gathered around an object within an image. Hence, this highlights the need for datasets that reflect such a complex and heterogeneous domain shift in real-world applications, especially in the area of satellite imagery.

2.2.2 Realistic DG Benchmark Datasets

Recently, WILDS (Koh et al., 2021) was introduced as a suite of tools to study domain generalisation through datasets, with domain and/or subpopulation shifts in real-life applications, and some DG algorithms to overcome these shifts. For example, Camelyon17-WILDS is a medical imagery dataset where each hospital is considered as a domain. GlobalWheat-WILDS is proposed to study the domain shift in plant characteristics from close-up photos of plant fields captured in twelve countries (i.e., domains). Similarly, Auto Arborist (Beery et al., 2022) is a drone and street-view imagery dataset that aims to address geographical domain shift in urban forest monitoring application, covering 23 cities in the US and Canada. Another recent dataset GeoNet (Kalluri et al., 2023) comes close to the intuition behind DSGR, in which the authors proposed several datasets to study multiple elements, one of which is used for evaluating SOTA unsupervised DA methods across two different regions, namely, Asia and USA. Likewise, in the remote sensing field, PovertyMap-WILDS (Koh et al., 2021) was designed for DA to regress the poverty levels in different African countries.

When it comes to domain shifts in land-use classification, FMoW-WILDS (Koh et al., 2021) was introduced as a DG dataset to investigate the performance of DG techniques under temporal domain shift across different domains defined in terms of years. Their study has shown that the performance of the DG techniques was negatively affected by the temporal domain shift. Although much has been learned about temporal domain shift in the recent literature (Yao et al., 2022a; Xie et al., 2023; Yao et al., 2022b), understanding the effects of shift across space on a global scale for this task remains unexplored. Therefore, as of the time of writing this paper, our proposed DSGR dataset is a novel land-use classification DG benchmark for remote sensing applications that is focused on the spatial domain shift with a global coverage. Table 1 highlights the uniqueness of DSGR in comparison to the standard DG benchmark datasets, where it can be observed that for land-use classification, DSGR is the only DG dataset that provides the definition of domains in terms of geographic regions across the globe.

3 Proposed DSGR Dataset

We focus on the problem of domain shift with a set of distinct but similar domains for land-use classification application. In particular, we focus on the spatial domain shift, where we associate domains with geographic regions. To ensure that DSGR reflects scenarios in real-life applications and can be used to assess the robustness of models to domain shift prior to their deployment, as mentioned previously, we use the geographic region categorisation defined by the Population Division of the United Nations Department of Economic and Social Affairs in the United Nations (Nations, 2022). Hence, DSGR consists of the six geographic regions highlighted in Figure 3: Asia (AS), Africa (AF), Oceania (OC), Europe (EU), Latin America and the Caribbean (LAC), and Northern America (NA). The main goal behind such design is that, in a real-life environment, a DL model would be trained on a limited dataset acquired from a geographic region or a set of geographic regions but would be deployed and expected to perform well worldwide, including previously unseen territories.

3.1 Data Preprocessing

Due to the scarcity of satellite imagery with large global coverage coupled with the high cost of obtaining them, we utilise the raw satellite imagery released publicly in fMoW (Christie et al., 2018). The raw images of fMoW were acquired temporally from the DigitalGlobe (now Maxar) constellation with a resolution of 30cm for land-use classification task with a total of 63 unique classes representing around 200 countries worldwide. Following the practice of FMoW-WILDS (Koh

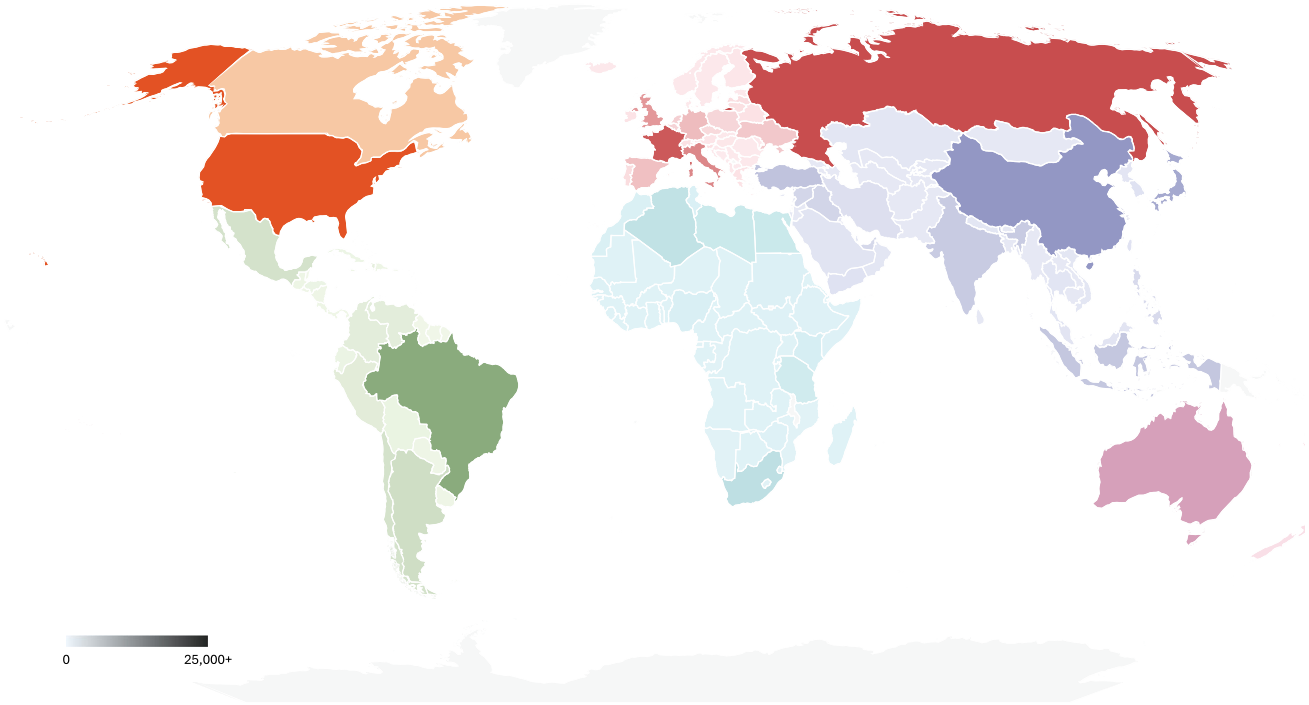


Fig. 3 The geographical distribution of DSGR which indicates a global representation. The different colours represent the six different geographic regions in DSGR individually. Whereas, the shades represent the number of samples per country within each region

et al., 2021) to reduce I/O usage, in the creation of DSGR, we used PNG compressed versions of the raw images that are resized to 224 × 224 pixels focusing only on the RGB bands of the satellite imagery.

We have performed careful cleaning, preprocessing and rebalancing of the raw images to reduce the influence of factors such as shortage of data for a specific domain, geographic region, or data imbalance on the overall performance of DL models. More notably, we attempted to address the issues of data imbalance (1) cross-splits within a domain, (2) cross-domains and (3) the cases where both cross-domain and cross-split discrepancies existed, through the steps illustrated in Figure 4. We provide the details of each step as follows:

Within-Domain Preprocessing

Initially, we followed the original dataset splits proposed by fMoW which consisted of *Train*, *Validation*, *Test* and *Seq* splits. The *Seq* split was omitted from our study. Reserving the original splits was necessary to ensure that there was no data leakage between the *Train* and *Test* splits since fMoW was designed as a temporal dataset, with individual sequences consisting of multiple images of the same location captured at different time instances. However, once we partitioned the raw images into distinct geographic regions, a mismatch between the number of classes cross-splits was observed. That is, some classes appeared in the *Train* split

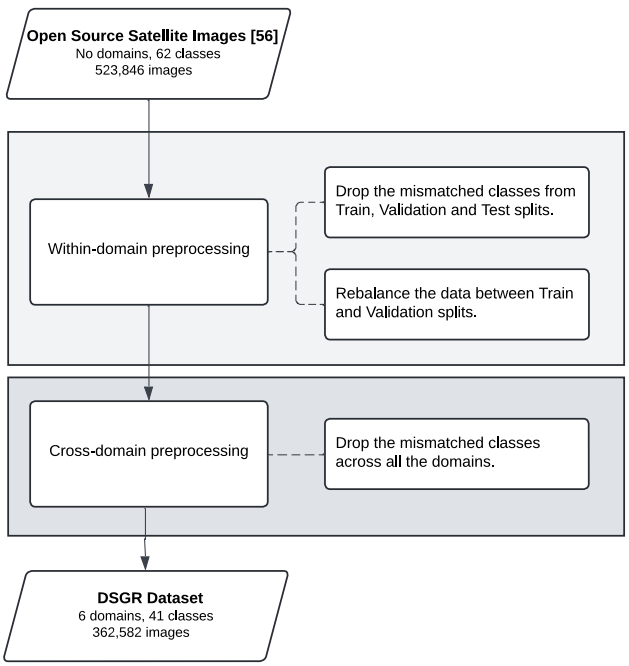


Fig. 4 Preprocessing pipeline used in the creation of our dataset, DSGR

but not in the *Test* split or vice versa. To address this issue, we carefully analysed the classes in each split within-domain and dropped the mismatched classes and those with the least number of samples. Afterwards, we noticed that while the

Table 1 Comparison of DG benchmark datasets

Dataset	Application	Image Style	Domains	Classes	Samples	Descriptions
PACS (Li et al., 2017)	Object Recognition	Mixed Type	4	7	9,991	Art, Photo, Cartoon, Sketch
OfficeHome (Venkateswara et al., 2017)	Object Recognition	Mixed Type	4	65	15,588	Art, ClipArt, Product, Real
DomainNet (Peng et al., 2019)	Object Recognition	Mixed Type	6	345	586,575	ClipArt, Infograph, Painting, Quickdraw, Real, Sketch
Rotated MNIST (Ghifary et al., 2015)	Handwritten Digit Recognition	Digits	6	10	70,000	Digits rotated from 0° to 90° with a stepsize of 15°
Coloured MNIST (Ghifary et al., 2015)	Handwritten Digit Recognition	Digits	3	2	70,000	Red or green colours and labels
VLCS (Fang et al., 2013)	Object Recognition	Real Objects	4	5	10,729	Four public datasets, each considered as a domain
Camelyon17-WILDS (Koh et al., 2021)	Medical Imaging	Medical Images	5	2	455,954	Hospitals
GlobalWheat-WILDS (Koh et al., 2021)	Object Localisation	Wheat Images	47	1	6,515	Acquisition sessions (time, location, sensor)
Auto Arborist (Beery et al., 2022)	Urban Forest Monitoring	Aerial & Street-View	3	334	9,100,000	Cities
PovertyMap-WILDS (Koh et al., 2021)	Poverty Prediction & Mapping	Satellite	46	$\mathbb{R}$	19,669	Countries
FMoW-WILDS (Koh et al., 2021)	Land-Use Classification	Satellite	16	62	523,846	Years
DSGR (ours)	Land-Use Classification	Satellite	6	41	362,582	Global geographic regions

Table 2 Classes that are included and excluded from DSGR during the preprocessing stage

Included in DSGR
Airport, Airport hangar, Airport terminal , Amusement park, Archaeological site, Barn, Burial site, Crop field, Dam, Educational institution, Electric substation, Factory or powerplant, Fire station, Flooded road, Gas station, Golf course, Ground transportation station, Helipad, Lighthouse, Office building, Oil or gas facility, Park, Parking lot or garage, Place of worship, Police station, Port, Prison, Railway bridge, Recreational facility, Road bridge, Shipyard, Shopping mall, Single-unit residential, Smokestack, Stadium, Storage tank, Surface mine, Swimming pool, Tower, Water treatment facility, Wind farm.
Excluded from DSGR
Aquaculture, Border checkpoint, Car dealership, Construction site, Debris or rubble, Fountain, Hospital, Impoverished settlement, Interchange, Lake or pond, Military facility, Multi-unit residential, Nuclear powerplant, Race track, Runway, Solar farm, Space facility, Toll booth, Tunnel opening, Waste disposal, Zoo.

number of classes in *Train* and *Test* splits became identical, the number of classes in the *Validation* split was notably lower than those of the *Train* and *Test* sets. Hence, for such unique cases, we further divided the *Train* split of the classes which were missing from the *Validation* using a 70 : 30 ratio, where 70% of the sequences in original *Train* split remained as they are and 30% of the sequences were added to the *Validation* split.

Cross-Domain Preprocessing

Besides the within-domain cross-split mismatch between the number of classes, we have observed that there was also an imbalance between the number of classes across domains. For example, the *Tunnel-Opening* class existed in the *Train* set of AS, but not in AF.

To resolve the issues of discrepancies in the number of classes cross-domains and the cases where both cross-domains and cross-split discrepancies existed, we dropped the classes that were not present in any of the splits (*Train*, *Validation* or *Test*) in a specific domain with respect to the other domains. Following the previous example, if the *Tunnel-Opening* class was not present in the *Train* set of AF but existed in the *Train* set of AS, then we dropped it. We also cross-checked all the different combination of splits within-domains and cross-domains for all the classes and rebalanced them using the aforementioned technique accordingly. Subsequently, it was observed that a number of classes in some domains contained very few data samples compared to other domains. Therefore, those classes were dropped from the dataset, as well. Table 2 shows the classes included in DSGR, and those dropped throughout the preprocessing stage.

Furthermore, it can be argued that even with the attempt of rebalancing and reducing the overall data imbalance, DSGR inherits these inevitable imbalance cases from the underlying biases of fMoW when it comes to the distribution of samples

Table 3 Geographic region-wise data partitions used in all experiments

Split	AS	AF	OC	EU	LAC	NA
Training	51, 266	20, 254	8, 610	97, 625	37, 603	65, 065
Validation	7, 243	2, 865	1, 420	14, 853	5, 072	9, 438
Test	7, 264	2, 883	1, 396	14, 644	5, 452	9, 629

per class and per geographic region. This is attributed to the way the classes were defined in fMoW, where there are inconsistencies in the classification granularity. Additionally, an important factor that effect the data imbalance is the scarcity of satellite imagery for underrepresented geographic regions in comparison to others.

To determine whether the performance changes of DL models on OOD data are due to data imbalance or domain shift, we conducted two controlled experiments. In the first, we capped the number of samples per region at 8,000 while maintaining the overall class sample distribution for each region (details in Section 5.7.1). In the second, we ensured a balanced distribution of samples across classes and regions (discussed in Section 5.7.2). Both experiments showed that class imbalance has a negligible impact compared to the significant performance drop caused by domain shift.

While data imbalance is a well-known issue in the DL community and beyond the scope of this paper, we aimed to reduce its effect on the overall OOD performance through the aforementioned techniques[2].

3.2 Final Dataset

This process yielded our dataset, DSGR, with global coverage across all geographic regions. In particular, DSGR includes samples from 49 countries in AS, 52 in AF, 6 in OC, 43 in EU, 43 in LAC, and 3 in NA, as shown in Figure 3, where the intensity of the shade represents the country-level distribution of samples with respect to the geographic region it belongs to. More specifically, DSGR comprises a total of 362, 582 samples from 41 classes across different splits and geographic regions. The total number of samples per split in every geographic region is indicated in Table 3. Whereas, a detailed breakdown on the number of samples per class in each geographic region is shown in Table 4. From this table, it can be observed that geographic regions such as AF and OC consist of smaller data set sizes whereas EU and NA have larger data sets followed by AS and LAC.

Similarly, Figure 5 illustrates the cross-domain training data distributions across the land-use classes in DSGR[3] An interesting observation from this figure is that such a discrepancy in the training set sizes cross-domains usually stems

Table 4 Breakdown of the DSGR dataset with number of samples per class for each geographic region. The heatmap representation is with respect to the samples of the same class across geographic regions

	Geographic Region					
Class	**AS**	**AF**	**OC**	**EU**	**LAC**	**NA**
Airport	730	638	15	70	408	5
Airport hangar	749	546	185	1, 348	867	2, 656
Airport terminal	1, 090	514	280	1, 343	1, 319	1, 243
Amusement park	1, 718	275	137	3, 073	957	1, 439
Archaeological site	1, 144	197	28	1, 496	470	92
Barn	303	190	126	5, 478	473	1, 433
Burial site	624	152	111	2, 891	710	958
Crop field	3, 732	1, 541	497	22, 606	1, 877	1, 660
Dam	1, 022	640	147	1, 025	501	2, 265
Educational institution	4, 207	590	319	3, 183	3, 600	3, 896
Electric substation	1, 300	398	131	2, 494	598	2, 384
Factory or powerplant	668	282	34	2, 226	513	1, 048
Fire station	411	70	140	3, 323	536	2, 362
Flooded road	247	92	130	866	276	1, 112
Gas station	1, 083	172	232	2, 232	1, 545	2, 012
Golf course	674	200	193	1, 226	340	2, 412
Ground trans. station	3, 587	872	1, 181	4, 110	3, 103	1, 258
Helipad	1, 358	221	130	1, 192	1, 131	1, 693
Lighthouse	673	224	238	1, 681	638	1, 366
Office building	920	208	171	3, 001	622	2, 687
Oil or gas facility	2, 199	268	39	1, 677	822	804
Park	1, 103	74	693	1, 828	1, 211	1, 545
Parking lot or garage	2, 643	445	914	6, 049	1, 624	5, 700
Place of worship	6, 845	3, 939	358	4, 278	5, 356	4, 271
Police station	1, 008	401	126	3, 193	1, 287	825
Port	670	244	107	827	428	140
Prison	522	333	88	2, 239	1, 375	1, 225
Railway bridge	1, 282	111	247	2, 555	228	1, 677
Recreational facility	4, 353	603	1, 469	10, 030	4, 389	16, 453
Road bridge	2, 104	190	263	1, 799	464	2, 083
Shipyard	174	34	72	345	47	193
Shopping mall	1, 198	280	561	2, 866	1, 373	1, 440
Single-unit residential	2, 495	7, 015	862	2, 347	807	497
Smokestack	758	224	75	4, 957	181	312
Stadium	1, 608	487	176	1, 346	1, 543	2, 448
Storage tank	638	345	150	3, 766	502	1, 706
Surface mine	1, 141	221	140	3, 185	651	1, 243
Swimming pool	646	1, 487	212	1, 332	3, 451	3, 798
Tower	2, 953	826	82	2, 702	635	2, 348
Water treatment facility	746	249	142	4, 224	338	1, 307
Wind farm	4, 447	204	225	713	931	136
Total	65, 773	26, 002	11, 426	127, 122	48, 127	84, 132

[2] Further information and discussion on DSGR dataset and data imbalance can be found in Section 6.2.

[3] A closer look at the class distribution per-geographic region can be found in A.

from the natural distribution of data availability of these geographic regions. For example, an oil or gas facility might be present in AS in a larger number than in OC.

Likewise, when it comes to cross-class discrepancies, a *Single-Unit Residential* class is expected to have a larger number of samples in comparison to an airport within a geographic region. Such cross-class discrepancies can be considered as an inherent difficulty of land-use and land-cover classification task. Hence, the ultimate goal is to have a model that is robust to both the imbalance of samples cross-class and cross-domains.

4 Domain Generalisation Background

In this section, we lay the foundation to our experiments by, first, introducing the mathematical formulation behind different types of DG in Section 4.1. Next, in Section 4.2, we describe the benchmark DG algorithms we selected for training the DL models. Finally, we explain the metrics used to evaluate and measure the spatial domain shift experienced by the DL models in Section 4.3.

4.1 Mathematical Formulation

Let X be the input feature space and Y be the target label space. One can define a domain, D, with P_{XY} as the joint distribution on $X \times Y$. The goal of DG is to learn a model $f : X \rightarrow Y$ using data samples drawn from the source domain(s) such that when the model is evaluated on the OOD target data, the error on both source (ID) and target (OOD) test data is minimal.

In the generic definition of domain generalisation, also known as *Multi-Source* DG, we assume that multiple (N) similar but distinct source domains, D_s, where $s \in \{1, ..., N\}$ indicates a unique source domain, are available during training. Therefore, the training set, D_{train}, is defined as:

$$D_{train} = \bigcup_{s=1}^{N} D_s$$
$$D_s = \{(x_i^s, y_i^s)\}_{i=1}^{M_s} \tag{1}$$

where x_i^s is the i^{th} sample with label y_i^s and M_s is the total number of training samples belonging to the domain D_s. Each source domain D_s is associated with a joint distribution P_{XY}^s. While the distributions of the source domains might be similar, they are not identical, i.e., $P_{XY}^s \neq P_{XY}^{s'}, s \neq s'$ and $s, s' \in \{1, ..., N\}$.

As a concrete example of the multi-source DG setup using DSGR, we train a DG model on the union of training sets of all geographic regions apart from AS.

Single-Source DG is a special case of this generic definition where $N = 1$. That is, we assume there is only one source domain available during training, and hence, define D_{train} as follows:

$$D_{train} = D_s \tag{2}$$

To demonstrate the single-source DG setup using DSGR, we exclusively use the training set of a single geographic region, for instance AF, for training a DG model.

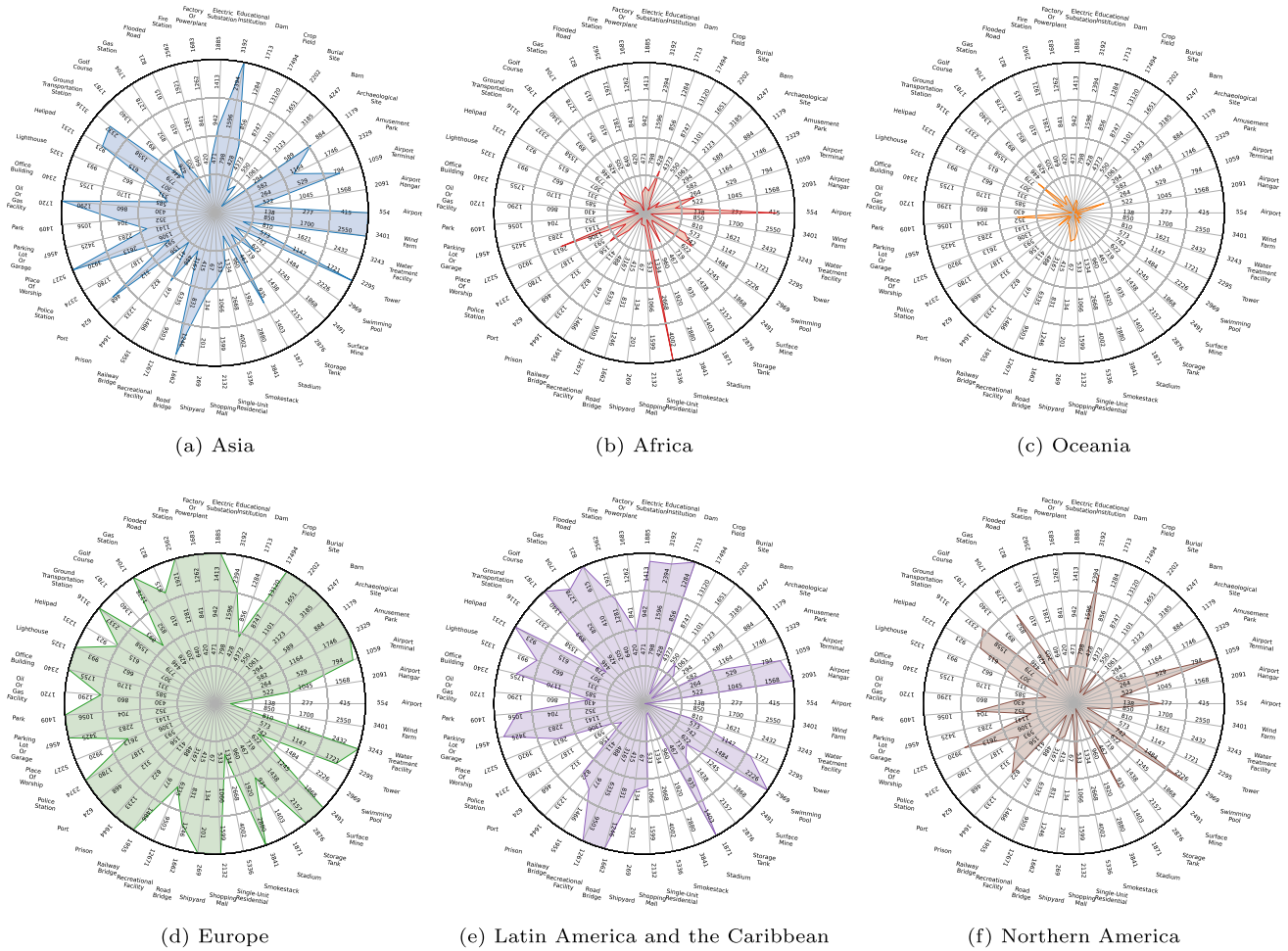
In DG, we define the OOD target domain(s) as D_t, where t represents a target domain such that $t \neq s$ and D_t has a joint distribution P_{XY}^t where $P_{XY}^t \neq P_{XY}^s, \forall s \in \{1, ..., N\}$. Hence, we define the test set as:

$$D_{test} = \{D_t | t \in \{1, ..., K\}\}$$
$$D_t = \{(x_i^t, y_i^t)\}_{i=1}^{M_t} \tag{3}$$

where K is the total number of target domains, x_j^t is the j^{th} sample with label y_j^t and M_t indicates the total number of test samples from the target domain, D_t.

4.2 Benchmark Domain Generalisation Methods

In this study, we evaluate the performance of five SOTA DG methods, namely ERM (Vapnik, 1999), IRM (Arjovsky et al., 2019), Deep CORAL (Sun & Saenko, 2016), group DRO (Sagawa et al., 2019), and CLIPood (Shu et al., 2023), on our DSGR dataset. As previously discussed in Section 2.1, ERM has shown superior generalisation performance compared to modern DG algorithms across various standard DG datasets (Gulrajani & Lopez-Paz, 2021). As a result of these findings, ERM has been established as a standard baseline in all DG studies. Similarly, when it comes to regularisation-based DG methods, IRM (Arjovsky et al., 2019) has become a classical DG method among the DG community. Furthermore, due to the success of Deep CORAL (Sun & Saenko, 2016) in DA, the utilisation of Deep CORAL has been extended to train DG models as well and since then has become another classical DG technique. Moreover, we use group DRO (Sagawa et al., 2019) as one of the recent techniques to address the DG problem by redefining the usage of groups to domains in our experimental analyses. Finally, due to the outstanding performance of CLIPood (Shu et al., 2023) in comparison to the SOTA DG methods on the classical DG benchmark datasets, we consider CLIPood as a foundation-model-based DG technique in our experimental analyses. While the authors attempted to address both DG and Open Class problems using CLIPood, we explicitly focus on investigating the capabilities of this algorithm on the DG problem. More specifically, we aim to examine its performance on our realistic and challenging remote sensing DG dataset, DSGR.



(a) Asia (b) Africa (c) Oceania

(d) Europe (e) Latin America and the Caribbean (f) Northern America

Fig. 5 The representations in this diagram provides an overview of the cross-class training data distribution for each of the geographic regions as well as cross-domain overall data distribution. There are 41 unique classes in DSGR which are indicated as the sectors in this diagram. Whereas, the number of samples per class are indicated dividers along the radius of the diagram

4.3 Evaluation metrics

Given that DG is an emerging field, the evaluation techniques for DG models are also an open area of research. However, researchers have adapted techniques to evaluate the performance of DL models on OOD datasets. *Leave-one-domain-out* is one of the most popular evaluation setups (Gulrajani & Lopez-Paz, 2021), in which one of the domains is left out of the training phase. The left-out domain is then used to test and evaluate the performance of the model without any further fine-tuning. To this end, one of the metrics used by the DG community is *Average Accuracy* (Shen et al., 2021), where the overall performance of the model is computed collectively over the set of target domains D_{test} as follows:

$$\text{Average Accuracy} = \frac{1}{|K|} \sum_{k \in K} \text{Accuracy}_k \tag{4}$$

where k represents a single domain and K is the total number of domains such that $k \in \{1, ..., K\}$.

Besides average accuracy, we introduce through this work another evaluation metric denoted as *performance drop*, which measures the percentage of performance hit experienced by the model when exposed to OOD data from a target domain. This is formulated as follows:

$$\text{Performance Drop (\%)} = -100 \times \frac{A_{OOD} - A_{ID}}{A_{ID}} \tag{5}$$

where A_{ID} and A_{OOD} represent the average accuracy of the combination of models tested on a specific geographic region's ID and OOD test sets, respectively.

Since we will be comparing ID versus OOD performance of DG algorithms using their accuracy values, we opt to use the well-known *harmonic mean* as another evaluation metric to measure the differences between these values. This is inspired by its usage as an evaluation metric in the recent

generalised zero-shot learning studies (Xian et al., 2017; Fu et al., 2019; Chen et al., 2020), where it is used to compute a joint score of the model's performance on training and test sets. The harmonic mean is defined as follows:

$$\text{Harmonic Mean (H)} = \frac{2 \times A_{OOD} \times A_{ID}}{A_{OOD} + A_{ID}} \tag{6}$$

5 Experiments

As introduced in the preceding sections, we aim to shed light on the DG problem for land-use classification task using satellite imagery. We first clarify the experimental settings and parameters in Section 5.1. Then, in Section 5.2, we create the ideal setup under the i.i.d. assumption, where samples from all domains, including the target domain, are available during the training phase. We use the results as an upper bound for comparison with our experiments on DG solutions. After that, we present the results of four experiments that evaluate the following aspects of DG, respectively:

(i) The role of single-source domain training in addressing DG for spatial domains in satellite imagery. This setup aims to understand whether DL models, while obvious to human eyes, interpret different geographic regions through distinctive features even for typical land-use classes such as recreational facilities, educational institutions, stadiums, etc. This provides us with an intuition about the domain shift, if any, experienced by the model (Section 5.3).
(ii) The effect of multi-source training, through collecting data from multiple geographic regions. With this setup, our aim is to explore the DG problem given multiple spatial domains in satellite imagery and whether such a scheme is sufficient for DL models to learn more universal representations of land-use classes that can translate to better generalisable models (Section 5.4).
(iii) The influence of foundation models on SOTA DG techniques. Our objective is to assess whether the strong performance of foundation model contributes to enhancing the SOTA DG models under spatial domain shift. To this end, we examine their performance using two versions of CLIP with different backbone sizes and evaluate the overall performance of these techniques on the proposed DSGR dataset (Section 5.5).
(iv) The effects of an OOD-aware training scheme, which uses a left-out source domain as an OOD validation set during training. Here our goal is to check whether such a scheme leads to an improvement in the models' generalisability across domains in contrast to using all the available source domains as part of the ID validation set (Section 5.6).

Table 5 ID data partitions used in the upper bound experiments

Source	Target	Training	Validation	Test
All Regions	AS	280, 423	40, 891	7, 264
	AF			2, 883
	OC			1, 396
	EU			14, 644
	LAC			9, 629
	NA			5, 452

Furthermore, to ensure comparability of the results, we fix the same training, validation, ID and OOD test sets for each geographic region in all experiments. We repeat each experiment three times with different random seeds, and report the average performance of each model on the ID and OOD test sets, to reduce the impact of randomness. Moreover, to evaluate the models' performance, we use the evaluation metrics: average accuracy, performance drop and harmonic mean, as described in Section 4.3. Along each experiment in the following sections, we discuss the customised setting under which these evaluation metrics are computed.

5.1 Experimental Setup

We trained ERM, IRM, Deep CORAL and group DRO using DenseNet121 (Huang et al., 2017), pretrained on ImageNet (Russakovsky et al., 2015), as the backbone model. Furthermore, we used Adam as the optimiser of choice, a batch size of 64, a learning rate of 0.0001 that decay with a factor of 0.96 per epoch and trained all the models for 50 epochs with early stopping. This setup was followed in all of our experiments. Furthermore, it is worth noting that we have experimented with other classical backbone architectures such as ResNet50 (He et al., 2016) and different hyperparameters, however, we have found that the effects they had on the performance of the techniques were negligible.

For CLIPood, we followed the recommended setup in the original paper (Shu et al., 2023) in all of our experiments since we aimed to evaluate the performance of the-off-the-shelf algorithm without rigorous fine-tuning. The reason behind this choice is that hard fine-tuning would hurt the generalisability of the model to other datasets. However, we have introduced an additional, larger backbone model of CLIP, ViT L/14. Moreover, we used the same hyperparameters for ERM with CLIP as a backbone model and CLIPood to warrant a fair comparison.

In terms of the computation power, we trained all of our models on NVIDIA A100 GPUs with 80GB memory, NVIDIA V100 GPUs with 32GB memory and NVIDIA V100 GPUs with 16GB memory.

Table 6 Accuracy results (in %) of the upper bound scenario

Algorithm	Target					
	AS	AF	OC	EU	LAC	NA
ERM	67.3 ± 0.2	73.2 ± 1.3	66.3 ± 1.2	68.3 ± 0.1	61.6 ± 1.1	67.4 ± 0.8
IRM	66.1 ± 1.0	73.4 ± 1.2	65.9 ± 1.0	67.8 ± 0.8	61.6 ± 1.1	68.3 ± 0.2
Deep CORAL	68.0 ± 1.2	75.3 ± 1.3	69.8 ± 0.2	69.0 ± 0.5	63.7 ± 0.3	68.8 ± 0.4
Group DRO	66.4 ± 0.5	70.7 ± 0.6	65.6 ± 2.1	68.9 ± 0.7	59.6 ± 0.8	67.7 ± 0.8
CLIPood	56.3 ± 0.4	62.3 ± 1.1	64.0 ± 0.5	57.3 ± 0.3	48.8 ± 0.6	56.4 ± 0.6

Table 7 Single-source DG accuracy results (in %) of ERM where the diagonal cells indicate the performance on the ID test sets. Whereas, the off-diagonal cells indicate the performance on the OOD test sets

Source	Target					
	AS	AF	OC	EU	LAC	NA
AS	**65.6 ± 0.3**	41.6 ± 0.2	42.1 ± 3.1	41.0 ± 1.3	37.7 ± 0.2	41.3 ± 0.9
AF	33.8 ± 0.9	**70.2 ± 0.6**	38.7 ± 2.2	30.1 ± 1.3	31.8 ± 0.8	29.5 ± 1.4
OC	25.0 ± 0.5	19.9 ± 1.5	**58.0 ± 1.0**	32.0 ± 0.7	25.9 ± 0.1	34.2 ± 0.2
EU	41.2 ± 0.9	28.6 ± 3.3	49.3 ± 0.7	**69.1 ± 0.3**	35.4 ± 1.4	50.8 ± 1.0
LAC	38.3 ± 1.6	34.9 ± 0.1	45.7 ± 2.5	35.0 ± 1.3	**57.5 ± 1.2**	38.3 ± 1.6
NA	33.0 ± 1.9	22.4 ± 1.4	46.4 ± 1.1	46.4 ± 1.3	30.9 ± 2.3	**68.1 ± 0.6**

5.2 Upper Bound Analysis

In the ideal scenario that satisfies the i.i.d. assumption, all the distributions that are seen by the model during test time are available during the training phase, i.e., the target domain is included as part of the training set. Hence, the model is always evaluated on an ID test set.

In order to understand the upper bound of the model's performance under the best-case scenario on DSGR, we trained the selected DG algorithms, namely, ERM, IRM, Deep CORAL, group DRO and CLIPood, on the combined training sets of all the geographic regions (domains). Similarly, we used the combined validation set to assess the performance of the model during the training phase. Afterwards, the models were tested on each geographic region (domain) separately as previously described in Table 3 (Section 3.2). Table 5 presents the number of samples in each set used in these experiments. Finally, the average accuracy was computed as shown in Table 6.

The results show that, under the i.i.d. assumption, the difference in performance between ERM, IRM, Deep CORAL, and group DRO is relatively minimal. However, CLIPood demonstrates a weaker ID performance in comparison to all the other DG algorithms on DSGR. This can be attributed to the more realistic and challenging nature of DSGR, which is also reflected in the overall low accuracy scores (i.e., mostly less than 70%) achieved by all the models across all geographic regions. We consider these results as the upper bound to compare against in the next set of experiments.

5.3 Single-Source Domain Generalisation

In this experiment, we explored the role of single-source training in addressing DG for spatial domains in satellite imagery. Following the definition mentioned in Eq. (2), we trained a model on a single source domain, D_{train}, and assessed its OOD performance on all the other *unseen* target domains, D_{test} as defined in Eq. (3), separately. For example, if the model was trained on the training set of the source geographic region AS, the validation set of AS was also used to assess the training progress. Whereas, during the testing phase, the trained model was evaluated on the test set of AS to obtain its ID performance as well as on the test sets of all the other unseen target geographic regions (i.e., AF, OC, EU, LAC, and NA) individually to obtain its OOD performance. The breakdown of the dataset for this experiment is presented in Table 3.

For each algorithm, we trained a total of six models, where the training for each model was on a unique geographic region. This resulted in a total of 108 evaluation experiments per algorithm. For brevity, we report only the average performance of ERM on each individual testing dataset in Table 7.[4] It can be observed that, for all the geographic regions, there is a significant generalisation gap between the OOD and ID performance. For example, for AF, the ID performance is around 70.2%, whereas the OOD performance ranges from 19.9% to 41.6% when OC and AS are used as the single-source domain, respectively, resulting in an average OOD performance of 29.5% on AF test set.

[4] The average performances of IRM, Deep CORAL, group DRO and CLIPood can be found in B.1.

Table 8 Single-source DG performance drop (%) analysis of ERM, IRM, Deep CORAL, group DRO and CLIPood

Algorithm	Target																					
	AS				AF				OC				EU				LAC				NA			
	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H
ERM	65.6	34.2	**48**	**45**	70.2	29.5	**58**	**42**	58.0	44.4	**23**	**50**	69.1	36.9	**47**	**48**	57.5	32.3	**44**	**41**	68.1	38.8	**43**	**49**
IRM	65.0	32.9	**49**	**44**	70.2	29.0	**59**	**41**	57.2	43.4	**24**	**49**	68.4	36.3	**47**	**47**	57.2	31.8	**44**	**41**	67.6	38.1	**44**	**49**
Deep CORAL	65.6	34.3	**48**	**45**	69.5	30.3	**56**	**42**	58.0	45.3	**22**	**51**	70.0	37.6	**46**	**49**	58.5	32.9	**44**	**42**	68.2	39.6	**42**	**50**
Group DRO	64.3	33.2	**48**	**44**	69.6	29.0	**58**	**41**	56.1	44.1	**21**	**49**	68.2	36.0	**47**	**47**	57.0	31.7	**44**	**41**	66.4	38.3	**42**	**49**
CLIPood	61.9	43.6	**30**	**57**	70.4	42.0	**40**	**56**	60.9	54.7	**10**	**64**	63.2	45.5	**28**	**60**	52.8	40.6	**23**	**54**	64.9	47.4	**27**	**62**

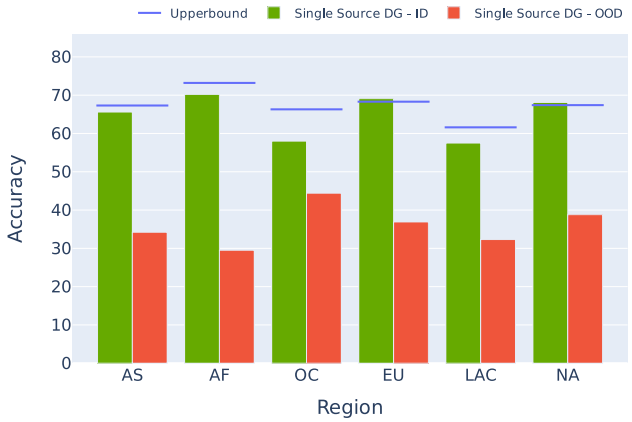


Fig. 6 Single-source DG for ERM evaluated on OOD data versus ID data with respect to the upper bound using DSGR

Furthermore, Figure 6 compares the ID and OOD performance of all the models to the upper bound. As anticipated, for all the geographic regions, the ID performance of the models is comparable to the upper bound. However, when faced with the OOD test set, it can be observed that, a model trained on a single source domain, always experiences a sharp decline in performance compared to the ID test sets which, consequently, yields a large generalisation gap between the OOD and the upper bound. One possible explanation for such a performance drop is that the representation of the OOD data is far from that of the ID data in the latent space.

To further understand this relation in the latent space, Figure 7 represents a t-SNE projection of test data samples of two classes, namely *Oil and Gas Facility*,—one of the most difficult classes in the dataset (further discussed in Section 5.7.6)—, and *Single-Unit Residential* from all geographic regions. The feature embeddings are obtained using an ERM model trained on EU. Since the feature extractor model is originally trained on EU as a single-source domain, we see that it can clearly distinguish the test samples from EU (ID) as two distinct and separable clusters corresponding to the two land-use classes. However, when we investigate the feature projections of the test data samples from the other geographic regions (OOD), we observe that the samples belonging to the same class but from OOD geographic regions are not aligned well (i.e., clustered tightly) with the corresponding class samples from the ID (EU) test set due to domain shift. Consequently, this phenomenon reduces the cross-class separation and increases the confusion between classes, leading to the significant drop in OOD performance of the model.

When it comes to the overall analysis of all the algorithms on DSGR, Table 8 summarises each algorithms' performance in terms of average accuracy on ID and OOD test sets as well as the performance drop and the harmonic mean. Put specifically, the ID performance is measured as the accu-

racy of the model trained and tested on the same geographic region. Whereas, OOD test is computed by averaging the accuracy of the models trained on all the different geographic regions except for the OOD target geographic region (Eq. (4)) and tested on the excluded OOD target geographic region. Finally, the performance drop (Eq. (5)) and the harmonic mean (Eq. (6)) indicate the overall performance between ID and OOD test results for each geographic region.

It can be observed from Table 8 that the results are consistent, with minor differences, among the four benchmark algorithms ERM, IRM, Deep CORAL and group DRO. Whereas, CLIPood performs worse than ERM on the ID test set on majority of the geographic regions apart from AF, where the average ID performance between CLIPood and ERM is comparable. However, CLIPood outperforms the other algorithms on the average OOD test sets where the performance drop of CLIPood is notably lower. One possible explanation for such a behaviour is that CLIPood is able to generalise to unseen domains but underperforms on ID data. This might be an undesirable characteristic, especially, when the model is deployed in an environment, where majority of the time, it will be faced with ID cases and occasionally OOD cases. Therefore, the aim of a good DG algorithm is to achieve a good OOD performance without jeopardising the ID performance.

To evaluate the impact of data imbalance on domain generalisation and to investigate whether the performance degradation of DL models on OOD data is due to data imbalance or domain shift, we conducted two controlled experiments. In the first experiment, the number of samples per region was capped at 8,000 while preserving the overall class distribution per region (further details can be found in Section 5.7.1). In the second experiment, we ensured that the sample distributions across classes and regions are equivalent and balanced (details can be found in Section 5.7.2).

5.4 Multi-Source Domain Generalisation

Unlike the single-source DG experiment discussed in Section 5.3, where the models were trained on single source domains, in this experiment, we shift our focus to a more realistic scenario, where we have multiple source domains available during the training phase. We follow the leave-one-domain-out setup and create different dataset groups defined in Eq. (1) with $C = \{\text{AS}, .., \text{NA}\}$ and each $s = \{C - t\}$ represents a combination of the all the geographic regions except for the left-out OOD target geographic region, t. Table 9 presents the number of samples per split for each group in this experiment.

For each benchmark algorithm, we trained a total of six models, each on one of the groups, $D_{\{C-t\}}, t \in \{AS, .., NA\}$, and validated the performance of each model using the ID validation set during training. Moreover, each model training was repeated three times with different random seeds to ensure the robustness of results. Then, we evaluated the trained models on both ID and OOD testing sets individually during the testing phase. This resulted in a total of 540 evaluation experiments.

Table 10 presents the detailed breakdown of ERM's performance results on each group.[5] From the outcomes of this set of experiments, one could observe a significant performance drop when testing the DG algorithms on OOD test data even when training the DG algorithms on multiple-source domains.

To analyse the performance with respect to the other algorithms, we compiled the average OOD and ID performances along with the performance drop and the harmonic mean between the ID and OOD test sets per geographic region for all five algorithms in Table 11. An interesting observation that stands out is that while CLIPood had achieved an outstanding performance on traditional datasets like PACS, where it outperformed the DG algorithms with around 97.1% OOD accuracy (Shu et al., 2023), the average OOD performance of CLIPood on majority of the geographic regions in DSGR, apart from OC, is lower than that of ERM. Whereas, its performance is comparable to ERM when evaluated on AS. Furthermore, when it comes to the ID performance of CLIPood trained on multiple source domains, we draw a similar conclusion to that made in the single-source DG experiment, where its performance is consistently significantly lower than that of the other SOTA DG algorithms.

When comparing the aforementioned results with those found in the single-source DG experiment, as presented for ERM in Table 12,[6] two crucial conclusions can be drawn. Firstly, comparing the results of the ID test sets in single-source DG with the results of ID test sets in multi-source DG, one can observe that in the majority of the cases, apart from EU, there is an improvement in the overall performance of the model. Secondly, a similar observation is found when evaluating on the OOD test sets, where the average performance drop of the models in multi-source DG is noticeably lower than that of in single-source DG. However, in this experiment, the improvement in the performance, with respect to the upper bound, is reflected in all the different geographic regions as illustrated in Figure 8. This indicates that training on multiple source domains improves the generalisability of DL models. As an example, in single-source DG, the average accuracy of all the models that are not trained on AS but tested on AS is 34.2%. Whereas, the model which is not trained on AS but tested on AS in multi-source DG has a accuracy of

[5] Since the difference between the performance of ERM in comparison to the other algorithms is small, we leave the tables with breakdown results of IRM, Deep CORAL, group DRO and CLIPood to B.2 for brevity.

[6] The comparison of ID versus OOD performance between the two experiments for the rest of algorithms can be found in B.2.

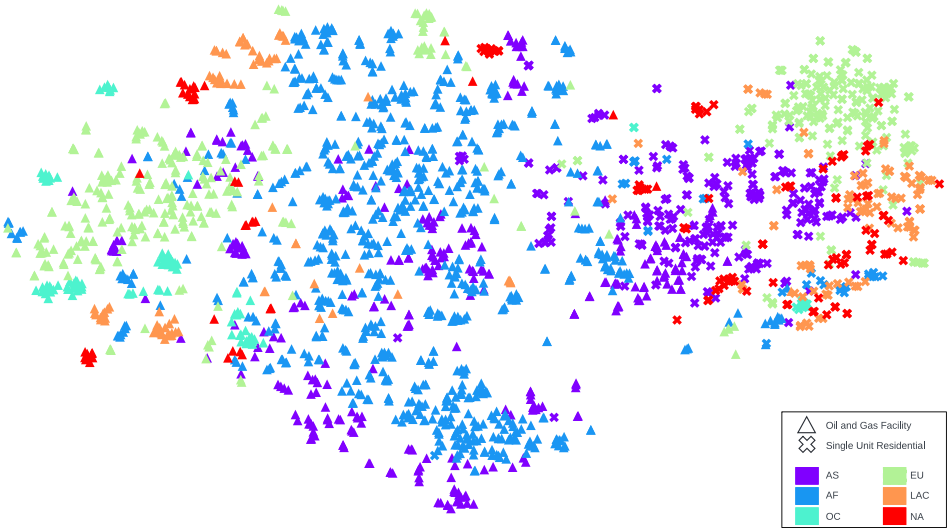


Fig. 7 t-SNE projection of two classes from DSGR where an ERM model is trained on EU (ID) and evaluated on the other OOD geographic regions

Table 9 Data partitions for multi-source domain generalisation experiment

Sources	Target	Training	ID Validation	ID Test	OOD Test
$D_{\{C-\text{AS}\}}$	AS	229, 157	33, 648	34, 004	7, 264
$D_{\{C-\text{AF}\}}$	AF	260, 169	38, 026	38, 385	2, 883
$D_{\{C-\text{OC}\}}$	OC	271, 813	39, 471	39, 872	1, 396
$D_{\{C-\text{EU}\}}$	EU	182, 798	26, 038	26, 624	14, 644
$D_{\{C-\text{LAC}\}}$	LAC	242, 820	35, 819	35, 816	5, 452
$D_{\{C-\text{NA}\}}$	NA	215, 358	31, 453	31, 639	9, 629

Table 10 Multi-source DG accuracy results (%) for ERM. The boldface cells indicate the performance of each model on the OOD test set. Whereas, the off-diagonal scores reflect the ID test performances

Source	Target					
	AS	AF	OC	EU	LAC	NA
$D_{\{C-\text{AS}\}}$	**50.8 ± 0.7**	72.0 ± 0.8	67.2 ± 0.6	69.1 ± 0.6	60.9 ± 0.3	68.6 ± 0.4
$D_{\{C-\text{AF}\}}$	67.7 ± 0.5	**49.3 ± 0.9**	66.0 ± 1.7	69.0 ± 0.2	61.8 ± 0.2	68.6 ± 0.8
$D_{\{C-\text{OC}\}}$	67.4 ± 0.3	73.6 ± 1.0	**58.2 ± 1.5**	68.8 ± 0.2	61.9 ± 0.6	68.4 ± 0.2
$D_{\{C-\text{EU}\}}$	67.0 ± 0.9	73.1 ± 1.1	66.2 ± 1.5	**51.7 ± 1.0**	61.4 ± 0.8	67.1 ± 0.7
$D_{\{C-\text{LAC}\}}$	66.9 ± 0.5	72.1 ± 1.5	65.8 ± 1.0	68.8 ± 0.3	**48.6 ± 1.3**	68.6 ± 0.2
$D_{\{C-\text{NA}\}}$	66.5 ± 0.7	71.5 ± 2.0	66.6 ± 1.3	68.8 ± 0.4	61.0 ± 0.7	**55.8 ± 1.1**

50.8% for ERM. This is equivalent to an improvement of approximately 49%.

Another interesting observation is that AF, an underrepresented region, had a significantly better performance in multi-source DG in comparison to single-source DG on the OOD data, with an increase of 67.1% in its average performance. A plausible explanation is that the satellite imagery captured of Africa looks very different from the other regions. Therefore, one could hypothesis that increasing the diversity of the training set will lead to better generalisation. We discuss in Section 6.4 potential ways to further improve the performance of the underrepresented regions for single-source DG.

To ensure that the observations were not a results of the discrepancy in training data size between single-source and multi-source experiments, we conducted additional experiments under a controlled setup, discussed further in Section 5.7.1, where we fixed the number of training samples for both experiments, single-source and multi-source DG.

5.5 Impact of Foundation Models on Domain Generalisation

Foundation models play a big role in generalisation, which is attributed to being pre-trained on large and diverse datasets in addition to having larger model capacity in comparison to the well-known DL models such as DenseNet (Huang et

Table 11 Multi-source DG performance drop (%) analysis of ERM, IRM, Deep CORAL, group DRO and CLIPood

Algorithm	Target AS				AF				OC				EU				LAC				NA			
	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H
ERM	67.1	50.8	**24**	**58**	72.5	49.3	**24**	**59**	66.3	58.2	**32**	**62**	68.9	51.7	**25**	**59**	61.4	48.6	**21**	**54**	68.2	55.8	**18**	**61**
IRM	66.6	50.7	**24**	**58**	73.2	47.3	**35**	**57**	66.0	59.4	**10**	**63**	67.9	51.4	**24**	**59**	61.1	48.6	**20**	**54**	67.6	56.3	**17**	**61**
Deep CORAL	67.5	50.4	**25**	**58**	75.2	49.4	**34**	**60**	68.5	58.0	**15**	**63**	68.2	52.1	**24**	**59**	62.9	47.4	**25**	**54**	69.2	56.0	**19**	**62**
Group DRO	66.5	50.0	**25**	**57**	69.7	49.4	**29**	**58**	64.5	59.5	**8**	**62**	68.9	51.7	**25**	**59**	59.7	47.7	**20**	**53**	67.5	55.3	**18**	**61**
CLIPood	56.7	51.1	**10**	**54**	63.5	45.7	**28**	**53**	64.2	61.2	**5**	**63**	58.0	51.0	**12**	**54**	49.7	45.0	**9**	**47**	57.3	52.2	**9**	**55**

Table 12 Comparison results of ERM: Multi-source vs. single-source DG for DSGR

	Target					
	AS	AF	OC	EU	LAC	NA
ID % Increase	2.3	3.3	14.2	−0.3	6.8	0.1
OOD % Increase	48.3	67.1	31.1	40.1	50.4	43.8

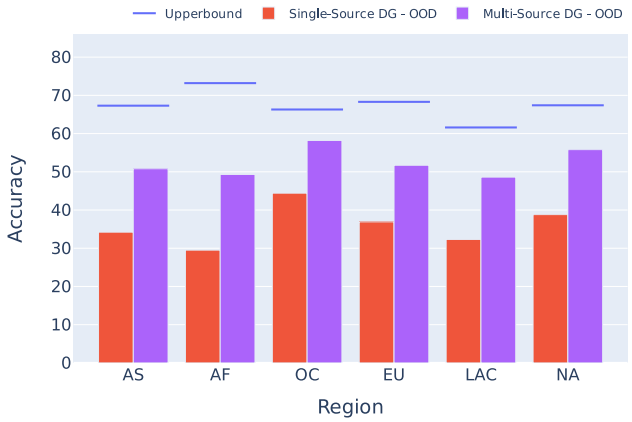


Fig. 8 Comparison results of ERM: Single-source vs. multi-source DG for DSGR

al., 2017) and ResNet (He et al., 2016). In CLIPood (Shu et al., 2023), the authors use CLIP (Radford et al., 2021) as the backbone model to their algorithm and show that their algorithm outperforms the SOTA DG techniques on the traditional DG benchmark datasets. However, we have observed, through the preceding experiments, that when faced with a challenging and realistic dataset such as DSGR, CLIPood does not consistently outperform the classical ERM with DenseNet121 as its backbone model. Following a similar intuition, in this section we examine the effect of introducing CLIP as a backbone to ERM, as opposed to DenseNet121, on its DG performance using a challenging DG dataset such as DSGR. We evaluate ERM with two different backbone CLIP models each with a vision transformer (ViT) of types B/16 and L/14 under single-source DG and multi-source DG experiments.

Table 13 presents the results of the single-source DG experiment on DSGR using different backbone models, namely, ERM with DenseNet, CLIP ViT B/16 and CLIP ViT L/14 as well as CLIPood with CLIP ViT B/16 and CLIP ViT L/14. One can draw the following three conclusions. Firstly, by analysing the effects of the different backbone models for ERM using the harmonic mean, shown in Figure 9, it is clear that ERM with either versions of CLIP as a backbone outperforms, by a large margin, the traditional ERM with a DenseNet121 backbone. Secondly, going from ViT B/16 to ViT L/14 improves the performance of both ERM and CLIPood. However, an interesting observation is that as

the size of ERM with CLIP as a backbone increases, the gain in the generalisation reduces when compared with the gain between either of the backbone models and ERM with DenseNet121. Finally, it can be observed, that ERM with CLIP ViT L/14 outperforms CLIPood with CLIP ViT L/14 on all the geographic regions, apart from AF, where the performance is considered comparable to that of CLIPood with CLIP ViT L/14.

In a similar vein, Table 14 presents the performance of ERM with DenseNet121, CLIP ViT B/16 and CLIP ViT L/14 as backbones as well as CLIPood with CLIP ViT B/16 and CLIP ViT L/14 as backbones in multi-source DG. We observed that ERM with the smaller version of CLIP ViT B/16 outperforms, in terms of ID and OOD, ERM with DenseNet121 and CLIPood with CLIP ViT B/16 in all the geographic regions and surprisingly beats CLIPood with CLIP ViT L/14 as well in majority of the geographic regions apart from OC, where the difference in performance is minor. Whereas, Figure 10 shows that ERM with CLIP ViT L/14 outperforms CLIPood with CLIP ViT L/14 and consequently all the other models in all the metrics. Moreover, it is important to note that the performance drop of training ERM with CLIP ViT L/14 in this experiment is lower than single-source DG training when evaluated on both ID and OOD respectively. Contrarily, CLIPood's performance deteriorates significantly when evaluated on ID data under the multi-source setup. One plausible explanation to such undesired performance hit is that CLIPood, as discussed in Sections 5.3 and 5.4, is not robust when evaluated on ID data, especially when trained using data samples from multiple domains, each with a different underlying distribution.

Given the impact of foundation models on DG algorithms for the land-use classification task is underexplored, we extend our analysis to investigate the effects of incorporating foundation models with DG algorithms for land-use classification task that suffers from temporal domain shift in Section 5.7.3. Our intuition behind such an analysis is to explore whether their impact will be consistent across spatial and temporal domain shifts.

5.6 OOD-Aware Multi-source Domain Generalisation

Previous experiments investigated the conventional experimental setup in which we validated the performance of the DL model on the ID validation set during training. However, the authors of FMoW-WILDS (Koh et al., 2021) put forward the idea of introducing a unique OOD validation set, which effects the early-stopping criteria and the learning rate, and consequently, the training process and the resulting model. Their findings suggests that using an OOD validation set helps in improving the generalisation of the model. While this observation might hold true for the unique categorisation of domains in FMoW-WILDS (Koh et al., 2021),

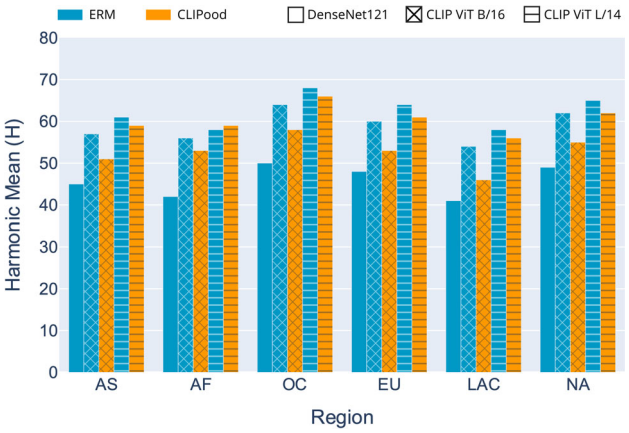


Fig. 9 Foundation models comparison as a backbone for ERM vs. CLIPood in single-source training using DSGR

such as temporal domain shift, we investigate if this claim holds also for our definition of domains with spatial domain shifts. Therefore, in this experiment, we omitted two unique domains from the training set. The first is used for model validation purposes, OOD validation set, and the second is reserved as OOD test set. We then repeated the experiments for all the different combinations of geographic regions. Each of these combinations is denoted as $D_t \setminus v$, which represents all the regions apart from the left out OOD test and validation regions, t and v respectively. Table 15 presents the number of data samples in each of these combinations. As an example of the training setup for the combination which has a unique left-out-domain, AF, reserved for OOD testing, with the remaining five geographic regions, we train five models each time using one of the five geographic regions as the OOD validation set, v.

For each benchmark algorithm, we trained a total of 30 models and repeated the experiment with three different seeds. This resulted in 450 trained models. We ran the evaluation experiments 2700 times to ensure that we cover all the different combinations. The initial observations from the results, presented in Table 16, are aligned with the findings of the single-source DG (Section 5.3) and multi-source DG (Section 5.4) experiments discussed previously, where there is a clear performance drop between the ID and OOD test sets.

However, unlike the observations found in FMoW-WILDS (Koh et al., 2021), our experiments revealed that using an OOD validation set does not improve the performance of the model on the target OOD test set. On the contrary, it hits the performance of the model in majority of the cases as illustrated for ERM in Table 17. One explanation to such a performance hit is related to the decrease in the number of domains used for training due to reserving one domain for OOD validation. Consequently, given that the model under this setup is exposed to fewer and less diverse data distribu-

Table 13 Foundation models as a backbone for DG algorithms in single-source training

Geographic Region	Test	Algorithm and Backbone				
		ERM			CLIPood	
		DenseNet121	CLIP ViT B/16	CLIP ViT L/14	CLIP ViT B/16	CLIP ViT L/14
AS	ID	65.60	73.02	**77.76**	61.87	71.48
	OOD	34.20	46.58	**50.56**	43.58	50.03
AF	ID	70.20	79.69	**82.02**	70.42	77.95
	OOD	29.50	42.66	44.87	41.95	**47.33**
OC	ID	58.00	69.79	**72.40**	60.91	69.10
	OOD	44.40	59.36	**63.62**	54.74	63.41
EU	ID	69.10	75.79	**77.80**	63.20	69.30
	OOD	36.90	49.91	**54.80**	45.47	**54.86**
LAC	ID	57.50	66.98	**70.67**	52.77	66.80
	OOD	32.30	44.81	**48.70**	40.63	47.71
NA	ID	68.10	75.29	**76.46**	64.95	70.02
	OOD	38.80	52.18	**57.09**	47.38	56.12

Table 14 Foundation models as a backbone for DG algorithms in multi-source training

Geographic Region	Test	Algorithm and Backbone				
		ERM			CLIPood	
		DenseNet121	CLIP ViT B/16	CLIP ViT L/14	CLIP ViT B/16	CLIP ViT L/14
AS	ID	67.09	75.65	**78.33**	56.65	63.62
	OOD	50.77	62.76	**64.71**	51.11	56.40
AF	ID	72.48	81.56	**84.04**	63.54	71.72
	OOD	49.27	57.10	**57.95**	45.67	56.11
OC	ID	66.33	76.46	**78.80**	64.17	70.53
	OOD	58.23	68.17	**70.70**	61.20	68.51
EU	ID	68.90	76.76	**78.95**	57.98	65.91
	OOD	51.70	63.15	**66.13**	51.03	59.84
LAC	ID	61.41	70.61	**73.91**	49.65	57.69
	OOD	48.63	59.03	**61.71**	44.99	52.08
NA	ID	68.25	75.94	**78.42**	57.34	64.26
	OOD	55.83	64.97	**67.43**	52.19	59.68

tions during training, it does not generalise as well as when a larger number of the source domains, with diverse data distributions, are used for training purposes. Furthermore, the same is said when analysing the ID performance. Therefore, we do not recommend reducing the number of source domains dedicated for training to create an OOD validation dataset, rather, it is encouraged to use more source domains, where possible, in the training dataset as it has shown to improve the accuracy of the model on the OOD test set.

Following the intuition behind the foundation model experiment presented in Section 5.5, we further explored OOD-aware setup while using different foundation models as backbone in Section 5.7.4.

5.7 Additional Analyses

5.7.1 Domain Generalisation under a Controlled Setup

In the multi-source DG experiment discussed in Section 5.4, D_{train} was defined as the union of multiple source domains which are available during training. Naturally, combining samples from multiple source domains results in a larger training set size in comparison to single-source domain train-

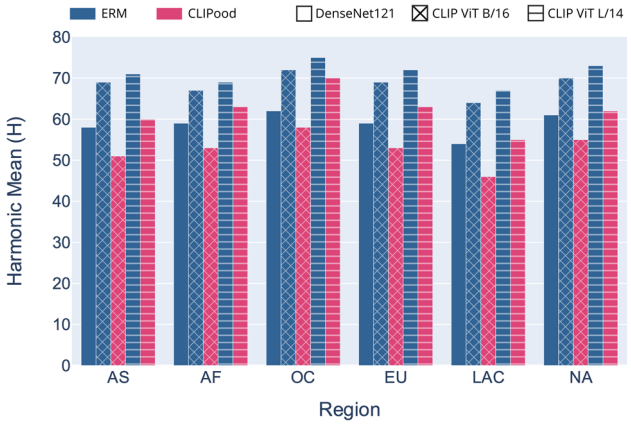


Fig. 10 Foundation models comparison as a backbone for ERM vs. CLIPood in multi-source training using DSGR

ing. Therefore, one might speculate that the improvement in performance is due to having a larger training set rather than a diverse training set. To better understand the main contribution of the aforementioned improvement in performance, we propose a controlled experimental setup, where the number of training samples in single-source domain training is identical to that of multi-source domain training. That is, we cap the number of training samples for both the single-source and multi-source experiments to 8, 000 samples, while maintaining their underlying distributions.

Table 18 presents the performance of the models trained on a single-source domain under this setup. We observe that there is a sharp domain shift between the ID and OOD test sets. This observation is consistent for the controlled multi-source training setup shown in Table 19. More importantly, when comparing the OOD performance of the models on the controlled single-source training versus multi-source training, as illustrated in Figure 11, it seems clear that, despite the *fixed* training set size, the performance of the model that has been exposed to a more diverse dataset during training, through multi-source training, yields a better performance in comparison to the model that have seen a single distribution through single-domain training. Hence, while a larger training set size enhances the performance of the model, so does having a diverse data distribution during training. In particular, for the controlled multi-source experiment we have 1600 samples per-geographic region among the 41 classes, then, the union of the training sets of each geographic region is used for training. Whereas, the single-source setup we have 8000 samples among the 41 classes from a single geographic region that is used during training.

Finally, it can also be observed from Figure 11 that the OOD performance of both single-source and multi-source DG under the controlled setup is lower than those presented for the uncontrolled setup. Such a decrease in performance is expected due to the reduced training set sizes in the controlled setup in comparison to the uncontrolled setup.

Table 15 Data partitions for OOD-aware multi-source domain generalisation experiment

Sources	OOD Val.	# OOD Val.	Target Domain	# OOD Test
$D_{AS} \setminus AF$	AF	2, 865	AS	7, 264
$D_{AS} \setminus OC$	OC	1, 420		
$D_{AS} \setminus EU$	EU	14, 853		
$D_{AS} \setminus LAC$	LAC	5, 072		
$D_{AS} \setminus NA$	NA	9, 438		
$D_{AF} \setminus AS$	AS	7, 243	AF	2, 883
$D_{AF} \setminus OC$	OC	1, 420		
$D_{AF} \setminus EU$	EU	14, 853		
$D_{AF} \setminus LAC$	LAC	5, 072		
$D_{AF} \setminus NA$	NA	9, 438		
$D_{OC} \setminus AS$	AS	7, 243	OC	1, 396
$D_{OC} \setminus AF$	AF	2, 865		
$D_{OC} \setminus EU$	EU	14, 853		
$D_{OC} \setminus LAC$	LAC	5, 072		
$D_{OC} \setminus NA$	NA	9, 438		
$D_{EU} \setminus AS$	AS	7, 243	EU	14, 644
$D_{EU} \setminus AF$	AF	2, 865		
$D_{EU} \setminus OC$	OC	1, 420		
$D_{EU} \setminus LAC$	LAC	5, 072		
$D_{EU} \setminus NA$	NA	9, 438		
$D_{LAC} \setminus AS$	AS	7, 243	LAC	5, 452
$D_{LAC} \setminus AF$	AF	2, 865		
$D_{LAC} \setminus OC$	OC	1, 420		
$D_{LAC} \setminus EU$	EU	14, 853		
$D_{LAC} \setminus NA$	NA	9, 438		
$D_{NA} \setminus AS$	AS	7, 243	NA	9, 629
$D_{NA} \setminus AF$	AF	2, 865		
$D_{NA} \setminus OC$	OC	1, 420		
$D_{NA} \setminus EU$	EU	14, 853		
$D_{NA} \setminus LAC$	LAC	5, 072		

5.7.2 Domain Generalisation under a Balanced Class Distribution Setup

Building on the discussion in Section 5.7.1, in this section, we investigate the impact of data imbalance on the performance of domain generalisation models. Specifically, we aim to understand whether eliminating class imbalance can mitigate or entirely eliminate domain shift.

For our evaluations, we selected the AS and EU regions. We began by rebalancing the dataset by capping the number of examples per class to the minimum observed within these

Table 16 OOD-Aware multi-source DG performance drop (%) analysis of ERM, IRM, Deep CORAL and group DRO

Algorithm	Target AS				AF				OC				EU				LAC				NA			
	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H	ID	OOD	%	H
ERM	65.3	48.8	**25**	**56**	71.9	44.9	**38**	**55**	64.8	56.8	**12**	**61**	67.5	49.2	**27**	**57**	59.8	46.8	**22**	**52**	67.1	53.5	**20**	**60**
IRM	66.0	48.8	**26**	**56**	72.8	45.9	**37**	**56**	65.3	56.7	**13**	**61**	67.7	49.6	**27**	**57**	60.4	46.4	**23**	**52**	67.0	53.5	**20**	**59**
Deep CORAL	67.0	48.8	**27**	**56**	74.5	46.2	**38**	**57**	67.5	57.8	**14**	**62**	68.0	50.4	**26**	**58**	62.2	46.7	**25**	**53**	68.5	54.2	**21**	**60**
Group DRO	65.7	48.1	**27**	**56**	70.5	44.9	**36**	**55**	63.4	55.2	**13**	**59**	68.2	49.3	**28**	**57**	59.2	46.0	**22**	**52**	67.2	53.6	**20**	**60**
CLIPood	57.7	50.5	**12**	**54**	64.5	47.6	**26**	**55**	64.7	61.0	**6**	**63**	58.8	50.9	**13**	**55**	50.7	44.6	**12**	**47**	58.4	52.0	**11**	**55**

Table 17 Comparison results of ERM: OOD-aware vs. multi-source DG

	Target					
	AS	AF	OC	EU	LAC	NA
ID % Increase	−2.7	−0.8	−2.2	−2.0	−2.5	−1.6
OOD % Increase	−3.8	−8.9	−2.4	−4.8	−3.8	−4.1

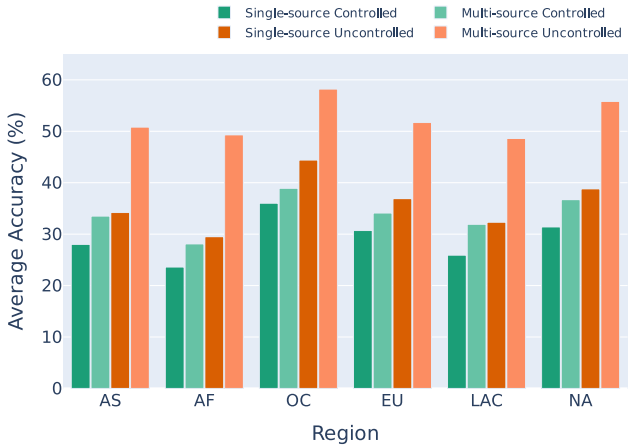


Fig. 11 Comparison results of controlled vs. uncontrolled setup using ERM for multi-source and single-source DG for DSGR

two regions while ensuring that no classes were omitted. As a result, the threshold was set at 58 samples per class, yielding a total of 2,378 training samples per region.

Next, we trained ERM on each region individually, under the single-source setup, and evaluated the model's performance across all regions. We repeated each training three times to ensure the robustness of the results. Table 20 presents the results of these experiments. Aligned with our previous observations, particularly those in Section 5.7.1, eliminating class and data imbalance does not alter the general trend where the model's performance still exhibits a sharp degradation in performance when transitioning from ID to OOD evaluation.

5.7.3 Impact of Foundation Models on Domain Generalisation under Temporal Domain Shift

To further investigate the influence of foundation models on DG algorithms, we repeated the experiments discussed in Section 5.5 for a temporal DG satellite imagery dataset, FMoW-WILDS (Koh et al., 2021). The experiments yielded the results presented in Table 21.

It can observed from the results that the performance of CLIPood with CLIP ViT B/16 on ID and OOD datasets are comparable, where it had the lowest performance followed by CLIPood with CLIP ViT L/14. Whereas, similar to the results previously found using DSGR in Section 5.5, ERM

Table 18 Controlled experiment for single-source DG accuracy results (in %) of ERM where the diagonal cells indicate the performance on the ID test sets. Whereas, the off-diagonal cells indicate the performance on the OOD test sets.

Source	Target AS	AF	OC	EU	LAC	NA
AS	**52.6 ± 0.7**	30.4 ± 0.9	27.5 ± 3.4	30.8 ± 1.4	30.0 ± 0.8	31.8 ± 3.4
AF	28.4 ± 0.5	**65.8 ± 0.7**	30.7 ± 2.4	26.8 ± 0.8	26.9 ± 1.8	27.3 ± 0.6
OC	23.3 ± 1.1	17.5 ± 1.5	**57.2 ± 0.7**	30.0 ± 0.6	24.6 ± 0.4	31.9 ± 1.6
EU	29.5 ± 0.7	19.8 ± 2.3	43.0 ± 1.1	**51.7 ± 0.5**	23.4 ± 2.0	35.7 ± 2.1
LAC	32.7 ± 0.9	31.3 ± 0.6	39.9 ± 1.6	30.0 ± 0.6	**49.5 ± 0.3**	30.3 ± 0.6
NA	26.2 ± 1.1	19.1 ± 1.4	38.8 ± 1.3	35.9 ± 1.1	24.6 ± 0.3	**57.1 ± 0.7**

Table 19 Controlled experiment for multi-source DG accuracy results (in %) of ERM where the diagonal cells indicate the performance on the ID test sets. Whereas, the off-diagonal cells indicate the performance on the OOD test sets

Source	Target AS	AF	OC	EU	LAC	NA
$D_{\{C-\text{AS}\}}$	**33.5 ± 0.8**	58.5 ± 1.2	52.9 ± 1.7	42.5 ± 1.1	40.7 ± 1.9	45.5 ± 0.9
$D_{\{C-\text{AF}\}}$	44.6 ± 0.6	**28.1 ± 1.6**	52.1 ± 2.4	42.3 ± 0.7	40.2 ± 0.4	45.4 ± 1.6
$D_{\{C-\text{OC}\}}$	45.8 ± 0.2	58.9 ± 0.9	**38.9 ± 2.0**	42.6 ± 0.2	39.3 ± 0.5	45.4 ± 0.4
$D_{\{C-\text{EU}\}}$	44.4 ± 0.4	58.5 ± 1.3	51.9 ± 1.2	**34.1 ± 0.8**	40.5 ± 0.3	45.4 ± 0.3
$D_{\{C-\text{LAC}\}}$	45.8 ± 0.4	56.8 ± 3.2	52.8 ± 1.6	43.5 ± 1.0	**31.9 ± 1.5**	46.8 ± 1.6
$D_{\{C-\text{NA}\}}$	44.2 ± 0.7	57.6 ± 0.3	50.0 ± 0.8	41.9 ± 0.8	39.9 ± 0.9	**36.7 ± 0.7**

Table 20 Controlled experiment with Balanced Class Distribution for single-source DG accuracy results (in %) of ERM where the diagonal cells indicate the performance on the ID test sets. Whereas, the off-diagonal cells indicate the performance on the OOD test sets.

Source	Target AS	AF	OC	EU	LAC	NA
AS	**33.4 ± 2.2**	21.6 ± 0.5	17.5 ± 2.4	17.5 ± 1.3	17.8 ± 1.1	19.6 ± 0.4
EU	23.1 ± 1.7	19.2 ± 2.5	28.8 ± 2.4	**35.9 ± 2.1**	18.8 ± 1.1	25.4 ± 1.6

with CLIP ViT L/14 outperforms CLIPood on OOD and ID datasets. Therefore, these results indicate that the performance of CLIPood versus ERM with CLIP is consistent in both spatially or temporally defined domains.

5.7.4 Foundation Models on OOD-Aware Multi-Source Domain Generalisation

We conducted an experiment similar to the one presented in Section 5.5 to examine the influence of foundation model as backbone to SOTA DG algorithm. However, in this experiment, we evaluate their performance under an OOD-aware setting similar to that discussed in Section 5.6.

Table 22 presents the outcomes of this experiment which are aligned with the observations made in Section 5.5, where ERM with either versions of CLIP outperforms both ERM with DenseNet121 as well as both versions of CLIPood significantly. Furthermore, similar to the findings of Section 5.6, having a separate OOD validation set does not positively impact the DG techniques' generalisability, even when coupled with foundation models as backbone.

5.7.5 Impact of Domain Shift on Specialised Remote Sensing Foundation Model

With the rise of general-purpose foundation models, researchers are exploring training and fine-tuning these models on large amount of remote sensing imagery to enhance their performance in downstream remote sensing applications. To understand whether these models are robust under domain shift and their generalisability to unseen domains, we assess the zero-shot of RemoteCLIP (Liu et al., 2024), a Multi-model Large Language Model (MLLM) trained on extensive open-source remote sensing datasets, on DSGR. Furthermore, we compare its performance with our fine-tuned versions of ERM and CLIPood under a multi-source setup, using CLIP ViT L/14 as a consistent backbone.

The experimental results, presented in Table 23, reveal that both the fine-tuned classical method, ERM, and the fine-tuned DG specialised method, CLIPood, significantly outperform RemoteCLIP. These findings are aligned with observations reported by the authors (Liu et al., 2024), who noted similar performance drops during zero-shot evaluations on certain datasets. This indicate that while specialised

Table 21 Foundation models as a backbone for DG algorithms using FMoW-WILDS

Metric	ERM			CLIPood	
	DenseNet121 (Koh et al., 2021)	CLIP ViT B/16	CLIP ViT L/14	CLIP ViT B/16	CLIP ViT L/14
ID	59.70	66.63	**71.02**	50.08	58.39
OOD	53.00	59.86	**64.57**	48.36	55.51
%	11	10	9	3	5
H	56	63	68	49	57

Table 22 Foundation models as a backbone for DG algorithms in OOD-aware multi-source training

Geographic Region	Test	ERM			CLIPood	
		DenseNet121	CLIP ViT B/16	CLIP ViT L/14	CLIP ViT B/16	CLIP ViT L/14
AS	ID	65.29	71.89	**75.07**	57.72	62.34
	OOD	48.84	59.89	**64.32**	50.53	55.75
AF	ID	71.92	76.63	**78.89**	64.50	69.65
	OOD	44.87	56.29	**58.93**	47.58	54.19
OC	ID	64.81	74.83	**77.46**	64.66	70.18
	OOD	56.83	67.15	**70.46**	60.99	67.70
EU	ID	67.50	75.68	**77.99**	58.84	65.66
	OOD	49.20	61.03	**65.53**	50.91	59.39
LAC	ID	59.81	68.99	**73.16**	50.65	58.35
	OOD	46.77	57.03	**60.30**	44.63	51.89
NA	ID	67.10	75.18	**78.16**	58.40	64.77
	OOD	53.53	63.03	**66.16**	51.96	59.35

Table 23 Comparison between a remote sensing MLLM, namely RemoteCLIP, and our fine-tuned versions, under multi-source setup, of ERM and CLIPood. We used a backbone of CLIP ViT L/14 across all the methods.

Setup	Method	Target					
		AS	AF	OC	EU	LAC	NA
Zero-shot	RemoteCLIP (Liu et al., 2024)	24.4	27.2	33.6	25.4	20.4	26.5
Multi-Source Fine-Tuning	CLIPood	56.4	56.1	68.5	59.8	52.1	59.7
	ERM	**64.7**	**57.9**	**70.7**	**66.1**	**61.7**	**67.4**

remote sensing models are versatile, their performance may degrade notably under domain shift.

5.7.6 Class-Wise Analysis of Single-Source DG

In order to further investigate the effects of the domain shift experienced by the model when faced with OOD test data, in this section, we provide a class-wise analysis of single-source DG, where we trained an ERM model on EU and evaluated its performance on both the ID and OOD test set, namely of AF, separately. The impact of the domain shift is immediately visible in Figure 12, in which the model achieves high performance, indicated through the strong diagonal in Figure 12a, on the ID test set as opposed to the case of an OOD test set, shown in shown Figure 12b, where its performance dropped significantly resulting in a high error rate. This is evident for the *Oil and Gas Facility* class for instance, where the model correctly predicted the ID samples 90% of the time and failed to predict the OOD samples correctly. This observation holds for other classes such as *Archaeological site*, *Burial site*, *Barn*, *Flooded road*, *Fire station*, and *Tower*. On the contrary, the classes that are anticipated to have a large similarity in both geographic regions in terms of the architectural features, for example in the case of the *Lighthouse* class, the model have maintained its high performance on the OOD test set. Another interesting observation is made when predicting the *Recreational Facility* class, where the model was able to identify 60% of the instances correctly and 36% misclassified as the class *Stadium* on ID test set. However, it had a better performance on the OOD test set, with 84% of samples

correctly classified. This behaviour could be attributed to the mismatch in the granularity of the original class definitions of fMoW discussed in Section 6.2.

5.7.7 The Influence of Urbanisation on Domain Shift: A Case Study

As discussed in Section 1, the problem of spatial domain shift in satellite imagery, such as DSGR, could be attributed to various domain-specific factors, such as the uniqueness of features of the target regions in terms of urbanisation, development, architectural designs, land-cover, etc. As a case study, we shed the light on the effect of urbanisation on the performance of DL models when evaluated on ID versus OOD data. Therefore, we consider the analysis from a multi-dimensional perspective.

First, we mapped the coordinates of the data in DSGR to their corresponding locations provided through Global Human Settlement Layer (GHSL) [7] in order to categorise whether each sample point is located within an urban or rural area. We observed a consistency between the ratios of training to testing sets across urban and rural areas[8].

Next, we evaluated the models' performance on the new testing sets, specifically designed to distinguish between rural and urban areas across different regions. It is important to note that the models analysed in the previous sections were trained on satellite imagery covering both urban and rural areas. To assess their accuracy across rural and urban areas, we calculated the error ratio using the number of misclassification mistakes the model makes with respect to each area individually. Table 24 presents a breakdown of these errors across rural and urban areas.

We selected the NA region as a case study and present the results as two separate heatmaps for urban and rural areas respectively in Figure 13. Firstly, when analysing the ID performance on NA across urban versus rural areas, we observed that the error rate increases from 27% on rural areas to 34% when evaluated on samples from urban areas. Secondly, in terms of OOD performance, we observed that the model made more errors on urban samples than on rural ones across the OOD testing sets. Additionally, we conducted a class-wise analysis by comparing the confusion matrices across areas for both ID and OOD testing sets and we found that the performance on rural samples was significantly higher than on urban samples for a class such as *Tower*.

This trend, where the model had a higher error rate on images from urban areas compared to rural, was evident across all the other regions, except for three outlier cases: one case where models showed similar performance on both urban and rural areas when evaluated on ID data from AF, and two cases where models trained on AS or AF and evaluated on OC. Furthermore, it is important to note that, apart from the outlier cases, this general trend held regardless of the distribution of urban and rural samples seen by the models during training.

A plausible explanation to the discrepancy between the performance of a model on urban versus rural areas could be attributed to the inherent characteristics of the satellite imagery. For example, in rural areas, images typically contain a single or few buildings with a relatively uniform background, such as a green landscape. In contrast, urban images often feature many buildings within a single image, accompanied by a more complex and cluttered background, resulting in the model's confusion between classes. This phenomenon is illustrated in Figure 1, in which two samples from the *Single-Unit Residential* class can be observed, one captured from a rural area (top) and the other from an urban area (bottom).

5.7.8 Coupling Domain Generalisation with Open-Set Recognition

In open-world scenarios, also known as open-set settings, some classes are not available during the training phase and the objective is to develop a model that remains robust to both seen and unseen classes. Therefore, in this section, we explore this problem while coupled with domain shift to analyse the impact of both new domains and classes on the performance of DG techniques.

To create this setup with DSGR, we followed the approach outlined in open-set literature, such as CoCoOp (Zhou et al., 2022b) and CLIPood (Shu et al., 2023), by randomly splitting classes, using a 50:50 ratio, into seen (*base*) and unseen (*new*) groups while ensuring consistency in classes across regions and splits. This resulted in 21 seen classes dedicated for training and validation, while 20 new classes were reserved for evaluating the model's generalisation. We refer to this variant of DSGR as **DSGR-OS**[9].

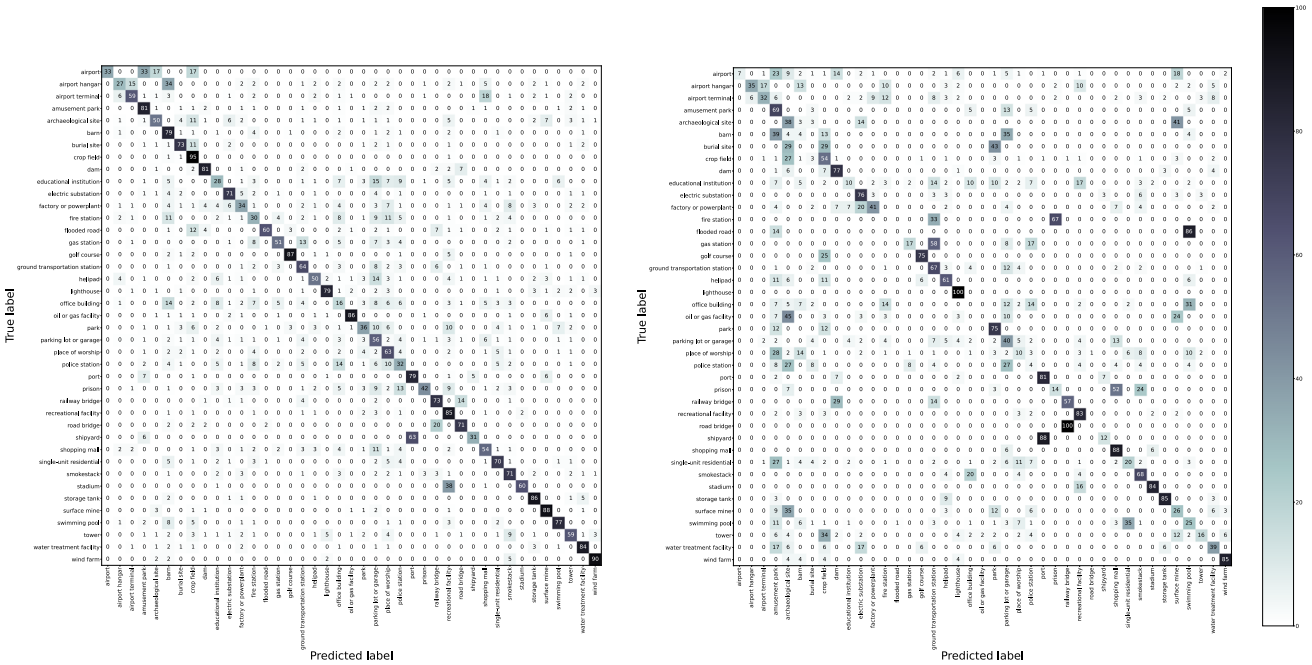
We evaluated two CLIPood variants with different backbone sizes, fine-tuned on the base classes, as it is the only model in this study that supports open-set classification and conducted the experiments in a similar fashion to the single-source DG experiment mentioned above.

Table 25 presents the performance of CLIPood with CLIP ViT L/14 on each region with respect to the base and new class groups. Analysing the base setup independently, where the model is evaluated on a smaller number of seen classes in comparison to the original DSGR, we can observe that domain shifts remain evident across geographical regions, which is aligned with our previous observations. Addition-

[7] https://human-settlement.emergency.copernicus.eu

[8] The distribution of the training and testing splits with respect to urban and rural areas can be found in Appendix D.

[9] The class distribution can be found in Appendix C.



(a) Test: EU (ID) (b) Test: AF (OOD)

Fig. 12 Single-Source DG for ERM per-class performance when the model is trained and evaluated on ID versus OOD. The confusion matrices are normalised and presented in percentages (%) with respect to the true labels

Table 24 Error (%) analysis of ERM with respect to urban and rural areas

	Target											
	AS		AF		OC		EU		LAC		NA	
Source	Urban	Rural	Urban	Rural	Urban	Rural	Urban	Rural	Urban	Rural	Urban	Rural
AS	**43**	22	**58**	47	54	**58**	**66**	51	**64**	48	**56**	55
AF	**68**	57	29	29	51	**57**	**76**	60	**66**	57	**71**	63
OC	**84**	73	**89**	81	**50**	40	**80**	70	**79**	71	**69**	62
EU	**66**	51	**72**	62	**53**	46	**42**	25	**67**	52	**53**	43
NA	**63**	53	**65**	54	**50**	46	**74**	55	**45**	29	**63**	55
LAC	**71**	59	**78**	69	**47**	45	**62**	47	**70**	57	**34**	27

ally, regardless of the backbone architecture, the model's performance is consistently better on ID than OOD under the open-set scenario (new-classes) across AS, EU, LAC, and NA. Whereas, this is not the case for AF or OC. This might be due to the small size and lack of diversity in the training set for those two regions in comparison to others. Another limiting factor could be that, as noted in prior open-set studies, such as CoCoOp (Zhou et al., 2022b) and CLIPood (Shu et al., 2023), the random division of base and new classes does not ensure balanced class difficulty.

More broadly, the combination of open-set recognition with domain generalisation remains an open area of research.

5.8 Summary of Findings

To conclude this section, we summarise the main findings of our experiments as follows:

DSGR dataset reflects the complex and real-world challenges in comparison to standard DG datasets, where the SOTA DG algorithms struggled to maintain a good performance when evaluated on its different OOD test sets. This observation was consistent among all the different experiments and setups conducted in this section.

Training under spatial domain shift with multiple source domains as opposed to a single source domain reduces the performance drop caused by the generalisation gap.

Examining the impact of foundation models on DG revealed that the classical ERM coupled with a foundation

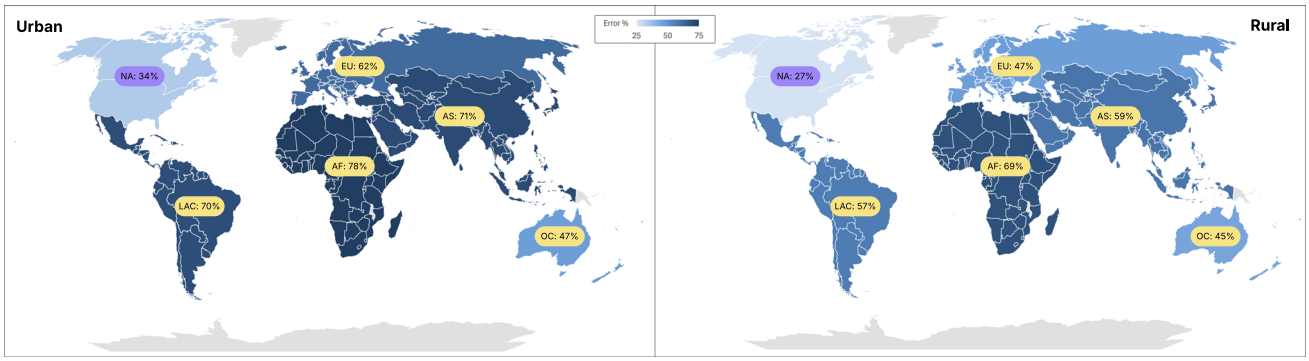


Fig. 13 A DG model trained on LAC and evaluated on ID and OOD regions with respect to urban (left) and rural (right) areas. The darker the shade is, the higher error rate

Table 25 DSGR in an open-set setting. We report single-source DG accuracy (%) for CLIPood ViT-L/14. Diagonal cells in *base* show ID test set performance, while in *new*, they represent open-set performance on ID regions. Off-diagonal cells show performance on test sets for both class groups across OOD regions

	Target								
	AS			AF			OC		
Source	Base	New	H	Base	New	H	Base	New	H
AS	**79.8 ± 0.4**	**41.7 ± 1.9**	55	65.7 ± 1.4	34.3 ± 1.5	45	74.0 ± 1.0	56.8 ± 0.5	64
AF	56.1 ± 1.0	36.8 ± 0.7	44	**83.1 ± 0.8**	31.4 ± 1.1	46	72.3 ± 1.0	51.5 ± 0.5	60
OC	52.2 ± 0.7	35.9 ± 0.2	43	45.0 ± 2.9	31.1 ± 1.4	37	**80.4 ± 1.1**	51.0 ± 0.3	62
EU	59.4 ± 0.5	35.7 ± 0.5	45	57.9 ± 1.9	**36.0 ± 2.8**	44	79.4 ± 0.2	54.3 ± 1.1	65
LAC	64.8 ± 0.1	37.5 ± 0.3	47	46.6 ± 0.1	**36.9 ± 0.8**	41	73.8 ± 0.8	**58.8 ± 1.0**	65
NA	56.2 ± 0.9	34.2 ± 0.8	43	47.3 ± 0.4	29.8 ± 1.6	37	77.9 ± 1.6	52.6 ± 0.2	63
	EU			LAC			NA		
Source	Base	New	H	Base	New	H	Base	New	H
AS	67.0 ± 0.5	**47.2 ± 0.6**	55	59.4 ± 1.0	41.9 ± 0.8	49	67.3 ± 0.5	49.1 ± 0.5	57
AF	58.8 ± 0.5	44.9 ± 0.3	51	51.5 ± 0.1	38.8 ± 0.8	44	60.1 ± 0.4	46.5 ± 1.3	52
OC	61.5 ± 0.5	43.0 ± 0.3	51	**47.8 ± 0.7**	39.7 ± 0.4	43	63.9 ± 0.8	45.9 ± 0.4	53
EU	**77.5 ± 0.0**	**47.3 ± 0.7**	59	53.5 ± 0.9	39.5 ± 0.7	45	68.7 ± 0.5	49.4 ± 0.4	57
LAC	65.4 ± 0.3	**47.5 ± 0.8**	55	**70.6 ± 0.9**	**42.7 ± 0.7**	53	66.7 ± 0.3	48.5 ± 1.2	56
NA	68.2 ± 0.1	42.5 ± 0.7	52	55.3 ± 1.1	38.4 ± 1.1	45	**78.1 ± 0.2**	**49.3 ± 0.3**	60

model, such as CLIP, as its backbone architecture yields an outstanding OOD performance in comparison to other SOTA DG algorithms that were designed explicitly around foundation models like CLIPood.

An OOD-aware training scheme does not improve the overall performance of the DG model. We deduced that rather than leaving one domain out for OOD validation, incorporating it as part of the training domains would yield a better OOD performance.

6 Discussion

6.1 Implications of DSGR in Real-World Applications

In designing DSGR, our aim was to create a DG dataset that reflects scenarios in real-life applications and can be used to assess the robustness of models to spatial domain shift prior to their deployment. Therefore, as previously mentioned, we opted to follow the geographic region categorisation defined by the United Nations (Nations, 2022) in DSGR. Based on our findings in Section 5, we recommend that practitioners consider evaluating their DL based solutions for land-use classification applications on DSGR in order to understand their behaviour when deployed in real-life. Furthermore, doing so can help in assessing and mitigating potential risks that arise when DL models are deployed in new, and many fault intolerant, environments.

The usage of DSGR extends beyond land-use classification to other real-world application. This can be achieved by learning a base-model that is good enough to generalise across geographic regions and transferring this generalisation capability through different types of learning, e.g, transfer learning, to other applications. For instance, DSGR can be used in critical infrastructure detection application, where the model is trained on data from a specific domain, for exam-

ple, region A and deployed in another OOD domain such as continent B. Other real-world examples of applications that can benefit from DSGR during the learning process are road extraction, flood and disaster detection, urban planning and development, among others.

6.2 Data Imbalance in DSGR

While data imbalance was not fully mitigated in DSGR, we aimed to reduce its effect on the overall OOD performance with our preprocessing approach, highlighted in Section 3.1. Also, we have designed the single-source DG experiment (Section 5.3) and the multi-source DG experiment (Section 5.4) to address this issue through the elimination of some geographic regions during training. Since we considered the average performance across all the different combinations of domains used for training, where there might be data imbalance between two specific domains, the OOD performance drop was still captured regardless of the combination of domains used in the experiments. This gave us an intuition that the problem of spatial domain shift exists under such a categorisation of domains.

Furthermore, we believe that incorporating additional classes to DSGR would lead to a more diversified dataset. For example, this enhancement can be achieved by including samples from the classes dropped due to the scarcity of data from these classes during the preprocessing stage (Section 3.1), such as hospitals, impoverished settlements, multi-unit residential, etc. Similarly, this goal can be achieved by adding new classes used in various real-world applications like farms, fisheries, wooded land, etc. Therefore, in our future work, we aim to enhance DSGR by including these classes, which are vital in evaluating spatial domain shifts in land-use classification tasks.

In the same vein, another direction towards enhancing DSGR is to address the imbalance in the granularity of the class categorisation of fMoW (Christie et al., 2018). For example, fMoW combines different recreational facilities, such as a tennis court and a soccer fields, under a single class, *Recreational-Facilities*. However, we believe these classes should not be merged due to the differences in their features. Whereas, the airport category in fMoW, for instance, is split into multiple classes, *Airport*, *Airport-Hanger* and *Airport-Runway*, resulting in a significant sample distribution imbalance between geographic regions.

Hence, it is important to note that the focus of this work evolves around the issue of domain shift and does not specifically approach to address data and class imbalances. However, the problems of data and class imbalances are orthogonal directions and we will investigate additional methods to address them in our future work.

6.3 Performance of DG Methods on DSGR

While in the popular DomainBed (Gulrajani & Lopez-Paz, 2021) benchmark study the authors have shown that ERM had a superior performance when evaluating on the standard DG datasets, we have observed, through benchmarking the SOTA DG algorithms on DSGR, that the standard ERM had negligible difference, yet weak, performance in comparison to other SOTA DG algorithms. Furthermore, our experiments on DSGR revealed that the SOTA DG algorithms are in their infancy in regards to their generalisability when it comes to real-world scenarios, therefore, DG remains an open area of research.

6.4 Limitations and Future Work

One of the main limitations of this work is the inherited biases of the original fMoW discussed in Section 6.2. While these subtle biases might be beneficial to mimic a real-life dataset, we aim in our future work to acquire more data samples from underrepresented regions, such as Oceania and Africa, which will also aid in creating fine-grained class categorisations in order to reduce the effects of these factors.

Another interesting direction for our upcoming work is defining the DG problem for remote sensing applications based on different categorisation of domains, such that, in addition to defining domains as geographic regions, we define them in terms of climate zones, poverty levels, world bank regions, seasonal changes, etc.

Furthermore, we aim to expand our analysis on the effect of domain shift, through proposing and investigating different DG datasets, for a diversity of remote sensing tasks, such as object detection, semantic segmentation, super-resolution and regression.

7 Conclusion

In this paper, we provided insights on the effects of spatial domain shift in land-use classification using our proposed dataset, DSGR, which aims to measure the domain shift of DL models on data samples gathered from different geographic regions. Our experiments showed that DL models suffer significant performance drops when evaluated on OOD data from an unseen geographic region. However, using multiple source domains instead of a single source domain during training improves the generalisability of DL models to OOD data. Furthermore, due to the effectiveness of foundation models in improving the performance of DL models, we explored their use as the backbone of the SOTA DG algorithms. Our experiments revealed that integrating CLIP with ERM, a decades-old algorithm, outperforms recent SOTA DG algorithms, including those directly built on CLIP.

Finally, we deduced that having an explicit OOD validation set during training does not further enhance the performance of the DL models on OOD test data. We also noted some limitations of our work due to the inherited biases of the original fMoW dataset and highlighted future research directions to mitigate them.

Funding Open Access funding provided by the Qatar National Library. No funds, grants, or other support was received.

Data Availability The data supporting the findings of this study will be made available upon the acceptance of this manuscript. During the review process, the editors and reviewers of the IJCV special issue may request access to the data via a private link.

Declarations

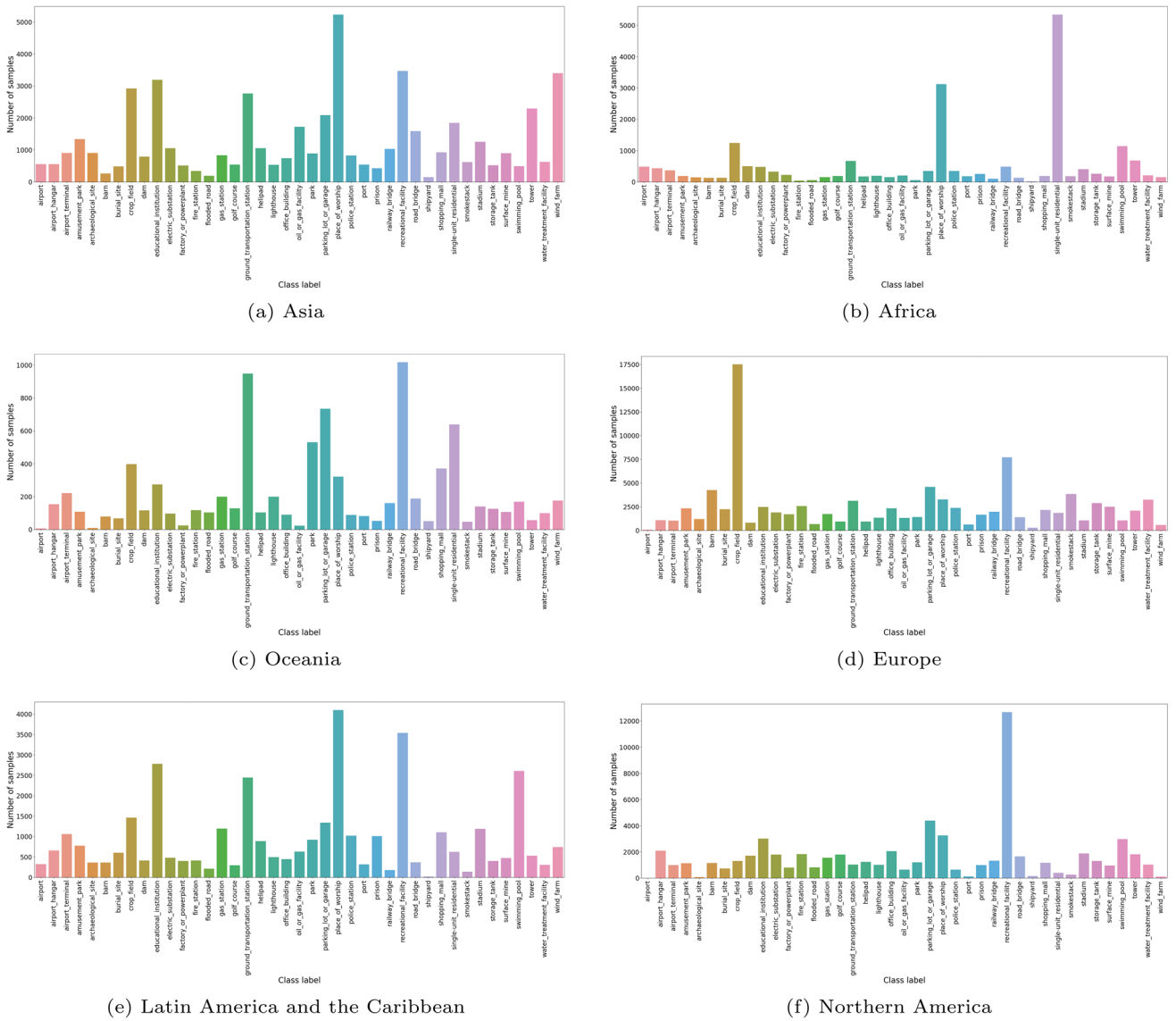
Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Materials availability Not applicable.

Code availability The code supporting the experiments conducted in this study will be made available following the acceptance of this manuscript.



(a) Asia (b) Africa

(c) Oceania (d) Europe

(e) Latin America and the Caribbean (f) Northern America

Fig. 14 The representation in this diagram provides an overview of the class distribution in the geographic region $D_s, s \in \{AS, .., NA\}$, for the training split. There are **41** unique classes in DSGR for the geographic region Asia which are indicated in the x-axis in this diagram. Whereas, the number of samples per class is indicated in y-axis of the diagram

Appendix A Class Distributions of the Geographic Regions in DSGR

In this section, similar to Figure 5, we show another visualisation of cross-class sample distribution of the training set for each geographic region, namely Asia (AS), Africa (AF), Oceania (OC), Europe (EU), Latin America and the Caribbean (LAC) and Northern America (NA), in Figure 14.

Table 26 Single-Source Domain Generalisation accuracy results on DSGR using IRM where the diagonal cells indicate the In-Distribution (ID) test performance of each model

Source	Target AS	AF	OC	EU	LAC	NA
AS	**65.0 ± 0.4**	42.1 ± 1.5	39.8 ± 2.7	40.5 ± 0.8	38.5 ± 0.7	41.0 ± 1.4
AF	32.8 ± 0.9	**70.2 ± 1.0**	36.3 ± 1.4	29.8 ± 0.7	31.2 ± 0.3	29.1 ± 0.5
OC	22.8 ± 0.7	18.7 ± 0.7	**57.2 ± 1.7**	30.7 ± 1.6	26.1 ± 1.7	32.9 ± 0.6
EU	39.1 ± 1.0	27.1 ± 1.9	47.0 ± 5.3	**68.4 ± 0.5**	33.3 ± 0.9	50.1 ± 0.7
LAC	37.7 ± 1.3	35.0 ± 2.2	46.8 ± 2.4	34.7 ± 0.3	**57.2 ± 1.1**	37.7 ± 0.6
NA	32.1 ± 1.0	22.1 ± 1.1	47.1 ± 1.7	45.8 ± 0.7	30.0 ± 1.6	**67.6 ± 0.5**

Table 27 Single-Source Domain Generalisation accuracy results on DSGR using Deep CORAL where the diagonal cells indicate the In-Distribution (ID) test performance of each model.

Source	Target AS	AF	OC	EU	LAC	NA
AS	**65.6 ± 0.5**	43.6 ± 1.8	41.8 ± 2.7	41.3 ± 0.8	39.1 ± 1.8	41.9 ± 0.5
AF	33.7 ± 0.3	**69.5 ± 1.9**	38.7 ± 3.9	30.9 ± 0.9	30.3 ± 1.4	31.2 ± 0.4
OC	25.4 ± 0.8	21.5 ± 4.3	**58.0 ± 2.5**	32.6 ± 0.6	27.0 ± 1.3	35.2 ± 1.1
EU	41.6 ± 1.0	30.2 ± 1.9	50.3 ± 2.4	**70.0 ± 0.2**	37.0 ± 0.7	51.6 ± 0.7
LAC	38.3 ± 2.2	34.1 ± 3.0	46.6 ± 1.3	36.2 ± 1.2	**58.5 ± 1.0**	38.1 ± 1.5
NA	32.6 ± 1.0	22.2 ± 0.9	48.9 ± 1.3	46.8 ± 1.1	31.0 ± 0.7	**68.2 ± 0.1**

Table 28 Single-Source Domain Generalisation accuracy results on DSGR using group DRO where the diagonal cells indicate the In-Distribution (ID) test performance of each model.

Source	Target AS	AF	OC	EU	LAC	NA
AS	**64.3 ± 0.1**	43.1 ± 0.6	43.1 ± 1.3	41.4 ± 0.5	37.8 ± 1.2	41.7 ± 0.6
AF	32.8 ± 0.7	**69.6 ± 1.1**	34.5 ± 2.4	29.2 ± 0.9	29.9 ± 0.8	28.9 ± 1.8
OC	23.2 ± 1.9	17.5 ± 0.5	**56.1 ± 4.1**	29.3 ± 1.6	23.9 ± 1.0	33.3 ± 1.7
EU	40.2 ± 0.4	26.5 ± 1.1	50.1 ± 3.0	**68.2 ± 0.3**	34.5 ± 1.4	50.1 ± 0.6
LAC	37.5 ± 0.2	35.2 ± 1.2	47.1 ± 0.9	35.4 ± 1.0	**57.0 ± 1.7**	37.2 ± 1.1
NA	32.1 ± 0.9	22.7 ± 3.7	45.4 ± 0.7	44.8 ± 0.5	32.3 ± 2.2	**66.4 ± 0.3**

Table 29 Single-Source Domain Generalisation accuracy results on DSGR using CLIPood where the diagonal cells indicate the In-Distribution (ID) test performance of each model.

Source	Target AS	AF	OC	EU	LAC	NA
AS	**61.9 ± 0.2**	48.3 ± 1.4	51.9 ± 0.7	47.7 ± 0.2	43.1 ± 0.3	49.3 ± 0.3
AF	43.1 ± 0.4	**70.4 ± 0.4**	47.9 ± 0.3	42.0 ± 0.4	38.8 ± 0.4	44.1 ± 0.2
OC	37.2 ± 0.4	38.5 ± 0.2	**60.9 ± 0.5**	42.3 ± 0.3	36.3 ± 0.4	44.6 ± 0.4
EU	46.4 ± 0.5	46.4 ± 1.7	59.9 ± 0.0	**63.2 ± 0.3**	41.8 ± 0.2	52.1 ± 0.6
LAC	47.2 ± 0.3	38.7 ± 0.8	55.1 ± 0.9	43.7 ± 0.1	**52.8 ± 0.8**	46.8 ± 0.3
NA	43.9 ± 0.2	37.9 ± 0.9	58.8 ± 0.8	51.7 ± 0.3	43.1 ± 0.6	**64.9 ± 0.3**

Appendix B Additional Experimental Results and Analyses

B.1 Single-Source Domain Generalisation

In this section, following the observations found in Section 5.3, we present below the experimental results for IRM, Deep CORAL, group DRO and CLIPood in Tables 26, 27, 28 and 29 respectively.

The results illustrated in Tables 26, 27, 28 and 29 are aligned with our findings in Section 5.3, where a notable generalisation gap between ID and OOD performance of the model trained using a single source domain.

B. 2 Multi-source Domain Generalisation

Following the observations found in Section 5.4, we present below the experimental results for IRM, Deep CORAL, group DRO and CLIPood in Tables 30, 31, 32 and 33 respectively.

Furthermore, for each algorithm, we provide in Tables 34, 35, 36 and 37 a comparison between the ID and OOD test results between Single-Source and Multi-Source Domain Generalisation on DSGR.

Similar to the observations found in Section 5.4, it can be deduced by comparing the results of the ID test sets in single-source DG with the results of ID test sets in multi-source DG, that in the majority of the cases for all the four algorithms, there is an improvement in the overall performance of the model. Likewise, when testing on the OOD test set in single-source DG versus multi-source DG where the average performance drop of the models in multi-source DG on the OOD testing set is noticeably lower than that of the corresponding OOD testing set in single-source DG. However, in this case, the improvement in the performance is reflected in all the different cases.

B. 3 OOD-Aware Multi-Source Domain Generalisation

In this section, we present the details and breakdown results of the OOD-aware multi-source DG experiment conducted in Section 5.6. Each of the five algorithms consist of a set of experiments. Each of these experiments is is defined as indicated in Table 15 in Section 5.6.

We trained a total of 30 models per algorithm for each of the three different seeds. This resulted in 450 trained models. We ran the evaluation experiments 2700 times to ensure that we cover all the different combinations. The breakdown of results of these experiments for ERM, IRM, Deep CORAL, group DRO and CLIPood are illustrated in Tables 38, 39, 40, 41 and 42 respectively. Moreover, for IRM, Deep CORAL, group DRO and CLIPood we provide in Tables 43, 44, 45 and 46 a comparison between the ID and OOD test results between OOD-aware multi-source DG and multi-source DG.

Appendix C Class Distribution of DSGR-OS

We provide, in Table 47, the class distribution of DSGR-OS.

Appendix D Urbanisation Distribution Across Splits in DSGR

Table 48 presents the distribution of the training and testing splits with respect to urban and rural areas.

Table 30 Multi-Source Domain Generalisation accuracy results on DSGR using IRM. The boldface cells indicate the performance of each model on the OOD test set

Source	Target AS	AF	OC	EU	LAC	NA
$D_{\{C-\text{AS}\}}$	**50.7 ± 0.3**	72.9 ± 0.7	67.4 ± 1.3	67.7 ± 0.2	60.9 ± 1.4	67.7 ± 0.4
$D_{\{C-\text{AF}\}}$	66.6 ± 0.7	**47.3 ± 1.7**	64.8 ± 1.9	68.2 ± 0.4	61.5 ± 0.5	68.0 ± 0.8
$D_{\{C-\text{OC}\}}$	67.6 ± 1.3	73.3 ± 1.3	**59.4 ± 1.9**	67.9 ± 0.4	61.5 ± 0.9	67.8 ± 1.1
$D_{\{C-\text{EU}\}}$	65.5 ± 0.4	72.8 ± 1.0	67.4 ± 1.5	**51.4 ± 0.6**	60.4 ± 0.9	66.3 ± 0.4
$D_{\{C-\text{LAC}\}}$	66.6 ± 0.2	74.2 ± 1.5	65.0 ± 4.0	67.8 ± 0.6	**48.6 ± 1.0**	68.2 ± 0.6
$D_{\{C-\text{NA}\}}$	66.5 ± 1.0	72.9 ± 1.8	65.4 ± 2.2	67.8 ± 0.5	61.1 ± 0.6	**56.3 ± 0.5**

Table 31 Multi-Source Domain Generalisation accuracy results on DSGR using Deep CORAL. The boldface cells indicate the performance of each model on the OOD test set

Source	Target AS	AF	OC	EU	LAC	NA
$D_{\{C-\text{AS}\}}$	**50.4 ± 0.8**	75.0 ± 1.8	68.5 ± 1.0	67.3 ± 1.3	62.9 ± 0.4	69.1 ± 0.2
$D_{\{C-\text{AF}\}}$	68.2 ± 0.5	**49.4 ± 3.3**	68.3 ± 1.2	68.1 ± 0.5	63.5 ± 0.6	69.4 ± 0.6
$D_{\{C-\text{OC}\}}$	68.2 ± 0.2	75.5 ± 0.8	**58.0 ± 2.4**	68.5 ± 0.9	64.4 ± 0.2	70.1 ± 0.4
$D_{\{C-\text{EU}\}}$	66.4 ± 0.8	75.4 ± 0.3	69.1 ± 2.6	**52.1 ± 1.3**	61.8 ± 1.0	67.9 ± 1.0
$D_{\{C-\text{LAC}\}}$	67.6 ± 0.5	74.7 ± 0.5	68.1 ± 2.0	69.1 ± 0.2	**47.4 ± 2.5**	69.3 ± 0.5
$D_{\{C-\text{NA}\}}$	67.2 ± 0.5	75.4 ± 0.3	68.8 ± 1.6	68.1 ± 0.6	62.1 ± 0.6	**56.0 ± 0.4**

Table 32 Multi-Source Domain Generalisation accuracy results on DSGR using group DRO. The boldface cells indicate the performance of each model on the OOD test set

Source	Target AS	AF	OC	EU	LAC	NA
$D_{\{C-\text{AS}\}}$	**50.0 ± 1.0**	69.3 ± 0.9	66.1 ± 1.6	68.7 ± 0.5	58.5 ± 1.3	67.7 ± 0.8
$D_{\{C-\text{AF}\}}$	66.9 ± 0.5	**49.4 ± 1.3**	63.5 ± 2.2	69.1 ± 0.1	59.9 ± 1.0	67.6 ± 0.5
$D_{\{C-\text{OC}\}}$	66.9 ± 0.4	69.7 ± 1.4	**59.5 ± 3.2**	69.5 ± 0.4	60.4 ± 0.2	68.0 ± 0.9
$D_{\{C-\text{EU}\}}$	65.3 ± 1.4	69.9 ± 0.9	64.7 ± 0.9	**51.7 ± 0.6**	60.4 ± 0.3	66.8 ± 0.3
$D_{\{C-\text{LAC}\}}$	67.3 ± 0.6	69.7 ± 0.6	64.7 ± 3.3	69.1 ± 0.6	**47.7 ± 1.5**	67.3 ± 0.7
$D_{\{C-\text{NA}\}}$	66.0 ± 0.8	70.1 ± 0.8	63.6 ± 0.9	68.1 ± 0.1	59.1 ± 0.6	**55.3 ± 0.7**

Table 33 Multi-Source Domain Generalisation accuracy results on DSGR using CLIPood. The boldface cells indicate the performance of each model on the OOD test set

Source	Target AS	AF	OC	EU	LAC	NA
$D_{\{C-\text{AS}\}}$	**51.1 ± 0.2**	63.3 ± 0.5	64.1 ± 0.8	58.3 ± 0.3	49.7 ± 0.1	57.7 ± 0.4
$D_{\{C-\text{AF}\}}$	56.0 ± 0.7	**45.7 ± 2.4**	63.4 ± 1.2	57.5 ± 0.5	49.6 ± 0.8	56.8 ± 0.2
$D_{\{C-\text{OC}\}}$	56.3 ± 0.4	62.4 ± 0.4	**61.2 ± 0.8**	57.4 ± 0.1	48.7 ± 0.8	56.5 ± 0.9
$D_{\{C-\text{EU}\}}$	57.5 ± 0.3	64.5 ± 0.2	64.5 ± 1.0	**51.0 ± 0.5**	50.8 ± 0.2	58.5 ± 0.3
$D_{\{C-\text{LAC}\}}$	56.3 ± 0.2	64.2 ± 0.6	64.0 ± 1.2	58.4 ± 0.3	**45.0 ± 0.5**	57.3 ± 0.6
$D_{\{C-\text{NA}\}}$	57.1 ± 0.2	63.2 ± 1.8	64.9 ± 0.5	58.2 ± 0.2	49.5 ± 0.7	**52.2 ± 0.4**

Table 34 Comparison results of IRM: Multi-Source vs. Single-Source Domain Generalisation on DSGR

	Target					
	AS	AF	OC	EU	LAC	NA
ID % Increase	2.4	4.4	15.4	−0.7	6.9	0.1
OOD % Increase	54	63	37	42	53	48

Table 35 Comparison results of Deep CORAL: Multi-Source vs. Single-Source Domain Generalisation on DSGR

	Target					
	AS	AF	OC	EU	LAC	NA
ID % Increase	2.9	8.2	18.3	−2.6	7.5	1.4
OOD % Increase	47	63	28	39	44	41

Table 36 Comparison results of group DRO: Multi-Source vs. Single-Source Domain Generalisation on DSGR

	Target					
	AS	AF	OC	EU	LAC	NA
ID % Increase	3.3	0.2	15	1.0	4.6	1.6
OOD % Increase	51	70	35	43	51	45

Table 37 Comparison results of CLIPood: Multi-Source vs. Single-Source Domain Generalisation on DSGR

	Target					
	AS	AF	OC	EU	LAC	NA
ID % Increase	−8.4	−9.8	5.4	−8.3	−5.9	−11.7
OOD % Increase	17.3	8.9	11.8	12.2	10.7	10.2

Table 38 OOD-aware Multi-Source Domain Generalisation accuracy results of ERM

Sources	Target AS	AF	OC	EU	LAC	NA
$D_{AS} \setminus AF$	**48.1 ± 1.6**	41.1 ± 1.6	62.9 ± 0.1	67.7 ± 0.7	60.5 ± 1.0	67.6 ± 0.9
$D_{AS} \setminus OC$	**49.9 ± 0.3**	73.2 ± 0.7	57.9 ± 1.0	67.0 ± 1.0	60.1 ± 1.3	67.1 ± 0.5
$D_{AS} \setminus EU$	**46.7 ± 1.4**	72.9 ± 0.8	65.6 ± 1.7	48.7 ± 0.6	59.6 ± 1.6	65.2 ± 1.5
$D_{AS} \setminus LAC$	**49.3 ± 1.1**	72.0 ± 0.2	65.8 ± 1.5	67.3 ± 1.1	46.1 ± 1.7	67.1 ± 1.4
$D_{AS} \setminus NA$	**50.2 ± 0.7**	71.8 ± 1.9	65.0 ± 0.7	67.4 ± 0.7	59.4 ± 0.7	54.8 ± 0.8
$D_{AF} \setminus AS$	48.0 ± 0.4	**37.6 ± 2.6**	65.7 ± 0.7	67.3 ± 1.0	60.3 ± 1.3	67.7 ± 1.0
$D_{AF} \setminus OC$	65.8 ± 2.0	**48.2 ± 2.6**	58.5 ± 1.0	67.5 ± 0.3	59.9 ± 1.0	68.1 ± 0.2
$D_{AF} \setminus EU$	65.0 ± 2.1	**45.4 ± 0.8**	66.0 ± 1.6	52.0 ± 1.1	60.2 ± 0.5	66.1 ± 0.5
$D_{AF} \setminus LAC$	66.9 ± 0.2	**48.4 ± 1.3**	64.3 ± 0.3	68.3 ± 0.9	48.3 ± 0.6	68.0 ± 0.8
$D_{AF} \setminus NA$	66.0 ± 0.9	**44.8 ± 2.1**	65.1 ± 1.5	67.4 ± 0.8	60.4 ± 0.9	56.2 ± 1.2
$D_{OC} \setminus AS$	51.1 ± 0.4	71.8 ± 2.1	**57.5 ± 1.2**	67.2 ± 0.6	60.1 ± 0.9	67.9 ± 0.3
$D_{OC} \setminus AF$	66.5 ± 1.3	51.6 ± 1.1	**60.0 ± 1.9**	68.0 ± 0.8	60.6 ± 1.4	68.3 ± 1.3
$D_{OC} \setminus EU$	65.2 ± 1.1	73.2 ± 1.4	**56.7 ± 2.2**	52.5 ± 0.8	59.6 ± 1.4	66.8 ± 0.4
$D_{OC} \setminus LAC$	66.8 ± 0.3	73.1 ± 0.9	**53.7 ± 0.6**	68.3 ± 0.2	48.4 ± 0.8	68.1 ± 0.6
$D_{OC} \setminus NA$	65.9 ± 0.3	73.1 ± 0.5	**56.2 ± 1.6**	67.9 ± 0.8	60.6 ± 0.6	56.2 ± 0.7
$D_{EU} \setminus AS$	45.4 ± 1.7	69.9 ± 3.9	65.0 ± 2.6	**47.8 ± 0.2**	59.1 ± 0.9	65.6 ± 0.9
$D_{EU} \setminus AF$	65.6 ± 0.7	48.5 ± 1.0	67.0 ± 0.5	**51.4 ± 0.6**	60.1 ± 1.0	66.9 ± 1.1
$D_{EU} \setminus OC$	63.3 ± 2.8	71.3 ± 4.3	54.4 ± 0.5	**50.1 ± 0.9**	58.8 ± 0.6	65.7 ± 0.5
$D_{EU} \setminus LAC$	64.4 ± 1.1	72.2 ± 1.7	65.7 ± 1.3	**51.0 ± 1.5**	47.1 ± 1.7	66.4 ± 1.6
$D_{EU} \setminus NA$	64.5 ± 0.3	71.5 ± 0.1	64.6 ± 0.8	**45.7 ± 1.0**	59.2 ± 1.2	50.1 ± 1.1
$D_{LAC} \setminus AS$	48.3 ± 0.7	69.2 ± 4.6	62.7 ± 2.3	66.8 ± 0.4	**45.5 ± 0.7**	67.1 ± 0.5
$D_{LAC} \setminus AF$	65.5 ± 0.3	51.1 ± 0.9	64.2 ± 0.9	68.1 ± 0.3	**47.4 ± 1.0**	67.5 ± 1.1
$D_{LAC} \setminus OC$	64.7 ± 2.0	71.4 ± 3.9	53.5 ± 0.4	68.3 ± 1.1	**47.0 ± 1.7**	68.1 ± 1.0
$D_{LAC} \setminus EU$	64.6 ± 1.0	73.5 ± 0.5	64.7 ± 1.4	51.7 ± 0.8	**47.2 ± 1.3**	66.7 ± 1.1
$D_{LAC} \setminus NA$	65.7 ± 1.0	72.1 ± 0.2	64.0 ± 2.2	67.4 ± 0.7	**46.7 ± 0.4**	55.5 ± 0.1
$D_{NA} \setminus AS$	50.0 ± 0.3	70.7 ± 1.6	63.9 ± 1.3	66.6 ± 0.7	58.6 ± 1.5	**53.0 ± 1.1**
$D_{NA} \setminus AF$	65.9 ± 0.8	48.7 ± 1.6	64.8 ± 1.4	67.7 ± 0.9	59.7 ± 0.7	**54.8 ± 1.0**
$D_{NA} \setminus OC$	64.4 ± 1.2	71.6 ± 2.4	56.1 ± 0.5	66.8 ± 1.9	59.6 ± 0.3	**55.7 ± 0.8**
$D_{NA} \setminus EU$	63.4 ± 1.2	71.9 ± 0.9	65.4 ± 1.0	46.3 ± 0.9	59.6 ± 0.5	**49.3 ± 1.4**
$D_{NA} \setminus LAC$	65.8 ± 1.5	72.0 ± 1.2	63.8 ± 0.3	67.2 ± 1.0	48.0 ± 2.0	**54.9 ± 0.8**

Table 39 OOD-aware Multi-Source Domain Generalisation accuracy results of IRM

Sources	Target AS	AF	OC	EU	LAC	NA
$D_{AS} \setminus AF$	**48.3 ± 0.1**	39.7 ± 0.5	66.8 ± 2.5	68.2 ± 0.1	60.7 ± 0.2	67.9 ± 0.9
$D_{AS} \setminus OC$	**50.4 ± 1.3**	72.9 ± 0.7	58.0 ± 3.7	67.2 ± 0.4	59.4 ± 1.4	67.5 ± 0.5
$D_{AS} \setminus EU$	**46.5 ± 0.8**	72.6 ± 1.4	65.5 ± 0.8	49.1 ± 0.3	59.8 ± 1.1	66.2 ± 0.6
$D_{AS} \setminus LAC$	**48.0 ± 0.6**	72.3 ± 0.7	66.5 ± 1.1	68.0 ± 0.1	45.4 ± 0.8	67.5 ± 0.9
$D_{AS} \setminus NA$	**50.8 ± 1.1**	72.1 ± 0.8	66.3 ± 2.1	67.5 ± 0.3	59.6 ± 0.7	54.7 ± 0.5
$D_{AF} \setminus AS$	49.3 ± 0.4	**38.8 ± 0.8**	64.7 ± 1.0	67.7 ± 0.9	61.1 ± 0.7	62.3 ± 9.4
$D_{AF} \setminus OC$	66.0 ± 0.8	**47.9 ± 1.1**	58.8 ± 2.2	66.4 ± 2.2	61.7 ± 1.0	67.4 ± 1.3
$D_{AF} \setminus EU$	66.0 ± 0.8	**47.0 ± 1.6**	66.2 ± 1.9	52.2 ± 1.4	61.5 ± 0.2	66.1 ± 0.6

Table 39 continued

Sources	Target AS	AF	OC	EU	LAC	NA
$D_{AF} \setminus LAC$	67.0 ± 0.3	$\mathbf{48.2 \pm 1.6}$	63.6 ± 3.2	68.0 ± 0.6	48.4 ± 1.8	67.5 ± 1.3
$D_{AF} \setminus NA$	66.5 ± 0.2	$\mathbf{47.4 \pm 0.7}$	66.2 ± 0.8	68.0 ± 0.6	61.2 ± 0.7	59.9 ± 6.7
$D_{OC} \setminus AS$	52.0 ± 0.7	72.1 ± 1.8	$\mathbf{58.5 \pm 1.7}$	67.8 ± 0.8	61.0 ± 0.8	68.2 ± 1.0
$D_{OC} \setminus AF$	65.7 ± 1.7	50.8 ± 1.6	$\mathbf{57.4 \pm 0.7}$	66.8 ± 2.9	60.7 ± 1.8	66.8 ± 1.7
$D_{OC} \setminus EU$	65.8 ± 1.0	74.8 ± 0.2	$\mathbf{55.7 \pm 1.6}$	51.7 ± 0.8	60.9 ± 0.0	67.3 ± 0.5
$D_{OC} \setminus LAC$	66.5 ± 0.3	72.8 ± 1.0	$\mathbf{53.9 \pm 2.3}$	68.3 ± 0.3	48.3 ± 1.1	68.0 ± 0.3
$D_{OC} \setminus NA$	67.2 ± 0.5	73.3 ± 1.5	$\mathbf{58.0 \pm 1.7}$	68.0 ± 0.4	60.8 ± 0.6	55.8 ± 0.2
$D_{EU} \setminus AS$	47.7 ± 0.6	71.8 ± 1.1	65.2 ± 1.2	$\mathbf{48.5 \pm 0.7}$	60.0 ± 1.0	66.5 ± 0.3
$D_{EU} \setminus AF$	65.8 ± 0.4	47.7 ± 0.9	67.0 ± 0.7	$\mathbf{50.7 \pm 0.6}$	60.4 ± 0.6	66.6 ± 0.1
$D_{EU} \setminus OC$	65.5 ± 0.4	73.6 ± 0.9	57.2 ± 1.5	$\mathbf{51.3 \pm 0.3}$	60.0 ± 0.5	66.4 ± 0.9
$D_{EU} \setminus LAC$	66.1 ± 0.8	75.2 ± 0.5	65.7 ± 1.2	$\mathbf{51.2 \pm 0.7}$	47.7 ± 0.4	66.2 ± 0.8
$D_{EU} \setminus NA$	62.9 ± 2.4	72.2 ± 0.7	64.8 ± 0.3	$\mathbf{46.3 \pm 0.8}$	59.0 ± 0.8	50.0 ± 0.1
$D_{LAC} \setminus AS$	49.1 ± 0.5	72.7 ± 0.8	65.6 ± 1.2	67.9 ± 0.6	$\mathbf{44.7 \pm 0.5}$	67.9 ± 0.3
$D_{LAC} \setminus AF$	66.6 ± 1.4	49.7 ± 1.3	63.6 ± 1.1	67.7 ± 0.3	$\mathbf{46.2 \pm 0.9}$	67.8 ± 0.8
$D_{LAC} \setminus OC$	66.5 ± 0.6	73.6 ± 0.5	53.6 ± 0.6	68.5 ± 0.7	$\mathbf{47.8 \pm 0.9}$	68.3 ± 0.3
$D_{LAC} \setminus EU$	65.2 ± 0.9	73.1 ± 1.1	65.4 ± 0.9	52.4 ± 0.5	$\mathbf{47.0 \pm 0.2}$	67.1 ± 0.6
$D_{LAC} \setminus NA$	66.6 ± 1.1	73.5 ± 0.9	63.7 ± 1.9	67.8 ± 0.2	$\mathbf{46.5 \pm 1.5}$	55.3 ± 0.6
$D_{NA} \setminus AS$	50.8 ± 1.5	72.2 ± 0.5	66.0 ± 0.9	67.0 ± 0.2	59.6 ± 1.0	$\mathbf{54.5 \pm 0.6}$
$D_{NA} \setminus AF$	66.1 ± 0.5	48.9 ± 0.7	64.7 ± 1.0	67.7 ± 0.7	61.1 ± 1.3	$\mathbf{55.0 \pm 0.2}$
$D_{NA} \setminus OC$	66.8 ± 1.2	72.0 ± 0.6	56.2 ± 0.8	67.5 ± 0.3	59.8 ± 1.0	$\mathbf{54.5 \pm 0.3}$
$D_{NA} \setminus EU$	65.1 ± 0.4	70.8 ± 0.7	64.8 ± 1.8	46.9 ± 0.7	59.7 ± 0.7	$\mathbf{48.6 \pm 1.0}$
$D_{NA} \setminus LAC$	66.2 ± 0.3	72.6 ± 1.3	62.9 ± 1.4	67.5 ± 0.4	46.9 ± 0.5	$\mathbf{55.0 \pm 0.1}$

Table 40 OOD-aware Multi-Source Domain Generalisation accuracy results of Deep CORAL

Sources	Target AS	AF	OC	EU	LAC	NA
$D_{AS} \setminus AF$	$\mathbf{47.6 \pm 1.4}$	39.5 ± 2.2	67.6 ± 0.5	68.0 ± 1.8	61.8 ± 0.8	69.6 ± 0.2
$D_{AS} \setminus OC$	$\mathbf{50.8 \pm 1.2}$	75.1 ± 1.9	59.3 ± 2.0	68.0 ± 1.3	62.2 ± 0.8	69.1 ± 0.5
$D_{AS} \setminus EU$	$\mathbf{48.0 \pm 0.7}$	74.3 ± 1.1	67.2 ± 1.0	49.9 ± 0.7	61.6 ± 0.6	67.2 ± 0.2
$D_{AS} \setminus LAC$	$\mathbf{49.6 \pm 0.9}$	74.2 ± 1.0	68.1 ± 2.4	67.7 ± 1.9	45.6 ± 0.9	67.7 ± 0.9
$D_{AS} \setminus NA$	$\mathbf{47.9 \pm 1.5}$	73.6 ± 1.1	68.2 ± 1.2	67.1 ± 0.7	61.5 ± 1.3	55.4 ± 0.8
$D_{AF} \setminus AS$	48.9 ± 0.3	$\mathbf{40.0 \pm 1.8}$	66.8 ± 2.4	68.3 ± 0.2	63.1 ± 1.2	69.7 ± 0.5
$D_{AF} \setminus OC$	68.0 ± 0.9	$\mathbf{48.6 \pm 3.9}$	60.0 ± 0.6	68.1 ± 1.1	63.3 ± 1.4	69.2 ± 1.1
$D_{AF} \setminus EU$	66.7 ± 0.9	$\mathbf{48.8 \pm 1.7}$	68.8 ± 0.7	53.2 ± 0.7	62.3 ± 1.6	68.3 ± 0.2
$D_{AF} \setminus LAC$	67.7 ± 0.6	$\mathbf{46.3 \pm 1.3}$	67.8 ± 1.2	68.1 ± 1.7	48.5 ± 1.7	69.0 ± 0.7
$D_{AF} \setminus NA$	67.9 ± 1.6	$\mathbf{47.3 \pm 1.2}$	65.3 ± 0.5	68.7 ± 0.6	61.9 ± 0.4	56.4 ± 0.6
$D_{OC} \setminus AS$	50.4 ± 1.3	74.9 ± 2.9	$\mathbf{59.3 \pm 2.0}$	67.7 ± 1.7	62.9 ± 0.1	68.6 ± 1.2
$D_{OC} \setminus AF$	68.4 ± 0.2	52.0 ± 1.5	$\mathbf{60.0 \pm 3.3}$	68.8 ± 0.8	62.7 ± 0.2	69.3 ± 0.8
$D_{OC} \setminus EU$	67.0 ± 0.8	76.2 ± 0.4	$\mathbf{57.8 \pm 1.5}$	53.1 ± 0.9	62.1 ± 0.1	68.0 ± 0.4
$D_{OC} \setminus LAC$	66.9 ± 0.3	74.1 ± 1.5	$\mathbf{57.5 \pm 0.4}$	67.7 ± 1.6	49.2 ± 1.6	68.6 ± 0.2
$D_{OC} \setminus NA$	68.3 ± 0.7	75.4 ± 2.1	$\mathbf{54.5 \pm 2.3}$	68.0 ± 1.0	62.4 ± 0.5	56.3 ± 1.2
$D_{EU} \setminus AS$	48.9 ± 0.1	74.2 ± 1.0	68.2 ± 1.9	$\mathbf{49.1 \pm 0.8}$	62.4 ± 0.8	67.4 ± 1.1
$D_{EU} \setminus AF$	66.1 ± 1.0	50.6 ± 1.9	69.0 ± 0.3	$\mathbf{52.4 \pm 0.7}$	61.4 ± 0.7	66.9 ± 0.6
$D_{EU} \setminus OC$	66.5 ± 1.6	74.8 ± 0.7	58.0 ± 1.2	$\mathbf{52.5 \pm 1.3}$	62.4 ± 0.9	67.7 ± 1.2
$D_{EU} \setminus LAC$	65.2 ± 0.4	74.4 ± 1.5	68.6 ± 0.5	$\mathbf{51.5 \pm 0.5}$	48.5 ± 1.1	67.4 ± 0.1

Table 40 continued

Sources	Target AS	AF	OC	EU	LAC	NA
$D_{EU} \setminus NA$	65.9 ± 0.6	74.1 ± 0.9	67.3 ± 2.1	$\mathbf{46.6 \pm 0.2}$	61.2 ± 0.8	50.4 ± 0.4
$D_{LAC} \setminus AS$	49.8 ± 0.6	73.2 ± 0.9	67.8 ± 1.7	67.9 ± 1.5	$\mathbf{44.4 \pm 1.3}$	68.7 ± 0.8
$D_{LAC} \setminus AF$	68.0 ± 1.5	52.4 ± 1.5	66.7 ± 2.3	68.1 ± 1.4	$\mathbf{46.6 \pm 1.9}$	69.1 ± 0.9
$D_{LAC} \setminus OC$	67.3 ± 0.9	75.7 ± 1.2	57.7 ± 1.6	68.9 ± 1.4	$\mathbf{48.7 \pm 1.3}$	69.9 ± 0.3
$D_{LAC} \setminus EU$	65.3 ± 0.4	73.5 ± 0.9	68.1 ± 1.8	51.7 ± 0.1	$\mathbf{47.1 \pm 0.9}$	67.8 ± 0.5
$D_{LAC} \setminus NA$	66.6 ± 1.0	74.6 ± 1.2	67.6 ± 0.8	67.7 ± 0.8	$\mathbf{46.6 \pm 0.3}$	56.7 ± 0.3
$D_{NA} \setminus AS$	50.2 ± 1.1	73.5 ± 0.9	66.3 ± 1.6	67.0 ± 1.3	61.9 ± 1.9	$\mathbf{53.4 \pm 0.7}$
$D_{NA} \setminus AF$	67.5 ± 1.2	50.0 ± 1.0	66.3 ± 1.6	67.9 ± 1.5	62.6 ± 1.3	$\mathbf{56.1 \pm 0.1}$
$D_{NA} \setminus OC$	68.5 ± 0.5	75.7 ± 1.2	56.5 ± 0.6	68.9 ± 0.1	62.3 ± 0.7	$\mathbf{55.4 \pm 0.3}$
$D_{NA} \setminus EU$	65.8 ± 0.5	73.7 ± 0.5	67.3 ± 1.7	46.9 ± 0.8	62.5 ± 0.8	$\mathbf{50.8 \pm 0.7}$
$D_{NA} \setminus LAC$	66.5 ± 1.5	74.0 ± 1.4	66.8 ± 0.7	66.7 ± 2.1	47.4 ± 1.5	$\mathbf{55.1 \pm 1.3}$

Table 41 OOD-aware Multi-Source Domain Generalisation accuracy results of group DRO

Sources	Target AS	AF	OC	EU	LAC	NA
$D_{AS} \setminus AF$	$\mathbf{48.4 \pm 0.4}$	40.1 ± 2.2	63.2 ± 2.8	67.3 ± 2.0	58.1 ± 1.0	67.1 ± 1.1
$D_{AS} \setminus OC$	$\mathbf{49.4 \pm 0.5}$	69.3 ± 1.8	58.0 ± 2.1	68.2 ± 0.3	57.7 ± 0.5	67.2 ± 0.8
$D_{AS} \setminus EU$	$\mathbf{46.2 \pm 1.1}$	70.7 ± 1.7	65.4 ± 0.3	49.6 ± 0.6	59.3 ± 1.2	66.7 ± 1.3
$D_{AS} \setminus LAC$	$\mathbf{47.4 \pm 1.6}$	70.8 ± 1.5	62.5 ± 1.3	67.6 ± 0.8	45.6 ± 1.8	67.5 ± 0.6
$D_{AS} \setminus NA$	$\mathbf{49.2 \pm 0.4}$	69.6 ± 1.0	63.0 ± 1.8	68.1 ± 0.2	59.1 ± 0.2	54.7 ± 1.0
$D_{AF} \setminus AS$	49.8 ± 0.6	$\mathbf{36.7 \pm 1.3}$	62.2 ± 1.4	67.8 ± 1.1	60.1 ± 0.3	67.1 ± 0.7
$D_{AF} \setminus OC$	66.5 ± 0.6	$\mathbf{48.1 \pm 2.5}$	57.1 ± 2.7	69.1 ± 0.3	59.5 ± 0.2	67.4 ± 0.8
$D_{AF} \setminus EU$	65.5 ± 0.8	$\mathbf{48.1 \pm 3.1}$	64.1 ± 2.3	51.7 ± 0.3	60.4 ± 1.9	67.0 ± 0.3
$D_{AF} \setminus LAC$	66.7 ± 1.0	$\mathbf{45.4 \pm 1.6}$	63.2 ± 1.0	68.8 ± 0.5	47.0 ± 0.8	67.2 ± 0.8
$D_{AF} \setminus NA$	66.0 ± 0.6	$\mathbf{46.2 \pm 2.3}$	63.7 ± 1.9	68.7 ± 0.6	58.9 ± 0.9	56.4 ± 1.2
$D_{OC} \setminus AS$	51.6 ± 0.9	70.5 ± 0.8	$\mathbf{56.1 \pm 0.8}$	68.5 ± 0.4	58.5 ± 1.2	67.4 ± 0.9
$D_{OC} \setminus AF$	66.2 ± 0.5	50.7 ± 2.7	$\mathbf{56.9 \pm 1.3}$	68.8 ± 0.1	60.1 ± 0.8	67.5 ± 0.3
$D_{OC} \setminus EU$	65.7 ± 0.9	70.4 ± 0.1	$\mathbf{56.9 \pm 2.0}$	52.3 ± 0.1	59.4 ± 0.7	66.9 ± 0.8
$D_{OC} \setminus LAC$	66.7 ± 0.6	70.9 ± 0.3	$\mathbf{52.6 \pm 0.6}$	69.2 ± 0.3	47.5 ± 0.7	67.9 ± 0.3
$D_{OC} \setminus NA$	65.6 ± 0.6	71.1 ± 0.2	$\mathbf{53.4 \pm 1.9}$	68.1 ± 1.0	58.4 ± 1.2	56.2 ± 0.5
$D_{EU} \setminus AS$	47.8 ± 0.5	69.9 ± 1.2	64.8 ± 1.1	$\mathbf{48.5 \pm 1.4}$	60.4 ± 1.1	67.2 ± 0.3
$D_{EU} \setminus AF$	65.7 ± 0.5	49.8 ± 1.6	65.5 ± 1.4	$\mathbf{50.7 \pm 0.5}$	59.7 ± 1.4	67.0 ± 0.3
$D_{EU} \setminus OC$	65.5 ± 0.7	71.3 ± 1.6	56.0 ± 1.2	$\mathbf{51.1 \pm 0.9}$	60.1 ± 0.3	66.9 ± 0.6
$D_{EU} \setminus LAC$	64.4 ± 0.8	72.4 ± 1.3	63.6 ± 0.9	$\mathbf{50.2 \pm 0.9}$	47.5 ± 1.4	66.4 ± 0.5
$D_{EU} \setminus NA$	65.1 ± 1.2	70.6 ± 0.8	64.5 ± 0.6	$\mathbf{46.0 \pm 1.1}$	59.7 ± 1.1	49.4 ± 0.4
$D_{LAC} \setminus AS$	48.9 ± 0.3	70.2 ± 1.5	61.9 ± 1.9	68.3 ± 0.2	$\mathbf{45.4 \pm 0.9}$	67.9 ± 0.4
$D_{LAC} \setminus AF$	65.9 ± 1.2	48.9 ± 2.2	61.4 ± 0.9	68.6 ± 0.6	$\mathbf{45.8 \pm 0.4}$	67.7 ± 0.7
$D_{LAC} \setminus OC$	66.7 ± 0.5	71.1 ± 2.2	54.5 ± 0.9	68.7 ± 1.4	$\mathbf{47.0 \pm 1.5}$	67.1 ± 1.7
$D_{LAC} \setminus EU$	66.1 ± 0.5	72.0 ± 1.3	65.3 ± 1.4	52.0 ± 0.9	$\mathbf{46.4 \pm 0.8}$	66.9 ± 0.2
$D_{LAC} \setminus NA$	64.8 ± 0.4	70.4 ± 0.8	61.9 ± 0.4	67.7 ± 0.3	$\mathbf{45.4 \pm 1.3}$	55.0 ± 0.8
$D_{NA} \setminus AS$	49.4 ± 0.4	68.5 ± 1.5	62.4 ± 1.0	67.7 ± 0.6	58.8 ± 0.9	$\mathbf{53.5 \pm 1.7}$
$D_{NA} \setminus AF$	65.5 ± 0.1	48.7 ± 0.4	62.8 ± 2.7	68.1 ± 1.1	57.6 ± 0.4	$\mathbf{55.4 \pm 1.1}$
$D_{NA} \setminus OC$	65.3 ± 0.1	69.9 ± 1.1	54.8 ± 1.6	67.7 ± 0.5	58.7 ± 1.0	$\mathbf{55.7 \pm 0.3}$
$D_{NA} \setminus EU$	64.8 ± 0.2	70.0 ± 0.2	63.8 ± 1.6	46.7 ± 0.1	58.6 ± 0.9	$\mathbf{48.5 \pm 0.9}$
$D_{NA} \setminus LAC$	65.8 ± 0.4	69.5 ± 0.4	62.2 ± 1.6	67.5 ± 0.8	45.8 ± 0.6	$\mathbf{55.0 \pm 0.6}$

Table 42 OOD-aware Multi-Source Domain Generalisation accuracy results of CLIPood

Sources	Target AS	AF	OC	EU	LAC	NA
$D_{AS} \setminus AF$	**50.7 ± 0.4**	43.4 ± 0.0	65.0 ± 0.3	59.1 ± 0.4	50.5 ± 0.1	58.5 ± 0.8
$D_{AS} \setminus OC$	**50.8 ± 0.7**	62.9 ± 0.3	60.9 ± 1.0	58.3 ± 0.3	49.4 ± 0.2	57.7 ± 0.5
$D_{AS} \setminus EU$	**51.5 ± 0.1**	66.6 ± 0.3	65.9 ± 0.5	51.2 ± 0.3	52.4 ± 0.5	60.8 ± 0.3
$D_{AS} \setminus LAC$	**48.7 ± 0.4**	64.9 ± 0.9	64.2 ± 0.9	59.0 ± 0.3	43.4 ± 0.6	58.2 ± 0.1
$D_{AS} \setminus NA$	**51.1 ± 0.2**	64.8 ± 0.1	64.0 ± 0.3	59.5 ± 0.3	50.5 ± 0.2	52.4 ± 0.5
$D_{AF} \setminus AS$	50.7 ± 0.4	**43.4 ± 0.0**	65.0 ± 0.3	59.1 ± 0.4	50.5 ± 0.1	58.5 ± 0.8
$D_{AF} \setminus OC$	56.3 ± 0.2	**48.6 ± 2.0**	61.3 ± 1.3	58.0 ± 0.1	49.1 ± 0.3	56.9 ± 0.4
$D_{AF} \setminus EU$	58.3 ± 0.2	**46.7 ± 1.0**	65.1 ± 0.2	51.3 ± 0.1	51.3 ± 0.4	59.1 ± 0.4
$D_{AF} \setminus LAC$	57.3 ± 0.2	**51.0 ± 1.6**	63.8 ± 0.8	58.9 ± 0.2	45.9 ± 0.4	57.5 ± 0.2
$D_{AF} \setminus NA$	57.6 ± 0.3	**48.1 ± 0.8**	64.3 ± 0.9	59.3 ± 0.1	50.6 ± 0.3	52.0 ± 0.4
$D_{OC} \setminus AS$	50.8 ± 0.7	62.9 ± 0.3	**60.9 ± 1.0**	58.3 ± 0.3	49.4 ± 0.2	57.7 ± 0.5
$D_{OC} \setminus AF$	56.2 ± 0.2	48.7 ± 2.0	**61.3 ± 1.3**	58.0 ± 0.1	49.2 ± 0.4	57.0 ± 0.4
$D_{OC} \setminus EU$	58.0 ± 0.2	64.4 ± 1.3	**59.9 ± 0.1**	51.0 ± 0.3	50.4 ± 0.3	58.7 ± 0.3
$D_{OC} \setminus LAC$	56.7 ± 0.2	63.6 ± 0.2	**60.6 ± 0.3**	58.5 ± 0.1	44.2 ± 0.3	57.5 ± 0.0
$D_{OC} \setminus NA$	57.1 ± 0.1	63.1 ± 0.4	**62.2 ± 0.8**	58.4 ± 0.2	49.3 ± 0.1	51.9 ± 0.4
$D_{EU} \setminus AS$	51.5 ± 0.1	66.6 ± 0.3	65.9 ± 0.5	**51.2 ± 0.3**	52.4 ± 0.5	60.8 ± 0.3
$D_{EU} \setminus AF$	58.3 ± 0.3	46.8 ± 1.0	65.1 ± 0.2	**51.3 ± 0.1**	51.3 ± 0.4	59.1 ± 0.4
$D_{EU} \setminus OC$	58.0 ± 0.2	64.4 ± 1.3	59.9 ± 0.1	**51.0 ± 0.3**	50.4 ± 0.3	58.7 ± 0.3
$D_{EU} \setminus LAC$	58.1 ± 1.2	64.6 ± 1.8	65.1 ± 1.1	**51.3 ± 0.6**	45.2 ± 1.4	59.0 ± 0.3
$D_{EU} \setminus NA$	60.1 ± 0.6	66.0 ± 0.3	65.4 ± 0.9	**49.6 ± 0.9**	53.0 ± 0.9	51.3 ± 0.5
$D_{LAC} \setminus AS$	48.7 ± 0.4	64.9 ± 0.9	64.2 ± 0.9	59.0 ± 0.3	**43.4 ± 0.6**	58.2 ± 0.1
$D_{LAC} \setminus AF$	57.3 ± 0.2	50.9 ± 1.5	63.8 ± 0.8	58.8 ± 0.2	**45.9 ± 0.4**	57.5 ± 0.2
$D_{LAC} \setminus OC$	56.7 ± 0.2	63.6 ± 0.2	60.6 ± 0.3	58.5 ± 0.1	**44.2 ± 0.3**	57.5 ± 0.0
$D_{LAC} \setminus EU$	58.1 ± 1.2	64.7 ± 2.0	65.0 ± 1.0	51.3 ± 0.6	**45.1 ± 1.3**	59.0 ± 0.2
$D_{LAC} \setminus NA$	57.8 ± 0.3	63.8 ± 0.3	63.8 ± 1.0	59.5 ± 0.0	**44.6 ± 0.5**	52.1 ± 0.2
$D_{NA} \setminus AS$	51.1 ± 0.2	64.9 ± 0.0	64.0 ± 0.3	59.5 ± 0.3	50.4 ± 0.2	**52.5 ± 0.5**
$D_{NA} \setminus AF$	57.6 ± 0.3	48.1 ± 0.8	64.3 ± 0.9	59.3 ± 0.1	50.6 ± 0.3	**52.0 ± 0.4**
$D_{NA} \setminus OC$	57.1 ± 0.1	63.1 ± 0.4	62.2 ± 0.8	58.4 ± 0.2	49.3 ± 0.1	**51.9 ± 0.4**
$D_{NA} \setminus EU$	60.0 ± 0.6	66.1 ± 0.4	65.4 ± 1.0	49.6 ± 0.9	53.1 ± 0.9	**51.3 ± 0.5**
$D_{NA} \setminus LAC$	57.8 ± 0.3	63.8 ± 0.3	63.8 ± 1.0	59.5 ± 0.0	44.6 ± 0.5	**52.1 ± 0.2**

Table 43 Comparison results of IRM: OOD-aware Multi-Source vs. Multi-Source Domain Generalisation on DSGR

	Target AS	AF	OC	EU	LAC	NA
ID % Increase	−0.8	−0.6	−1.1	−0.3	−1.2	−1.0
OOD % Increase	−3.7	−3.0	−4.6	−3.5	−4.5	−5.0

Table 44 Comparison results of Deep CORAL: OOD-aware Multi-Source vs. Multi-Source Domain Generalisation

	Target AS	AF	OC	EU	LAC	NA
ID % Increase	−0.8	−1.0	−1.6	−0.4	−1.1	−1.0
OOD % Increase	−3.3	−6.4	−0.2	−3.2	−1.5	−3.3

Table 45 Comparison results of group DRO: OOD-aware Multi-Source vs. Multi-Source Domain Generalisation

	Target					
	AS	AF	OC	EU	LAC	NA
ID % Increase	−1.1	1.1	−1.8	−1.0	−0.8	−0.4
OOD % Increase	−3.7	−9.1	−7.2	−4.6	−3.6	−3.0

Table 46 Comparison results of CLIPood: OOD-aware Multi-Source vs. Multi-Source Domain Generalisation

	Target					
	AS	AF	OC	EU	LAC	NA
ID % Increase	1.9	1.5	0.8	1.5	2.0	1.8
OOD % Increase	−1.1	4.2	−0.3	−0.2	−0.8	−0.4

Table 47 Geographic region-wise data partitions used in all experiments

Split	AS	AF	OC	EU	LAC	NA
Training	33, 732	14, 660	5, 486	60, 361	23, 42	42, 183
Validation	4, 922	2, 027	893	9, 154	3, 234	59, 88
Base Testing	4, 934	2, 215	863	8, 847	3, 311	63, 56
New Testing	2, 330	668	533	5, 797	2, 141	3, 273

Table 48 Sample-wise distribution of urban and rural in DSGR

Region	Urbanisation	Training (%)	Testing (%)
AS	Urban	59	60
	Rural	41	40
AF	Urban	63	65
	Rural	37	35
OC	Urban	50	58
	Rural	50	42
EU	Urban	36	37
	Rural	64	63
LAC	Urban	62	65
	Rural	38	35
NA	Urban	45	48
	Rural	55	52

References

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893 .

Arpit, D., Wang, H., Zhou, Y., & Xiong, C. (2022). Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems, 35*, 8265–8277.

Beery, S., Wu, G., Edwards, T., Pavetic, F., Majewski, B., Mukherjee, S., Chan, S., Morgan, J., Rathod, V., & Huang, J. (2022). The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Monitoring Under Domain Shift, In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA. pp. 21262–21275. https://doi.org/10.1109/CVPR52688.2022.02061.

Bui, M. H., Tran, T., Tran, A., & Phung, D. (2021). Exploiting domain-specific features to enhance domain generalization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems* (pp. 21189–21201). Curran Associates: Inc.

Chen, X., Lan, X., Sun, F., & Zheng, N. (2020). A boundary based out-of-distribution classifier for generalized zero-shot learning, In: European conference on computer vision, Springer. pp. 572–588.

Christie, G., Fendley, N., Wilson, J., & Mukherjee, R. (2018). Functional map of the world, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6172–6180.

Ding, Y., Wang, L., Liang, B., Liang, S., Wang, Y., & Chen, F. (2022). Domain generalization by learning and removing domain-specific features. arXiv preprint arXiv:2212.07101 .

Drakonakis, G. I., Tsagkatakis, G., Fotiadou, K., & Tsakalides, P. (2022). Ombrianet-supervised flood mapping via convolutional neural networks using multitemporal sentinel-1 and sentinel-2 data fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15*, 2341–2356. https://doi.org/10.1109/JSTARS.2022.3155559

Eastwood, C., Robey, A., Singh, S., Von Kügelgen, J., Hassani, H., Pappas, G. J., & Schölkopf, B. (2022). Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems, 35*, 17340–17358.

Fang, C., Xu, Y., & Rockmore, D.N. (2013). Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias, In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1657–1664.

Fu, Y., Wang, X., Dong, H., Jiang, Y. G., Wang, M., Xue, X., & Sigal, L. (2019). Vocabulary-informed zero-shot and open-set learning. *IEEE transactions on pattern analysis and machine intelligence, 42*, 3136–3152.

Ghifary, M., Kleijn, W.B., Zhang, M., & Balduzzi, D. (2015). Domain generalization for object recognition with multi-task autoencoders, In: Proceedings of the IEEE international conference on computer vision, pp. 2551–2559.

Gulrajani, I., & Lopez-Paz, D. (2021). In search of lost domain generalization. International Conference on Learning Representations .

Harary, S., Schwartz, E., Arbelle, A.. Staar, P. Abu-Hussein, S., Amrani, E., Herzig, R., Alfassy, A., Giryes, R., Kuehne, H. others. (2022). Unsupervised domain generalization by learning a bridge across domains, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5280–5290.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition, In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Huang, B., Bradbury, K., Collins, L.M., & Malof, J.M. (2020). Do Deep Learning Models Generalize to Overhead Imagery from Novel Geographic Domains? The xGD Benchmark Problem, In: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, IEEE, Waikoloa, HI, USA. pp. 1476–1479. https://doi.org/10.1109/IGARSS39084.2020.9323080.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K.Q. (2017). Densely connected convolutional networks, In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Kalita, I., Kumar, R., & Roy, M. (2021). Deep learning-based cross-sensor domain adaptation under active learning for land cover classification. *IEEE Geoscience and Remote Sensing Letters, 19*, 1–5.

Kalluri, T., Xu, W., & Chandraker, M. (2023). Geonet: Benchmarking unsupervised adaptation across geographies, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15368–15379.

Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S.M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., & Liang, P. (2021). Wilds: A benchmark of in-the-wild distribution shifts, In M. Meila, & T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, PMLR. pp. 5637–5664.

Li, C., Kim, D. J., Park, S., Kim, J., & Song, J. (2023). A self-evolving deep learning algorithm for automatic oil spill detection in Sentinel-1 SAR images. *Remote Sensing of Environment,299*, Article 113872. https://doi.org/10.1016/j.rse.2023.113872

Li, D., Yang, Y., Song, Y.Z., & Hospedales, T.M. (2017). Deeper, broader and artier domain generalization, In: Proceedings of the IEEE international conference on computer vision, pp. 5542–5550.

Li, Z., Ren, K., JIANG, X., Shen, Y., Zhang, H., & Li, D., (2023b). SIMPLE: Specialized model-sample matching for domain generalization. The Eleventh International Conference on Learning Representations .

Lin, C., Yuan, Z., Zhao, S., Sun, P., Wang, C., & Cai, J., (2021). Domain-invariant disentangled network for generalizable object detection, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8771–8780.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., & Zhou, J. (2024). Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing, 62*, 1–16. https://doi.org/10.1109/TGRS.2024.3390838

Lu, X., Zhong, Y., & Zhang, L. (2022). Open-source data-driven cross-domain road detection from very high resolution remote sensing imagery. *IEEE Transactions on Image Processing, 31*, 6847–6862. https://doi.org/10.1109/TIP.2022.3216481

Luo, M., & Ji, S. (2022). Cross-spatiotemporal land-cover classification from vhr remote sensing images with deep learning based domain adaptation. *ISPRS Journal of Photogrammetry and Remote Sensing, 191*, 105–128.

Ma, Y., Chen, S., Ermon, S., & Lobell, D. B. (2024). Transfer learning in environmental remote sensing. *Remote Sensing of Environment,301*, Article 113924. https://doi.org/10.1016/j.rse.2023.113924

Nations, U. (2022). Definition of Regions, World Population Prospects 2022 - Population Division - United Nations. https://population.un.org/wpp/DefinitionOfRegions/.

Nguyen, T. A., Rußwurm, M., Lenczner, G., & Tuia, D. (2024). Multi-temporal forest monitoring in the Swiss Alps with knowledge-guided deep learning. *Remote Sensing of Environment,305*, Article 114109. https://doi.org/10.1016/j.rse.2024.114109

Peng, D., Guan, H., Zang, Y., & Bruzzone, L. (2022). Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–17. https://doi.org/10.1109/TGRS.2021.3093004

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., & Wang, B. (2019). Moment matching for multi-source domain adaptation, In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 1406–1415.

Pott, L. P., Amado, T., Schwalbert, R. A., Corassa, G. M., & Ciampitti, I. A. (2021). Satellite-based data fusion crop type classification and mapping in Rio Grande do Sul, Brazil. *ISPRS Journal of Photogrammetry and Remote Sensing, 176*, 196–210. https://doi.org/10.1016/j.isprsjprs.2021.04.015

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. others. (2021). Learning transferable visual models from natural language supervision, In: International conference on machine learning, PMLR. pp. 8748–8763.

Rame, A., Dancette, C., & Cord, M. (2022). Fishr: Invariant gradient variances for out-of-distribution generalization, In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, PMLR. pp. 18347–18377.

Robey, A., Pappas, G. J., & Hassani, H. (2021). Model-based domain generalization. *Advances in Neural Information Processing Systems, 34*, 20210–20229.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision, 115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sadiq, R., Akhtar, Z., Imran, M., & Ofli, F. (2022). Integrating remote sensing and social sensing for flood mapping. *Remote Sensing Applications: Society and Environment,25*, Article 100697. https://doi.org/10.1016/j.rsase.2022.100697

Sagawa, S., Koh, P.W., Hashimoto, T.B., & Liang, P. (2019). Distributionally robust neural networks. International Conference on Learning Representations .

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., & Cui, P. (2021). Towards out-of-distribution generalization: A survey. arXiv .

Shi, Y., Seely, J., Torr, P., N, S., Hannun, A., Usunier, N., & Synnaeve, G. (2022). Gradient matching for domain generalization. International Conference on Learning Representations .

Shu, Y., Guo, X., Wu, J., Wang, X., Wang, J., & Long, M. (2023). CLIPood: Generalizing CLIP to Out-of-Distributions, In: Proceedings of the 40th International Conference on Machine Learning, PMLR. pp. 31716–31731. ISSN: 2640-3498.

Singha, M., Jha, A., Bose, S., Nair, A., Abdar, M., & Banerjee, B. (2024). Unknown prompt the only lacuna: Unveiling clip's potential for open domain generalization, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13309–13319.

Suel, E., Bhatt, S., Brauer, M., Flaxman, S., & Ezzati, M. (2021). Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. *Remote Sensing of Environment,257*, Article 112339. https://doi.org/10.1016/j.rse.2021.112339

Sun, B., & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In G. Hua & H. Jégou (Eds.), *Computer Vision - ECCV 2016 Workshops* (pp. 443–450). Cham: Springer International Publishing.

Tasar, O., Giros, A., Tarabalka, Y., Alliez, P., & Clerc, S. (2021). Daugnet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images. *IEEE Transactions on Geoscience and Remote Sensing, 59*, 1067–1081. https://doi.org/10.1109/TGRS.2020.3006161

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks, 10*, 988–999.

Venkateswara, H., Eusebio, J., Chakraborty, S., & Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation, In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5018–5027.

Volpi, R., & Murino, V., (2019). Addressing Model Vulnerability to Distributional Shifts Over Image Transformation Sets, In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South). pp. 7979–7988. https://doi.org/10.1109/ICCV.2019.00807.

Voreiter, C., Burnel, J.C., Lassalle, P., Spigai, M., Hugues, R., & Courty, N. (2020). A cycle gan approach for heterogeneous domain adaptation in land use classification, In: IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 1961–1964.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., & Yu, P. (2022). Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering,1–1*. https://doi.org/10.1109/TKDE.2022.3178128

Wu, J., Tang, Z., Xu, C., Liu, E., Gao, L., & Yan, W. (2020). Super-resolution domain adaptation networks for semantic segmentation via pixel and output level aligning. arXiv .

Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly, In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4582–4591.

Xie, M., Li, S., Yuan, L., Liu, C., & Dai, Z. (2023). Evolving Standardization for Continual Domain Generalization over Temporal Drift. *Advances in Neural Information Processing Systems, 36*, 21983–22002.

Xu, Q., Shi, Y., & Zhu, X. (2022a). Universal domain adaptation without source data for remote sensing image scene classification, In: International Geoscience and Remote Sensing Symposium (IGARSS), Institute of Electrical and Electronics Engineers Inc.. pp. 5341–5344. https://doi.org/10.1109/IGARSS46834.2022.9884889.

Xu, S., Zhang, S., Zeng, J., Li, T., Guo, Q., & Jin, S., (2020). A Framework for Land Use Scenes Classification Based on Landscape Photos. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing13*, 6124–6141. https://ieeexplore.ieee.org/document/9210779/, https://doi.org/10.1109/JSTARS.2020.3028158.

Xu, T., Sun, X., Diao, W., Zhao, L., Fu, K., & Wang, H. (2022). Fada: Feature aligned domain adaptive object detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–16.

Yao, H., Choi, C., Cao, B., Lee, Y., Koh, P., & Finn, C. (2022). Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time. *Advances in Neural Information Processing Systems, 35*, 10309–10324.

Yao, H., Choi, C., Cao, B., Lee, Y., Koh, P., & Finn, C. (2022). Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time. *Advances in Neural Information Processing Systems, 35*, 10309–10324.

Zhang, H., Cisse, M., Dauphin, Y.N., & Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. International Conference on Learning Representations.

Zhang, X., He, Y., Xu, R., Yu, H., Shen, Z., & Cui, P. (2023). NICO++: Towards Better Benchmarking for Domain Generalization, In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Vancouver, BC, Canada. pp. 16036–16047. https://doi.org/10.1109/CVPR52729.2023.01539.

Zheng, J., Wu, W., Fu, H., Li, W., Dong, R., Zhang, L., & Yuan, S. (2020). Unsupervised mixed multi-target domain adaptation for remote sensing images classification. *International Geoscience and Remote Sensing Symposium (IGARSS),1381–1384*. https://doi.org/10.1109/IGARSS39084.2020.9323602

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2022.3195549

Zhou, K., Yang, J., Loy, C.C., & Liu, Z. (2022b). Conditional Prompt Learning for Vision-Language Models, In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA. pp. 16795–16804. https://ieeexplore.ieee.org/document/9879913/, https://doi.org/10.1109/CVPR52688.2022.01631.