International Journal of Digital Earth

ISSN: 1753-8947 (Print) 1753-8955 (Online) Journal homepage: www.tandfonline.com/journals/tjde20

RSplitzero: generalized zero-shot learning in remote sensing across attribute splits with single and multi-modal representations

Lorenzo Stacchio, Lindo Nepi, Marina Paolanti & Roberto Pierdicca

To cite this article: Lorenzo Stacchio, Lindo Nepi, Marina Paolanti & Roberto Pierdicca (2025) RSplitzero: generalized zero-shot learning in remote sensing across attribute splits with single and multi-modal representations, International Journal of Digital Earth, 18:2, 2551869, DOI: 10.1080/17538947.2025.2551869

To link to this article: https://doi.org/10.1080/17538947.2025.2551869

Published online: 14 Sep 2025.

INTERNATIONAL JOURNAL OF DIGITAL EARTH
2025, VOL. 18, NO. 2, 2551869
https://doi.org/10.1080/17538947.2025.2551869

RSplitzero: generalized zero-shot learning in remote sensing across attribute splits with single and multi-modal representations

Lorenzo Stacchio[a], Lindo Nepi[b], Marina Paolanti[a] and Roberto Pierdicca[c]

[a]Department of Political Sciences, Communication and International Relations, University of Macerata, Macerata, Italy; [b]Dipartimento di Ingegneria dell'Informazione (DII), Universita' Politecnica delle Marche, Ancona, Italy; [c]Dipartimento di Ingegneria Civile, Edile e dell'Architettura (DICEA), Universita' Politecnica delle Marche, Ancona, Italy

ABSTRACT

Zero-shot learning (ZSL) emerged as a way to classify unseen categories using semantic knowledge from known ones. While widely studied in computer vision, its use in remote sensing (RS) is still limited. Given RS's high intra-class variability, fine-grained distinctions, and scarce labeled data, ZSL presents a promising classification solution. We investigate the generalisability of ZSL methods in the RS domain, focusing on attribute-level annotations. We build upon existing knowledge in Attribute-based ZSL (ABZSL) to evaluate deep-learning backbone generalizability and robustness across different semantic splits. We extend this framework to RS considering the WHU-RS19 dataset with novel attribute-level annotations, defining the WHU-RS19 ABZSL dataset. These annotations include 38 attributes, providing the first attribute-based benchmark for ZSL in RS. We evaluate generative ZSL methods under different classes and attribute splitting strategies, using features extracted by vision and multimodal backbones. Our results show that ZSL performance is sensitive to the backbone and splitting strategy. We found that DINOv2-based backbones achieved the highest generalization and robustness scores when using specific generative ZSL approaches (i.e. TFVAEGAN) attribute splitting strategies (i.e. PCA attributes splitting) on unseen classes (with a Generalized Harmonic Accuracy Mean of 84.30 and 70.55, respectively, on seen-unseen classes splits of 15-4 and 13-6).

ARTICLE HISTORY
Received 18 May 2025
Accepted 7 August 2025

KEYWORDS
Zero-shot learning; attribute-based zero-shot learning; remote sensing; deep learning

1. Introduction

In recent years, remote sensing (RS) has evolved with the availability of high-resolution satellite imagery and multispectral/hyperspectral data (MS/HS), contributing to major Earth Observation (EO) tasks in land use classification, environmental monitoring, and disaster evaluation (Aleissaee et al. 2023; Wang et al. 2019; Xu et al. 2023). Despite the growth of spectral RS data, key challenges remain. Indeed, a lack of labeled data limits the use of supervised learning, and MS/HS data are costly and complex to manage. RGB imagery is more accessible, but RS still faces difficulties due to the diversity of land cover and object types (Paheding et al. 2024). The vast number of classes and the high cost of expert labeling make it impractical to train supervised models for all categories. Additionally, real-world changes demand models that can recognize previously unseen classes (e.g. unobserved regions). To overcome this limitation, several approaches have been proposed to exploit existing knowledge and minimize the dependence on labeled samples, including semi-supervised learning (Reddy, Viswanath, and Reddy 2018), transfer learning (Tan et al. 2018), and zero-shot learning (ZSL) (Wang et al. 2020). ZSL addresses label scarcity by teaching models to classify unseen categories by leveraging semantic rela-tionships between known and unknown classes, exploiting auxiliary information, such as semantic attributes or textual descriptions. This makes it especially suitable for RS, where novel land cover types, emerging environmental patterns may emerge. Unlike unsupervised learning, which identifies patterns in completely unlabeled data, ZSL is a supervised approach that leverages labeled examples from seen classes and semantic auxiliary information to classify previously unseen categories (Wang et al. 2020).

CONTACT Lorenzo Stacchio lorenzo.stacchio@unimc.it

Despite the vast adoption of ZSL, its application in RS remains relatively underexplored (Liu et al. 2024). The challenges inherent in RS imagery – such as high intra-class variability, complex spatial relationships, and fine-grained class distinctions make it difficult to adapt ZSL methods developed for natural/satellite images. Recent approaches have explored the concept of semantic properties or attributes in RS (Rambabu et al. 2024; Xu et al. 2023). As visually depicted in Figure 1, unseen classes are identified based on shared semantic attributes from known classes. The Train Set typically consists of seen class images (e.g. football fields, airports, ponds, beaches), from which key attributes are extracted, including colors (e.g. gray in airports and football fields), objects (e.g. water in ponds and beaches), and shapes (e.g. lines in airports and beaches). The Test Set includes unseen classes, classified using these attributes and, optionally, additional ones.

However, modern approaches remain constrained by issues such as class/attribute bias, domain adaptation difficulties, and semantic loss. Recent research efforts have attempted to mitigate these limitations (Rossi et al. 2024; Xu et al. 2023). However, the robustness and generalization capability of these methods across different data-splitting strategies remain largely unexamined.

Another overlooked issue in ZSL research is the reliance on fixed seen-unseen splits, which can introduce biases and limit real-world applicability (high performance variation according to the split) (Rossi et al. 2024). This is especially problematic in RS, where interdependent and spatially correlated land cover classes need robust evaluation frameworks. Therefore, a systematic investigation into attribute-based data splits and their influence on ZSL model performance is necessary to develop more robust and interpretable classification frameworks. In conventional zero-shot learning (CZSL), the test set consists only of samples from unseen classes. However, this scenario is often impractical in real-world applications, where both seen and unseen classes coexist during training. To account for this, generalized zero-shot learning (GZSL) has been introduced, where models are evaluated on a combined set of seen and unseen classes (Rahman, Khan, and Porikli 2018; Wang et al. 2020; Ye, Hu, and Zhan 2021). Although GZSL has shown promising results (Liu and Ozay 2023), its performance remains limited by several challenges, including domain change, class bias, limitations in cross-domain transfer, and semantic loss (Pourpanah et al. 2022). Recent advances in generative methods have addressed some of these issues by synthesizing visual features for unseen classes, thus improving accuracy (Sun, Gu, and Sun 2021; Xian et al. 2018). In this case, a critical issue is the lack of robust and generalizable models across different data splits.

The evaluation of ZSL models typically relies on fixed class benchmark splits (Xian et al. 2018). While these splits provide a basis for a fair comparison, they may unintentionally limit the generalization of models to real-world scenarios. Authors of (Rossi et al. 2024) highlight this limitation, arguing that reliance on fixed benchmark splits raises concerns about whether higher performance reflects true model improvement or overfitting to specific splits. They emphasized the need to evaluate ZSL models under varied and realistic conditions and proposed new splitting strategies and metrics to assess generalisability and robustness, discussing generative attribute-based approaches as a possible solution to the generalization problem.

Considering previous discussions, this study aims to address a crucial question: *to what extent can ZSL methods generalize across different attribute splits in RS, and how does the choice of split and visual backbone affect the model's performance?*

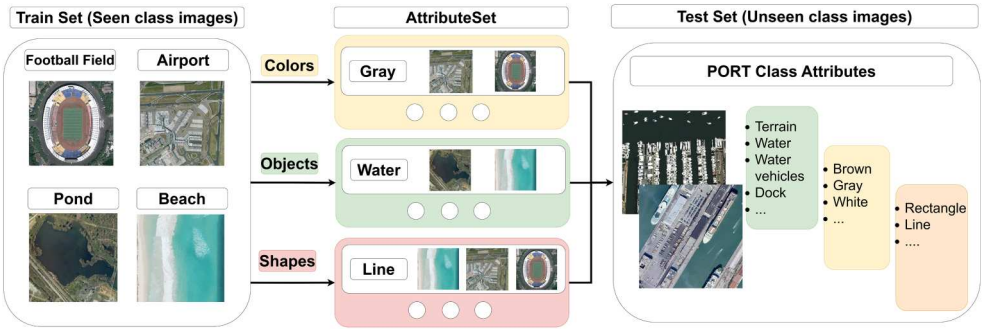


Figure 1. Visual description of attribute (semantic) based ZSL for RS. Exploiting attributes like colors, objects, and shapes from seen classes, a model is able to recognize unseen classes from them.

To answer this question, we introduce RSplitZero a comprehensive framework for evaluating the generalization and robustness of ZSL models in RS. Our work builds upon existing theoretical principles and extends them to the RS domain, introducing a novel attribute-annotated version of the WHU-RS19 dataset (Xia et al. 2010), named WHU-RS19 ABZSL. This dataset features 38 attribute-level annotations that capture key visual and contextual elements of RS images (and their classes), providing the first attribute based benchmark specifically designed for ZSL experiments in this field. We systematically analyze multiple class and attribute splitting strategies, including random splits, Clustered Class Splits (CCS), Class Greedy Splits (GCS), Minimal Attribute Splits (MAS), and PCA-based Attribute Splits (PAS). Moreover, we benchmark the performance of state-of-the-art generative ZSL models, including TFVAEGAN, CLSWGAN, and FREE, using CNN, Vision Transformer (ViT), and multimodal deep learning backbones. Our results demonstrate that ZSL performance in RS is sensitive to the choice of data splits and feature extraction strategies, leading to significant variations in standard ZSL evaluation metrics. By introducing RSplitZero and the WHU-RS19 ABZSL dataset, we lay the groundwork for more reliable and generalizable ZSL methods in RS, thereby paving the way for future advancements in attribute-driven learning frameworks.

To summarize, the key contributions of our work are listed below:

(i) We introduce a new dataset, *WHU-RS19 ABZSL*, comprising detailed image level attribute annotations for remote sensing imagery. To the best of our knowledge, this is the first dataset in the RS domain explicitly tailored for AB-ZSL tasks.
(ii) We adapt and extend a recent theoretical framework for AB-ZSL (Rossi et al. 2024), originally developed for natural images, to the RS domain. This includes a systematic evaluation across diverse attribute/class split strategies, generative ZSL methods, and visual backbones (CNNs, Transformers, and Mamba-based models), high lighting their relative strengths and weaknesses in handling RS-specific semantic representations.
(iii) We propose a novel *decremental evaluation protocol*, which progressively removes attribute information to simulate real-world scenarios with limited semantic descriptors. This approach provides deeper insight into model robustness and gen-eralization, especially under attribute sparsity, which is a common challenge in RS applications.

These contributions and the analysis made on the obtained results provide scientific evidence that could lead to novel methodologies to address the challenges associated with AB-ZSL in RS. This work paves the way for more robust and generalizable ZSL solutions in RS, addressing key challenges and advancing the state of the art in this field.

This paper is organized as follows: Section 2 reviews related work in ZSL with their applications in remote sensing. Section 3 describes the WHU-RS19 dataset, its novel attribute-level annotations, and their role in enabling attribute-based ZSL experiments in RS. Section 4 details the proposed methodology, including the RSplitZero framework, feature extraction using vision and multimodal backbones, and robustness evaluation of ZSL models across different attribute splits. Section 5 presents the experimental setup, including dataset details, evaluation metrics, and results obtained using the proposed framework. Finally, a thorough analysis of the performance of different models, backbones, and splitting strategies is provided, along with a conclusion of our work in Section 6.

2. Related works

The focus of ZSL is on the development of a model that is capable of recognizing objects from categories that have not been previously encountered. This is achieved by leveraging knowledge acquired from categories that have been previously encountered, with semantic information provided for both categories that have been seen and categories that have not been seen. The semantic information can be derived from pre-defined attribute vectors (Lampert, Nickisch, and Harmeling 2009), word or context-based embeddings (Fu et al. 2017; Socher et al. 2013), or combinations thereof (Song et al. 2020). These semantics are then used to establish links between known and novel categories.

In recent years, many approaches to ZSL have been developed. Early methods relied on carefully designed semantic representations for different categories. Typically, three main types of manually designed

semantic descriptors were used: visual attributes (Lampert, Nickisch, and Harmeling 2009; Palatucci et al. 2009), lexical descriptions (Ma, Cambria, and Gao 2016; Palatucci et al. 2009), and text-based keywords (Elhoseiny, Saleh, and Elgammal 2013; Lei Ba, Swersky, and Fidler 2015). For example, (Lampert, Nickisch, and Harmeling 2009) proposed an attribute-driven ZSL approach for classifying animals. These methods require the laborious and impractical process of human annotation of visual attributes, a process that is especially challenging in the context of large datasets. Recent research shows a shift toward using learning-based semantic representations, often leveraging pre-trained language models to extract semantic embeddings (Wang et al. 2016; Wang et al. 2019). By utilizing semantic embeddings, ZSL methods showed promising results, and can be categorized into the following three groups (Wang et al. 2019):

- *Mapping visual features to semantic space:* These methods project visual features into a semantic embedding space and utilize a process of comparison to determine the similarities of semantic embeddings (Bucher, Herbin, and Jurie 2016; Norouzi et al. 2013).
- *Mapping semantic features to visual space:* These techniques project semantic representations into the visual feature space (Pan et al. 2020; Zhang, Xiang, and Gong 2017).
- *Transforming both visual and semantic features into a common subspace:* This approach aligns visual and semantic representations in a shared embedding space (Demirel, Gokberk Cinbis, and Ikizler-Cinbis 2017; Ding, Shao, and Fu 2017)

Recently, ZSL has been shown to be a highly effective approach for the RS domain, due to its ability to recognize unseen classes by leveraging knowledge from seen classes, a capability that is crucial for this domain (Tan et al. 2024). However, in contrast to natural images, objects in RS images frequently exhibit significant structural and contextual diversity, which can make it challenging for models to learn reliable visual features for scene interpretation (Tan et al. 2024). Traditional ZSL methods in remote sensing (RS) use embedding-based techniques that map visual and semantic features into a shared space to recognize unseen classes (Tan et al. 2024). The first ZSL method for RS scene classification (RSSC) was proposed in (Li et al. 2017), using Word2Vec embeddings and a semantic graph to model class relationships. This was extended in (Quan et al. 2018.) with a semi-supervised Sammon embedding (Sammon 1969) to better align visual and semantic features. Other works include (Sumbul, Cinbis, and Aksoy 2017), which introduced a compatibility function to link image and semantic features, and (Wang, Peng, and Baets 2021), which proposed a distance-constrained semantic autoencoder for improved alignment.

A different approach based on GANs to zero-shot RSSC was proposed in (Li et al. 2022), where a generator synthesizes image features from class semantics, reframing ZSL as a standard classification task. Semantic embeddings were extracted using language models like Word2Vec (Wang et al. 2016) and BERT (Kenton and Toutanova 2019). Recent work in (Li et al. 2023) examined few-shot learning (FSL) for RS scene classification, comparing eight methods – three inductive and five transductive – across six datasets. Results showed transductive methods outperform inductive ones, and using five support samples yields better performance than using one. Interestingly, high similarity between unseen and training classes did not guarantee better results. In the textual domain, (Damalla et al. 2023) introduced a self-supervised method that generates word embeddings to describe RS scene classes, improving recognition of unseen categories by leveraging semantic relationships and addressing the scarcity of labeled data.

Employing CLIP multimodal models, (Al Rahhal et al. 2023) evaluated thirteen vision-language models (VLMs) on three RS scene datasets using a prompt-based classification approach, showing that larger VLMs perform better. Authors of (Liu et al. 2024) introduced RemoteCLIP, a foundation model tailored for RS tasks that learns rich visual and textual representations. By applying data scaling and incorporating UAV imagery, RemoteCLIP achieved SOTA ZS results despite limited pre-training data. Focusing on attribute-level predictions, (Rambabu et al. 2024) introduced a cross-semantic attribute-guided Transformer framework for zero-shot RS scene classification, combining semantic attribute localization with mutual learning between visual and semantic features to improve recognition of unseen classes. In (Xu et al. 2023), the authors proposed the Deep Semantic-Visual Alignment (DSVA) model, which automates the prediction of visually detectable attributes by measuring semantic-visual similarity. DSVA leverages transformer self-attention to relate local image regions and background context, mapping visual features into the attribute space for ZSL.

Despite the recent advancements in the current state of the art for ZSL and RS, the present study addresses the following gaps.

- *Gap 1: Lack of Attribute-Based ZSL Frameworks in RS.* Although significant progress has been made in ZSL for RSSC, existing approaches are class-centered, with a little exploration of attribute-based approaches (Rambabu et al. 2024; Xu et al. 2023), where classes are defined using fixed domain-specific semantic attributes. However, these approaches did not rely on human-aligned class attributes and did not explore the robustness of different DL backbones for non-fixed benchmark splits.
- *Gap 2: Dependence on fixed benchmark splits.* Current RS ZSL methods frequently depend on predetermined seen-unseen splits for evaluation that restricts their applicability to real-world scenarios, where class distributions are subject to variation. The lack of studies evaluating multiple attribute-splitting strategies outlined a limited understanding of model robustness.
- *Gap 3: Limited use of generative ZSL in RS.* Despite their vast adoption, generative ZSL methods (e.g. GANs and VAEs) remain underutilized in RS. These methods potentially address challenges such as high intra-class variability and fine-grained distinctions (Rossi et al. 2024). However, there is a lack of systematic benchmarking of these methods in RS, especially with robust vision and multimodal feature extractors.
- *Gap 4: Latent space fitness analysis.* Contrary to the findings of preceding studies, the present research utilizes representative visual backbones with an evaluation of the generalizability of various CNNs, Transformers, multimodal backbones, and Mamba-based ones across unseen classes and attributes, addressing the need to understand the robustness of their learned feature spaces in ZSL.

Considering these gaps, we present a novel attribute-annotated dataset here, evaluate several attribute-splitting strategies, and benchmark generative ZSL methods using state-of-the-art architectures in the RS context.

3. WHU-RS19 ABZSL dataset

We adopted the WHU-RS19 dataset (Xia et al. 2010), which contains 1,005 aerial images across 19 balanced classes (50 images per class) with 600 × 600 resolution. It's controlled intra- and inter-class variance, along with newly added attribute-level annotations, makes it a suitable benchmark for ZSL experimentation and backbone evaluation in remote sensing. Exploiting WHU-RS19 as our knowledge base, we initiated a data labeling process in March 2024 in collaboration with the University of Macerata and the Polytechnic University of Marche. In this data labeling phase, we annotate each example in this dataset according to 38 attributes to support the analysis of our attribute-based ZSL setting. Unlike previous datasets, our annotations were provided for each image in the dataset and then aggregated, specifically for the objectives of this study, differently from the previous ZSL attribute-based dataset, where the values were defined for each class without image-level annotations. Those novel attribute-level labels were furnished by expert remote-sensing annotators. Such a dataset is, to the best of our knowledge, the first dataset to explore and analyze the role of attribute based ZSL in the context of remote sensing. The novel dataset defined from this process, namely WHU-RS19 ABZSL, amounts to a foundational 'toy' one due to its balanced class distribution and manageable size, providing an ideal starting point for developing and testing methodologies in attribute-based ZSL within RS.

3.1. Attributes design

To enable ZSL analysis on the WHU-RS19 dataset, we developed a comprehensive set of 38 attributes that encapsulate the diverse elements present across its 19 scene classes. These attributes, designed along with remote sensing experts, encompass 19 various labels for objects (i.e. natural elements, man-made structures, transportation modes), 11 for geometric shapes, and 8 for dominant colors. The selected labels are:

> aircraft, road, bridge, road vehicles, houses, meadow, trees, mountains, terrain, water, sand, water vehicles, dock, stadium, football field, buildings, parking spots, rails, trains, red, orange, ochre, beige, green, light blue, blue, brown, gray, white, black, rectangle, triangle, square, sine wave, line, dashed line, curve, closed curve.

Each image within the dataset was meticulously annotated with these attributes, assigning a value between 0 and 1 to indicate the presence and prominence of each feature. This attribute-based labeling

scheme enabled us to investigate the relevance of different deep learning backbones for our considered attribute-based ZSL framework. By encompassing a broad spectrum of potential elements in the WHU-RS19 dataset, our attribute set ensures comprehensive coverage of the visual and contextual features inherent to each scene class (Xu et al. 2023).

3.2. Image annotation process

The annotation process followed a simple but strict protocol involving the following steps: (a) Attributes rules study and (b) Image Multi-label Annotation.

Our expert annotators (a) studied each definition of the possible attributes that can be present in the image. To date, we show them the presence of each attribute in exemplar images. Then, our experts catalog the presence of each attribute for each image, using Microsoft Excel software. We resort to 4 expert annotators (b), each cataloging 251 images, sampled balancing by the classes already exposed in the WHU RS19, for each of the designed 38 attributes. This results in a novel dataset that provides, per each of the 1, 005 images, a binary vector $v \in R^{38}$, resulting in a final matrix of $D \in R^{1005 \times 38}$. Here, we highlight that all these described processes show the uniqueness of these collected datasets. Since only trained annotators could have the ground truth, it is not possible to resort to just any standard labeling services (e.g. Amazon SageMaker).[1] This novel dataset was named WHU-RS19 ABZSL.

Qualitative examples of this novel attribute label set are reported in Figure 2. This depicts structured qualitative annotation of remote sensing imagery across four distinct classes (on a total of nineteen): Airport, Forest, Football Field, and Port. As mentioned, each image is annotated with three key dimensions: Objects, Dominant Colors, and Geometric Shapes. The Airport sample, includes annotations to prominent objects in the scene such as Aircraft, Roads, Meadows, Terrain, and Buildings, reflecting the infrastructure and open spaces typical of airport environments. The dominant colors, including Green, Brown, Gray, and White, represent a mix of natural and artificial elements. Finally, the Geometric shapes attributes amount to Rectangles, Triangles, Sine Waves, Lines, Dashed Lines, Curves, and Closed Curves highlight the structured layout and complex designs of runways, terminals, and other features. Similarly, in the Port category, annotations focus on objects such as Roads, Water, Sand, Water Vehicles, Docks, and Parking Spots, capturing the essential components of maritime infrastructure. The dominant colors, including Red, Beige, Light Blue, Blue, Brown, and Gray, reflect the interaction of natural water and sand with man-made structures. The primary geometric shape annotated amounts to a Rectangle only, emphasizing the linear and organized layout of docks, parking areas, and vessel arrangements.

4. Methodology

This section outlines the proposed methodology, RSplitZero, designed to evaluate the generalisability and robustness of ZSL models in RS across different attribute splits and DL backbones, visually depicted in Figure 3. We detail its key components in Section 4.1 and Section 4.3.

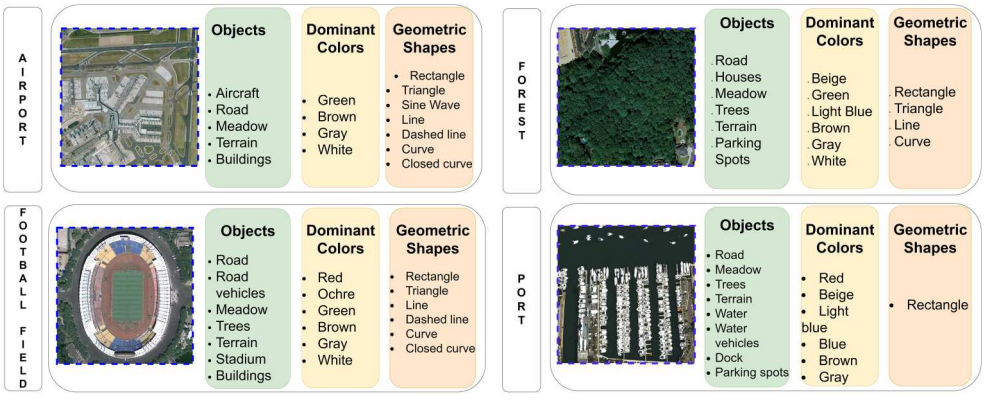


Figure 2. Qualitative examples of images labeled according to our attribute system. Top: labels for images in the Airport and Forest classes. Bottom: labels for images from the Football Field and Port classes.

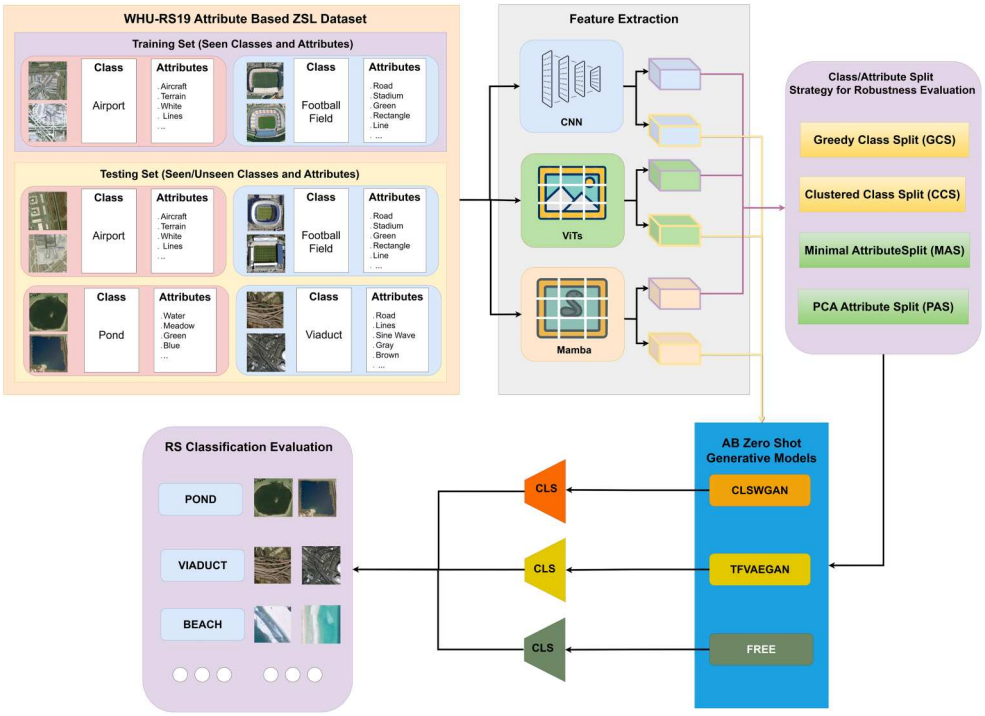


Figure 3. Overview of the RSplitZero Framework. The framework has three main components: Extraction of Vision and Multi-Modal Embeddings: Remote sensing images are passed through several feature extraction backbones, including CNNs (ResNet101, InceptionV3), vision transformers (ViT-B16, ViT-H14), multimodal models (ViT-CLIP-B16, ViT-CLIP-H14), and Mamba-based backbones (MambaVision-T). ZSL model classification and evaluation: Generative ZSL models (TFVAE-GAN, CLSWGAN, and FREE) are used to synthesize visual features for unseen classes. Class/Attribute splitting strategies: Following the methodology of (Rossi et al. 2024), the dataset is split into seen and unseen classes based on different attribute splitting strategies.

4.1. Extraction of vision and multi-modal embeddings

The first step of RsplitZero consists of projecting into the latent space all the images in our knowledge base (in our case the WHU-RS19 Attribute ZSL). Since one of our main contributions involves the analysis of different deep learning backbones for attribute based ZSL, we considered both vision and multi-modal architectures for the feature extraction process.

RsplitZero incorporates a diverse set of backbones for feature extraction, categorized into vision and multimodal approaches. Vision backbones include traditional convolutional neural networks (CNNs) such as ResNet50 (He et al. 2016) and InceptionV3 (Szegedy et al. 2016), as well as Vision Transformers (ViTs), which leverage transformer-based architectures like ViT (Dosovitskiy et al. 2021) and DinoV2 for advanced feature representation (Oquab et al. 2023). On the other hand, multimodal backbones focus on combining vision and language capabilities. ViT CLIP utilizes transformer-based architectures trained with a contrastive and multimodal vision-language approach (Radford et al. 2021). Finally, we also considered Mamba-based backbone (Hatamizadeh and Kautz 2025), which recently demonstrated SOTA performance on several vision tasks, maintaining high efficiency in terms of parameters.

This approach was adopted to compare the ZSL robustness vision backbones, trained with both supervised and self-supervised approaches, and multimodal backbones, in the case of attribute ZSL. We so adopted such backbones one at a time, to map the images in our dataset to the embedding space, which serves as the foundation for connecting visual features to semantic attributes in ZSL tasks. The robustness of these embedding spaces will be used to analyze their knowledge transfer capabilities to unseen classes, exploiting the method detailed in the following. Such feature extraction steps can be formalized as follows.

Let $X = \{\mathbf{X}_i\}_{i=1}^{n}$ represent a dataset of n remote sensing images, where each image $\mathbf{X}_i \in R^{H\times W\times C}$. The dimensions H and W vary according to the requirements of the specific backbone model used for feature

extraction. For example, ResNet50 requires input images of size 224×224 as ViT Base. The feature extraction function f parameterized by θ, representing the model's parameters:

$$f_\theta : \mathrm{R}^{H \times W \times C} \rightarrow \mathrm{R}^{\mathrm{d}},$$

where d is the dimensionality of the embedding vector space. The function f_θ maps an input image $\mathbf{X}_i$ to a d-dimensional embedding vector $\mathbf{z}_i = f_\theta(\mathbf{X}_i)$.

Applying f_θ to the entire dataset X results in a feature matrix $\mathbf{Z}$ of shape $n \times d$, where each row corresponds to the embedding of an image in the dataset. The dimensionality d of the embedding vector varies with the chosen backbone model. For example, ResNet50 and InceptionV3 both produce embeddings of dimension $d = 2048$. For ViT, the embedding dimension depends on the specific variant, such as $d = 768$ for ViT-B/16. Similarly, DINOv2 has embedding dimensions that vary by model size, such as $d = 768$ for ViT-B. The dimensionality for CLIP depends on its ViT backbones, which are the ones mentioned before. The same goes for the selected Mamba-based backbone.

4.2. Generative ZSL methods

In recent years, different generative approaches for ZSL were defined, in particular those that, by construction, synthesize features for unseen classes, thereby enabling the training of classifiers for both seen and unseen classes, in a supervised fashion. Among these methods, following insights from (Rossi et al. 2024), we considered FREE, CLSWGAN, and TF-VAEGAN. Those methods represent generative architectures to address domain shift and class bias in GZSL. FREE exploits a VAEGAN Feature Generator, where synthesized features are learned to match the real data distribution through adversarial and classification losses (Chen et al. 2021). CLSWGAN leverages Conditional WGAN to synthesize visual features conditioned on class attributes (Xian et al. 2018). TF-VAEGAN extends this paradigm by employing a two-stage approach, where it first generates visual features using a VAE-GAN and then learns a semantic embedding decoder that enforces semantic alignment (Narayan et al. 2020).

4.3. Zero-shot learning robustness analysis

In this section, we provide a full description of the proposed Attribute-based ZSL methodology visually depicted in Figure 3, illustrating the splitting methods that we adopted, inspired by (Rossi et al. 2024) to evaluate feature robustness.

4.3.1. Preliminary of attribute-based zero-shot learning problem

Let m represent the number of attributes, d the number of features, and C the total number of classes, which include both seen and unseen categories. The set C contains all classes, with C_s denoting the count of seen classes ($C_s = |S|$) and C_u representing the number of unseen ones. We define A as the semantic or attribute space and Z as the feature space. The subset C_s consists of observed classes, whereas C_u includes those that have not been seen during training. Similarly, Z_s corresponds to the feature representations of seen classes, while Z_u pertains to the feature representations of unseen ones. Each class c is identified by a function $\phi(c)$, which maps it to an attribute vector within the semantic space. Each attribute a_k indicates whether it is present (1) or absent (0) in a given class (we employed a continuous encoding scheme for attributes (range [0, 100]) to reflect their relevance in a given class). The labeled dataset is defined as triples (z, y, a_k) for both seen and unseen classes, but training relies only on a subset $D_s* \subseteq D_s$ of (feature, label) pairs from the seen classes.

In the standard ZSL framework, we define an unknown function

$$\widehat{f} : Z \rightarrow C$$

that maps feature representations to class labels while minimizing the classification error $\mathrm{E} = c - c\text{^}$. The training set consists exclusively of pairs from seen classes:

$$D_s = \{(z, y) | z \in Z_s, y \in C_s\}.$$

The model's performance on these seen classes is quantified using:

$$\text{Accuracy}_{Cs},$$

which is computed solely on the seen-class test samples.

In the generalized setting (GZSL), evaluation includes both seen (C_s) and unseen (C_u) classes. The performance is assessed by computing the harmonic mean between the two accuracies:

$$\text{H-Mean} = \frac{2 \cdot \text{Accuracy}_{Cs} \cdot \text{Accuracy}_{Cu}}{\text{Accuracy}_{Cs} + \text{Accuracy}_{Cu}}$$

The *H-Mean* metric outlines possibly uneven performance, providing high value only with high seen and unseen accuracies. It is used to check whether a model exhibits a more robust evaluation of ZS generalization. In this work, we adopted those approaches exploiting different visual backbone embeddings to refine the function $\hat{f}$ by leveraging relationships between the feature space Z and the attribute space A. On the other hand, generative methods, such as those employed in this work, involve training a generator G to synthesize a dataset D_{gen}. The generated samples are then merged with the original training data to form an augmented set $T_c = D_t \cup D_{gen}$, subsequently used to train a conventional classifier.

4.3.2. Split and performance gap

We evaluate the ZSL framework by applying different models and dataset *partitions*, established according to the guidelines and segmentation methods detailed in (Rossi et al. 2024). We categorize these partitions into two groups: *class partitions* and *attribute partitions*. Class partitions separate classes into two exclusive groups: seen classes, which are included during training, and unseen classes, which are reserved for testing. Conversely, attribute partitions define a configuration of the semantic space by selecting a subset of semantic descriptors, known as *attributes*, that uniquely characterize each class.

Utilizing these partitions, we examine the notion of Performance Gap (PG), which quantifies the discrepancy in accuracy between a ZSL model and its theoretical *Upper Bound (UB)*. Let M denote the set of all ZSL models, D the dataset collection, and Σ the complete set of partitions. For a dataset $d \in D$ and a partition $\sigma \in \Sigma$, let $g \in M$ be a ZSL model, and $\text{acc}(g, \sigma, d)$ its accuracy measured on d under partition σ. We further define $g^* \in M$ as an equivalent non-ZSL model, and $\sigma^* \in \Sigma$ as the *full partition*, in which all classes are considered seen. The upper bound is thus expressed as:

$$\text{UB} = \text{acc}(g^*, \sigma^*, d).$$

The Performance Gap for model m, partition σ, and dataset d is then defined as:

$$\text{PG}(m, \sigma, d) = \text{E}\,[\text{UB} - \text{acc}(m, \sigma, d)].$$

By computing PG across different models, partitions, and datasets, we can objectively determine how much performance is lost due to the absence of unseen-class supervision or semantic knowledge: a smaller PG indicates that a ZSL model is effective at bridging the semantic and data gaps.

4.3.3. Splitting methods and robustness evaluation

To analyze the stability of feature representations extracted from visual backbones and ZSL models, we introduce a collection of partitioning techniques: Clustered Class Split (CCS), Greedy Class Split (GCS), PCA Attribute Split (PAS), and Minimal Attribute Split (MAS) (Rossi et al. 2024). For clarity, each approach is described in terms of binary attributes. The attribute/class partitioning method restructures seen and unseen attributes/classes by ranking the classes and selecting the top subset top_S. The seen attribute/class set is initially defined as a function of an ordered set $\tilde{T}$: $T_S = \{t_i \in \tilde{T} \mid i \leq top_S\}$, where $T = C$ when the partitioning method is for classes and $T = A$ for attributes. In the following, we describe the adopted class and attribute splitting methods. It is worth mentioning that the split methods over attributes (MAS, PAS) and classes (GCS, CCS) will be compared to our theoretical UB, which corresponds to the highest achievable accuracy under a toy ZSL setting (where only 20% of the classes are unseen and all the attributes were adopted) (Rossi et al. 2024).

Greedy Class Split This partitioning strategy is designed to retain as much semantic consistency as possible within the seen class set C^S. The goal is to prevent cases where, for instance, a *zebra* is characterized as a *horse with stripes*, yet no training instances include the attribute *stripes*. To increase the informativeness of the semantic space, GCS maximizes the number of present attributes (i.e. entries with value 1) in C^S by sorting classes based on the norm of their binary attribute vectors, yielding the final ordered class set.

Clustered Class Split The CCS selects C^S and C^U as separate clusters with the objective of reducing intra-cluster variation while enhancing inter-cluster separation. When the classes in C^S exhibit strong similarity, the model can better interpret attribute representations as multiple classes share overlapping examples. Conversely, if C^S contains highly diverse classes, the risk of overfitting diminishes. Clusters are generated by ranking classes based on the sum of their distances. The top_S classes, having the lowest overall distances, form a cluster in the semantic space and are assigned as seen classes. In contrast, the remaining n_U classes, which are farther from this cluster, are grouped separately.

Minimal Attribute Split The attribute partitioning methods, Minimal Attribute Split (MAS) and PCA Attribute Split (PAS), introduced in (Rossi et al. 2024), aim to optimize the attribute set by refining its structure. The MAS approach restructures the semantic space by eliminating attributes that exhibit high redundancy due to strong correlations. This selection prioritizes attributes that provide the most distinctive information.

PCA Attribute Split The Principal Component Analysis Attribute Split (PAS) transforms the attribute space, reducing dimensionality by generating a new set of attributes that encapsulate the most relevant information from the original dataset. The initial attribute matrix, which contains attribute vectors for all classes, undergoes normalization. Then the covariance matrix is computed and decomposed into its eigenvalues and eigenvectors. By sorting the eigenvalues in descending order, the top eigenvectors corresponding to the largest eigenvalues are chosen to construct the final attribute matrix, forming the semantic space for PAS. Both classes and attribute splitting strategies follow the definition from (Rossi et al. 2024). It is important to note that class splits are used to benchmark our models and to evaluate the goodness of the different visual feature backbones to improve the performance of a ZSL model. The attribute splits can be used for testing real-world scenarios where only a few attributes are known, which is key for the remote sensing domain.

5. Experiments

The following section provides a comprehensive analysis and discussion of a series of experiments designed to evaluate the generalizability and robustness of the selected visual feature backbones exploiting the following ABZSL models: CLSWGAN (Xian et al. 2018), TF-VAEGAN (Narayan et al. 2020), and FREE (Chen et al. 2021). The analysis of our experimental campaign is organized as follows. Section 5.1 we report our experimental setting. Then, Section 5.2, we report the results obtained evaluating our UB, for several backbones and the adopted generative methods. In Section 5.3 analyzes the role of the selected class split against a random one, to ensure a proper generalization analysis setting. On top of that, Section 5.4 reports the results obtained for our upper bound analysis, which involves using the prefixed class split to analyze the performance of the different backbones and generative approaches. Selecting the best backbone and splitting strategy, we then report in Section 5.6, the comparison of our selected split (15 seen classes, 4 unseen) with a harder one, which better reflects the Remote Sensing domain under analysis (13 seen, 6 unseen) to explore which backbone and splitting strategy, reports the lower drop in performance (Performance Gap).

5.1. Experimental setting

As mentioned, in all our experiments we adopted the here introduced WHU-RS19 ABZSL dataset and we evaluated the performance of both CZSL and GZSL approaches, employing three generative methods: CLSWGAN, TFVAEGAN, and FREE. Each method was tailored to leverage the specific characteristics of WHU-RS19 ABZSL, which includes 19 classes and 38 attribute dimensions. We followed the main indications from (Rossi et al. 2024), adapting the framework to deal with our custom dataset. For all the considered generative methods, we adopted all the sets of attributes (the adopted splitting strategies will then split that) and a hidden layer size of 4096, generating a number of 1800 features per class, following the same

approach of (Rossi et al. 2024). A batch size of 512 was selected, with a gradient penalty weight of 10 and a pre-classifier loss weight of 0.02. As an optimizer, we adopted Adam, with a beta1 value set to 0.05, for a total of 60 epochs for the training.

Specifically for FREE, we adopted one critic iteration for each generator update and two training loops per epoch. Learning rates for both the generator and classifier were 0.00001 and 0.001, respectively. Regarding CLSWGAN, the discriminator was updated in five iterations for every generator update, and training utilized two loops per epoch. The generator employed a learning rate of 0.00001, while the classifier used a learning rate of 0.001. For TFVAEGAN, feedback-based learning was adopted, employing learning rates of 0.0001 for the feedback module, while the classification learning rate was 0.001. The decoder was frozen during stability training.

All three generative paradigms and settings were employed using features coming from the following visual backbones: ResNet101, InceptionV3, VITB16-224, VITH14-224, CLIP-B16, CLIP-H14-224, DINOv2-B14, DINOv2-L14, MambaV-B-1 K, MambaV-L-1 K. In particular, to compare generalization over unseen classes and attributes, we put this entire experimental setting under two increasing levels of class splitting:

- 15 seen classes, 4 unseen classes (20%), with 787 training samples (1005 total);
- 13 seen classes, 6 unseen classes (30%), with 687 training samples (1005 total);

This splitting strategy was built on top of (Rossi et al. 2024), where the authors picked a 10% class sample to analyze both CZSL and GZSL. In this work, considering that we have at our disposal fewer classes and attributes (i.e. 19) and considering the real-world applications of RS image analysis, we increased this percentage to 20% and 30%.

5.2. Upper bound analysis

We report results obtained for the defined CZSL accuracy across different feature extractors and generative models under a class split of 15-4, with the attribute split fixed to a null split. By employing a null split, no division occurs within the class/attribute space, creating a controlled environment for evaluation. The overall results are reported in Table 1. In this case, we will comment only on the results obtained with a fixed null split (i.e. where split type equals nosplit).

The choice of feature extractor significantly impacts CZSL accuracy, as evidenced by the obtained results. Among the evaluated models, ResNet101 achieves its highest accuracy with the TFVAEGAN model, reaching 59.662 and outperforming both the FREE baseline and CLSWGAN. Similarly, Inceptionv3 benefits from CLSWGAN, reaching its highest ZSL accuracy of 50.092 and GZSL harmonic mean of 54.35, highlighting the role of generative models in enhancing the performance of CNN-based extractors, as discussed in (Rossi et al. 2024).

Transformer-based architectures, such as VIT and DINOv2 models, consistently outperform traditional CNNs in CZSL tasks. VITB16 achieves its best accuracy of 73.023 with TFVAEGAN, significantly outperforming its results with the FREE and CLSWGAN paradigms. DINOv2 models further extend this trend, with DINOv2-B14 and DINOv2-L14 reaching 83.014 and 80.093, respectively, with TFVAEGAN. These results underscore the superior feature representation capabilities of transformer-based models, particularly when combined with generative methods that enhance their ability to generalize across unseen categories.

Multimodal models, particularly CLIP-based architectures, demonstrate better performance, achieving high CZSL seen accuracies. CLIP-B16, for instance, attains its best accuracy of 75.184 with TFVAEGAN, while CLIP-H14 has comparable results. The same goes for the GZSL harmonic mean, where CLIP-B16 achieves a score of 71.562 with the TFVAEGAN paradigm. However, the variation in performance across generative models indicates that CLIP-based architectures are more sensitive to the choice of training methodology.

Finally, DINOv2 achieves the overall best results. Indeed, DINOv2-B14 reached a CZSL of 83.014 and a harmonic mean of 77.13. Its larger version, DINOv2-L14, obtained an overall harmonic mean GZSL of 78.361, but slightly lower CZSL at 80.093. The TFVAEGAN approach yielded all the best-obtained results.

Table 1. Comparison of ZSL Accuracy across Different Models and Methods, for the split 15-4. FE stands for Feature Extractor, GM stands for Generative Method, S-ACC, and U-ACC, respectively, for seen and unseen class accuracies. H-Mean represents the harmonic mean between the S-ACC and U-ACC.

			Metrics (%)			
FE	ZSLParadigm	SplitType	ZSL ACC	GZSL S-ACC	GZSL U-ACC	GZSL H-Mean
ResNet101	FREE	nosplit	54.53	**87.92**	32.365	47.314
	FREE	rnd	**54.652**	75.775	**40.755**	**53.003**
	CLSWGAN	nosplit	50.647	65.838	37.64	47.9
	CLSWGAN	rnd	58.764	77.179	40.893	53.46
	TFVAEGAN	nosplit	**59.662**	70.439	**42.325**	52.877
	TFVAEGAN	rnd	51.393	**92.875**	38.606	**54.54**
Inceptionv3	FREE	nosplit	45.767	61.862	37.919	47.017
	FREE	rnd	**47.41**	**73.436**	**44.624**	**55.514**
	CLSWGAN	nosplit	**48.467**	74.125	**42.904**	**54.35**
	CLSWGAN	rnd	47.903	77.236	26.756	39.74
	TFVAEGAN	nosplit	50.092	**88.069**	37.164	52.271
	TFVAEGAN	rnd	41.981	72.889	**42.099**	**53.371**
VITB16	FREE	nosplit	59.108	59.258	54.48	**56.769**
	FREE	rnd	**66.488**	**63.926**	**48.587**	55.211
	CLSWGAN	nosplit	43.25	49.652	28.931	36.56
	CLSWGAN	rnd	54.193	71.00	39.355	50.64
	TFVAEGAN	nosplit	**73.023**	79.499	56.973	66.377
	TFVAEGAN	rnd	64.051	**76.00**	**59.181**	**66.544**
VIT-H14	FREE	nosplit	54.152	**81.084**	**47.789**	**60.135**
	FREE	rnd	**68.921**	76.667	43.895	55.827
	CLSWGAN	nosplit	**51.628**	58.294	32.022	41.34
	CLSWGAN	rnd	51.153	61.501	34.918	44.55
	TFVAEGAN	nosplit	66.083	83.711	46.453	59.749
	TFVAEGAN	rnd	59.704	77.977	**51.248**	**61.849**
CLIP-B16	FREE	nosplit	61.952	75.244	**49.038**	59.378
	FREE	rnd	**61.996**	**80.026**	47.43	**59.56**
	CLSWGAN	nosplit	45.257	55.664	**35.86**	43.62
	CLSWGAN	rnd	53.337	79.538	29.864	43.42
	TFVAEGAN	nosplit	**75.184**	**86.591**	**60.978**	**71.562**
	TFVAEGAN	rnd	70.755	82.396	56.794	67.24
CLIP-H14	FREE	nosplit	66.567	89.158	**55.275**	68.242
	FREE	rnd	**69.421**	**90.467**	59.044	**71.453**
	CLSWGAN	nosplit	55.797	**84.984**	40.517	54.87
	CLSWGAN	rnd	65.257	84.801	48.325	61.57
	TFVAEGAN	nosplit	69.893	87.664	50.684	64.232
	TFVAEGAN	rnd	**72.257**	**91.493**	**62.345**	**74.158**
DINOv2-B14	FREE	nosplit	65.783	83.019	55.403	66.456
	FREE	rnd	**68.033**	**88.194**	**55.77**	**68.33**
	CLSWGAN	nosplit	54.579	**65.206**	**31.575**	**42.55**
	CLSWGAN	rnd	58.139	60.698	25.431	35.84
	TFVAEGAN	nosplit	**83.014**	**87.197**	**69.147**	**77.13**
	TFVAEGAN	rnd	70.45	79.154	61.811	69.416
DINOv2-L14	FREE	nosplit	**66.198**	**91.788**	**57.344**	**70.589**
	FREE	rnd	65.607	89.997	50.709	64.868
	CLSWGAN	nosplit	54.963	**80.6**	**31.082**	**44.86**
	CLSWGAN	rnd	56.846	59.923	20.266	30.29
	TFVAEGAN	nosplit	**80.093**	**94.331**	**67.016**	**78.361**
	TFVAEGAN	rnd	78.293	95.256	63.508	76.208
MambaV-B-1K	FREE	nosplit	54.889	70.902	41.301	52.197
	FREE	rnd	**60.383**	**72.746**	**49.856**	**59.164**
	CLSWGAN	nosplit	**51.655**	39.176	**40.716**	39.930
	CLSWGAN	rnd	50.553	59.578	37.702	46.180
	TFVAEGAN	nosplit	62.269	56.064	**59.688**	57.820
	TFVAEGAN	rnd	**63.022**	**77.359**	50.820	**61.342**
MambaV-L-1K	FREE	nosplit	50.773	**86.277**	39.569	54.255
	FREE	rnd	**55.280**	68.607	**48.672**	**56.945**
	CLSWGAN	nosplit	50.394	52.729	29.216	37.600
	CLSWGAN	rnd	55.164	68.536	34.741	46.110
	TFVAEGAN	nosplit	62.034	**84.309**	46.468	**59.914**
	TFVAEGAN	rnd	**66.226**	54.151	**61.405**	57.550

Considering now the role of the generative ZSL approaches (FREE, CLSWGAN, and TFVAEGAN), these exhibit distinct performance patterns. FREE generally surpasses the CLSWGAN baseline, confirming its ability to synthesize realistic class representations that improve CZSL performance (Rossi et al. 2024). However, its performance remains inconsistent across different feature extractors, often trailing behind TFVAEGAN. TFVAEGAN emerges as the most effective generative model, consistently achieving the best ZSL accuracy across most feature extractors. Its robustness and ability to exploit our rhetorical UB setting (null split) configuration make it the preferred choice for improving unseen class generalization.

5.3. Generalizability evaluation pre-requisite analysis

We analyzed how random class splits affected CZSL and GZSL. We so picked classes randomly with different seeds (keeping the UB split fixed), the seen/unseen classes. Since such a split influences both generative training and downstream classification, every run gets a different synthetic dataset to train against. If changes in accuracy are introduced by the splits, and those are too large, it may affect robustness analysis. The results are reported in Table 1.

The results indicate that applying a random split instead of the fixed 'nosplit' configuration does not substantially alter the performance across different feature extractors and generative ZSL paradigms. In most cases, variations in ZSL accuracy and GZSL metrics remain within a small range, suggesting that the choice of the split does not introduce significant bias and that the models exhibit relative stability across different data partitioning strategies. However, some deviations are observed, particularly in ResNet101 with CLSWGAN, where ZSL accuracy improves from 50.647 in the 'nosplit' setting to 58.764 in the random split, alongside an increase in GZSL harmonic mean from 47.9 to 53.46. Similarly, for VITB16 with FREE, the GZSL harmonic mean improves from 36.56 to 50.64 when using the random split. All the other improvements/decreases amount to a maximum of 12 points of accuracy.

TFVAEGAN achieves the highest ZSL and GZSL accuracies across almost every feature extractor, in any split type (nosplit or random), strengthening its being a truly general model. FREE often surpasses CLSWGAN, especially with ZSL and GZSL unseen accuracy, eg, with ViT-B16 (random split), FREE achieves 66.49 vs 54.19 for CLSWGAN. TFVAEGAN performs so clearly better than CLSWGAN, e.g. with DINOv2-B14 (nosplit: 83.01 vs 54.58, random: 70.45 vs 58.14), which outlines that the extracted features are more robust for generation. Again, GZSL confirms this result: TFVAEGAN with CLIP-H14 (random split) reaches 74.16 vs. 71.45 for FREE; with DINOv2-L14, the gap is a bit smaller between base and random split (76.21 vs. 64.87 for FREE). When evaluating MambaVision, similar performances were observed. TFVAEGAN consistently yields better ZSL and GZSL performance compared to CLSWGAN across both MambaVision-B and MambaVision-L. For instance, in the Mamba-B nosplit setting, TFVAEGAN achieves 62.27% vs. 51.66% for CLSWGAN (ZSL), and 57.82% vs. 39.93% (GZSL-H). Even in random splits, TFVAEGAN maintains an edge (61.34% vs. 46.18% GZSL-H). Similar trends are seen for Mamba-L: TFVAEGAN achieves 62.03% vs. 50.39% (ZSL, nosplit) and 66.23% vs. 55.16% (ZSL, random), with corresponding gains in GZSL as well. Generally, DINOv2-L14 performed best, even though CLIP-B16 and MambaVision-B are more robust under randomness.

We will now proceed to analyze the robustness of the different generative methods, under the selected class (GCS, CCS) and attribute (MAS, PAS) splits, considering the most performant models.

5.4. ZSL splitting strategies robustness

For simplicity, we will compare the different selected class and attribute splitting methods, fixing the same ZSL-AB generative approach only for those backbones that exhibited the highest performance, according to their architectural class (CNN, ViT, Multimodal, and Self-Supervised ViT). We here recall that the 'nosplit' type, amounts to our theoretical UB, where all the available attributes (38/38) and classes (15/19) we used. Results for FREE, CLSWGAN, and TFVAEGAN are respectively reported in Tables 2–4.

The results obtained with the FREE method (Table 2), highlighted a clear difference in feature extractors and splitting strategy robustness. Indeed, DINOv2-L14 consistently achieved the highest ZSL accuracy (82.488%) and harmonic mean (79.422%) under the PAS19 split, demonstrating its superior robustness in transferring knowledge to unseen classes. Similarly, CLIP-H14 also exhibited

Table 2. Comparison of ZSL Accuracy across Different Models under the FREE method, for the split 15-4. FE stands for Feature Extractor, S-ACC, and U-ACC, respectively for seen and unseen class accuracies. H-Mean represents the harmonic mean between the S-ACC and U-ACC.

		Metrics (%)			
FE	Split Type	ZSL ACC	GZSL S-ACC	GZSL U-ACC	GZSL H-Mean
ResNet101	nosplit	54.53	87.92	32.365	47.314
	ccs	53.625	**89.653**	28.992	43.815
	gcs	49.41	72.929	36.459	48.614
	mas19	**59.673**	66.261	38.084	48.368
	pas19	57.139	78.854	**41.519**	**54.397**
VIT-H14	nosplit	54.152	**81.084**	47.789	60.135
	ccs	58.683	63.196	39.527	48.635
	gcs	51.033	57.59	44.738	50.357
	mas19	57.334	79.936	41.18	54.357
	pas19	**65.749**	80.126	**57.702**	**67.09**
CLIP-H14	nosplit	**66.567**	89.158	55.275	68.242
	ccs	64.461	91.301	56.008	69.427
	gcs	61.418	**96.247**	46.795	62.973
	mas19	61.994	91.784	55.551	69.212
	pas19	64.909	83.992	**61.897**	**71.271**
DINOv2-L14	nosplit	66.198	91.788	57.344	70.589
	ccs	68.66	83.176	51.998	63.992
	gcs	56.615	88.499	48.221	62.427
	mas19	79.501	**96.315**	58.481	72.774
	pas19	**82.488**	89.53	**71.366**	**79.422**
MambaV-B-1K	nosplit	54.889	70.902	41.301	52.197
	ccs	57.704	70.641	36.078	47.763
	gcs	45.877	63.262	43.213	51.350
	mas19	55.879	64.285	49.648	56.027
	pas19	**59.862**	**64.762**	**54.727**	**59.323**

strong performance, particularly in unseen class accuracy (61.897%) and overall generalization (H-Mean: 71.271%), reinforcing the effectiveness of multimodal vision transformers in the ZSL setting. For VIT-H14, the PAS19 split emerged as the most effective, leading to the highest ZSL accuracy

Table 3. Comparison of ZSL Accuracy across Different Models under the CLSWGAN method, for the split 15-4. FE stands for Feature Extractor, GM stands for Generative Method, S-ACC, and U-ACC respectively for seen and unseen class accuracies. H-Mean represents the harmonic mean between the S-ACC and U-ACC.

		Metrics (%)			
FE	Split Type	ZSL ACC	GZSL S-ACC	GZSL U-ACC	GZSL H-Mean
ResNet101	nosplit	50.647	65.838	37.64	47.9
	ccs	57.711	65.158	40.509	49.96
	gcs	**62.172**	**86.576**	**41.197**	**55.83**
	mas19	55.43	79.024	33.296	46.85
	pas19	58.026	55.476	36.641	44.13
VIT-H14	nosplit	51.628	58.294	32.022	41.34
	ccs	46.707	42.875	39.465	41.1
	gcs	44.877	60.63	30.393	40.49
	mas19	**58.07**	73.024	**41.999**	**53.33**
	pas19	41.003	**77.061**	25.714	38.56
CLIP-H14	nosplit	55.797	84.984	40.517	54.87
	ccs	48.724	86.559	**45.557**	**59.7**
	gcs	49.861	75.607	44.016	55.64
	mas19	58.983	**87.524**	33.481	48.43
	pas19	**60.982**	63.949	30.60	41.39
DINOv2-L14	nosplit	54.963	80.6	31.082	44.86
	ccs	**58.131**	78.586	38.809	51.96
	gcs	50.303	80.573	30.049	43.77
	mas19	52.492	**85.614**	37.286	51.95
	pas19	55.233	80.89	**47.341**	**59.73**
MambaV-B-1K	nosplit	**51.655**	39.176	**40.716**	39.930
	ccs	46.277	56.046	20.826	30.370
	gcs	44.303	47.456	38.754	42.670
	mas19	43.856	**74.226**	38.605	**50.790**
	pas19	38.766	50.774	29.003	36.920

Table 4. Comparison of ZSL Accuracy across Different Models under the TFVAEGAN method, for the split 15-4. FE stands for Feature Extractor, S-ACC, and U-ACC respectively for seen and unseen class accuracies. H-Mean represents the harmonic mean between the S-ACC and U-ACC.

		Metrics (%)			
FE	Split Type	ZSL ACC	GZSL S-ACC	GZSL U-ACC	GZSL H-Mean
ResNet101	nosplit	59.662	70.439	42.325	52.877
	ccs	51.177	70.124	31.286	43.268
	gcs	44.049	59.324	43.467	50.172
	mas19	61.421	80.765	41.743	55.039
	pas19	**66.364**	**83.91**	**46.204**	**59.594**
VIT-H14	nosplit	66.083	**83.711**	46.453	59.749
	ccs	67.462	78.647	51.662	62.36
	gcs	65.525	64.892	45.303	53.357
	mas19	64.274	81.166	52.415	63.696
	pas19	**68.603**	76.297	**66.734**	**71.196**
CLIP-H14	nosplit	69.893	87.664	50.684	64.232
	ccs	77.292	**94.254**	57.499	71.425
	gcs	61.074	87.569	47.656	61.722
	mas19	**77.305**	93.521	58.388	71.892
	pas19	71.635	89.507	**63.809**	**74.504**
DINOv2-L14	nosplit	80.093	94.331	67.016	78.361
	ccs	63.478	94.464	57.008	71.105
	gcs	61.721	91.693	50.615	65.225
	mas19	**89.471**	**96.184**	74.77	84.136
	pas19	83.698	93.784	**76.562**	**84.302**
MambaV-B-1K	nosplit	62.269	56.064	59.688	57.820
	ccs	64.776	72.914	45.312	55.891
	gcs	47.410	76.457	38.656	51.350
	mas19	61.141	**78.700**	39.320	52.440
	pas19	**68.986**	64.775	**63.248**	**64.003**

(65.749%) and harmonic mean (67.09%), further validating the benefits of attribute-driven splitting strategies. Conversely, ResNet101, a CNN-based model, performed best under the MAS19 and PAS19 splits, achieving 59.673% and 57.139% ZSL accuracy, respectively, indicating its relatively lower generalizability compared to transformer-based architectures. Finally, MambaV-B-1 K demonstrates better performance than Resnet101 but lower than the others: the best ZSL accuracy (59.862%) and harmonic mean (59.323%) observed under the *pas19* split. Across all models, the attribute-based PAS19 split generally outperformed other strategies, suggesting that carefully structured attribute splits improve unseen class generalization more effectively than class-based splits.

Considering now the CLSWGAN method, reported in Table 3, ResNet101 achieved the highest ZSL ACC of 62.172% under the GCS split, indicating that class-based splitting strategies can significantly affect model performance. Similarly, DINOv2-L14 performed best with the PAS19 split, achieving an unseen accuracy of 47.341% and an H-Mean of 59.73%, reinforcing the effectiveness of attribute-based splits and the robustness of Dino features for generalization. VIT-H14 showed improved results under the MAS19 split, with a ZSL ACC of 58.07% and an H-Mean of 53.33%, highlighting its capacity to leverage semantic attributes for unseen class classification. CLIP-H14 exhibited strong performance, particularly under the CCS split, achieving an H-Mean of 59.7% and an unseen accuracy of 45.557%, suggesting that multimodal architectures benefit from structured class-splitting methods. Finally, we outline that MambaV-B1K's performance shows the lowest degree of generalization: it achieved its highest ZSL accuracy (51.655%) and H-Mean (50.790%) under the *nosplit* and *mas19*. Moreover, performance dropped notably on *ccs* splits, indicating that the Mamba backbone may be more sensitive to attribute sparsity under the CLSWGAN.

Among all evaluated splits, attribute-based strategies (PAS19, MAS19) generally improved unseen accuracy across models, while class-based splits (CCS, GCS). The results indicate that CLSWGAN exhibits higher sensitivity to the choice of splitting strategy.

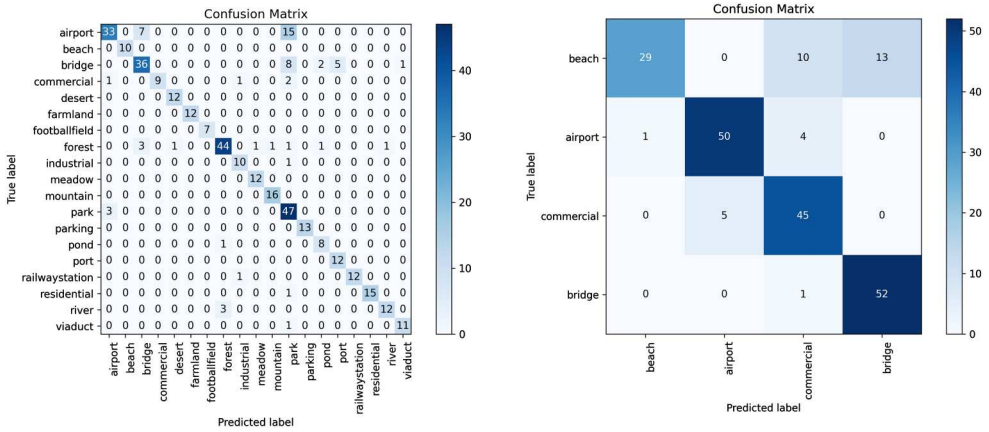
Table 4 highlights the performance of different feature extractors under the TFVAEGAN generative ZSL method across various attribute and class-based splits. Among the evaluated models, DINOv2-L14 consistently achieved the highest performance, with a peak ZSL ACC of 89.471% and an H-Mean of 84.136% under the MAS19 split, demonstrating its superior generalization capabilities. Similarly, the PAS19 split yielded high accuracy across different backbones, with DINOv2-L14 attaining the best-unseen class accuracy (U-

ACC: 76.562%) and overall harmonic mean (H-Mean: 84.302%). Among the vision transformer models, CLIP-H14 exhibited strong generalization, particularly under the PAS19 split, achieving an H-Mean of 74.504% and unseen accuracy of 63.809%, reinforcing the effectiveness of multimodal feature extraction. VIT-H14 performed best with PAS19, attaining a ZSL ACC of 68.603% and an H-Mean of 71.196%, further supporting the role of attribute-based splits in enhancing model robustness. MambaV-B-1 K performs competitively under the TFVAEGAN approach, with its highest ZSL accuracy (64.986%) and harmonic mean (64.003%) achieved under the PAS19 split. The model also exhibits strong unseen accuracy (63.248%), indicating robustness when attribute information is well-distributed. Conversely, ResNet101, the CNN-based model, showed lower performance compared to transformer-based extractors but still benefited from the PAS19 split, achieving a peak ZSL ACC of 66.364% and H-Mean of 59.594%, indicating its relative effectiveness in structured attribute-based learning.

The results suggest that attribute-driven splitting strategies (PAS19, MAS19) consistently outperform class-based splits (CCS, GCS), demonstrating the critical role of semantic information in ZSL for remote sensing. Overall, the study confirms that TFVAEGAN combined with DINOv2-L14 under attribute-based splits provides the most reliable and interpretable ZSL framework for remote sensing applications.

Different performance variations of both ZSL and GZSL were observed to conclude the analysis across the different generative methods. DINOv2-L14 consistently outperforms other feature extractors, achieving the highest GZSL H-Mean across all methods. In particular, CLIP-H14 follows closely, particularly excelling in seen-class accuracy (S-ACC), while VIT-H14 and ResNet101 underperform, especially in unseen-class generalization (U-ACC). Across generative models, TFVAEGAN demonstrates superior performance, significantly outperforming both FREE and CLSWGAN in GZSL, particularly in U-ACC, highlighting its capability to generate more informative representations for unseen categories. DINOv2-L14 achieves the best H-Mean (84.30) under TFVAEGAN, reinforcing the advantage of modern vision transformers over CNN-based models like ResNet101, which consistently yields lower performance across all methods. Regarding the choice of class/attribute split, the 'pas19' emerges as with a high and consistent H-Mean values, likely due to its balanced distribution of seen and unseen classes. 'mas19' also performs well, while 'gcs' improves unseen-class accuracy compared to 'ccs', suggesting that grouping strategies significantly impact generalization.

The confusion matrix for the GZSL evaluation model, which considered both unseen and seen classes (Figure 4a), demonstrates strong classification performance for well defined categories such as airport, bridge, forest, park, and residential, indicating a robust attribute selection. However, classes like bridge (unseen) and park (seen class) exhibit a higher rate of misclassification: those are mostly confused with airport, which shares some common attributes. Considering now the confusion matrix for the ZSL evaluation model, which considered only unseen classes (Figure 4b), we can outline an overall good discriminative performance. However, confusion was exhibited among unseen classes with similar attributes and features, like



(a) GZSL Evaluation Confusion Matrix (Seen and Unseen classes).

(b) ZSL evaluation Confusion Matrix (Unseen classes).

Figure 4. Comparison of ZSL and GZSL Confusion Matrices.

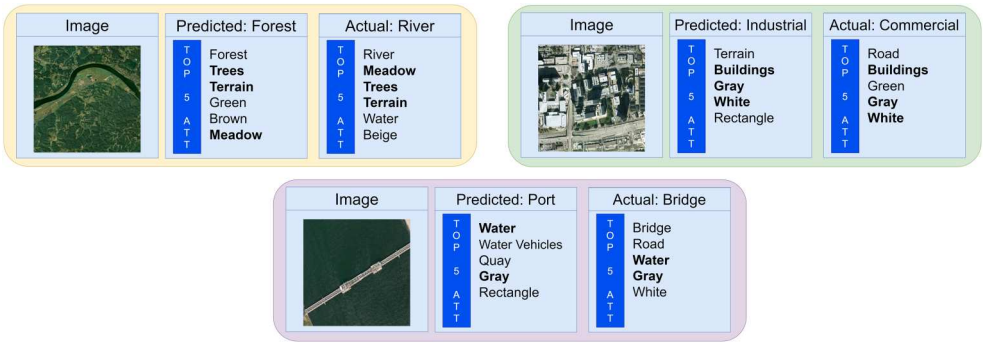


Figure 5. Qualitative examples of wrongly predicted images according to three different classes and their Top-5 attributes.

commercial, airport, and beach. We then performed a qualitative analysis of the PAS model, highlighting key factors related to attribute selection and classification errors for GZSL settings. Those results are reported in Figure 5.

Misclassifications between classes indicate that the PAS-selected attributes, while effective in the majority of cases, do not always provide sufficient discriminative power when categories share attributes and visual characteristics. For instance, the misclassification of a River with the class Forest (top-left, Figure 5), could arise from a high weight, related to Tree and Terrain attributes, which overcomes the presence of River-related attributes (e.g. water, gray, meadow). Similarly, the confusion between commercial and industrial areas (top-right, Figure 5) is likely due to shared attributes such as the presence of buildings and gray/white tones, which do not fully capture the spatial and functional differences between these classes. Additionally, confusion between Port and Bridge classes may arise with PAS prioritizing attributes related to water and gray, leading to difficulties in distinguishing areas where a bridge is in the middle of the water, for example. These findings suggest that our best model has some biases that cause it to be over-reliant on certain visual attributes.

5.5. Ablation study

To dissect the contribution of the different components, we here perform an ablation analysis by comparing the worst-performing configuration to the best one, changing one factor of the framework at a time: Backbone, Generative ZSL method, and Class/attribute split. Results are reported in Table 5.

The 'baseline' configuration amounts to ResNet101 backbone, the FREE generative model, and the gcs class split and achieves a ZSL accuracy of 49.41% and a GZSL H-Mean of 48.61%. Replacing ResNet101 with DINOv2-L14 backbone results in a performance increase: ZSL accuracy improves by +7.2%, while the GZSL H-Mean increases by approximately +14%. Introducing TFVAEGAN, we obtained an additional +5.1% gain in ZSL accuracy and +2.8% in H-Mean. The biggest improvement is observed when modifying the class split strategy: employing pas19 leads to an increase of +22% in ZSL accuracy and +19% in H-Mean.

5.6. ZSL hard setting analysis

Given all the results previously obtained, we here furnish a final analysis of how the most performant backbones (according to backbone type) and splitting strategies perform in a harder ZSL setting. To this date, we create a novel setting where the class split corresponds to 13 seen classes and 6 unseen (70%/30%), while comparing it against our previous baseline of 15 seen classes and 4 unseen (80%/20%). Results are reported in Table 6.

Table 5. Ablation Study from Worst to Best Configuration.

Backbone	ZSL Method	Split	ZSL ACC	GZLS (H-Mean)
(✗) ResNet101	(✗) FREE	(✗) gcs	49.41	48.614
(✓) DINOv2-L14	(✗) FREE	(✗) gcs	56.615	62.427
(✓) DINOv2-L14	(✓) TFVAEGAN	(✗) gcs	61.721	65.225
(✓) DINOv2-L14	(✓) TFVAEGAN	(✓) pas19	**83.698**	**84.302**

Table 6. Comparison of ZSL Accuracy across best generative ZSL method (TFVAEGAN) and backbones, comparing class split 15-4 against 13-6. FE stands for Feature Extractor, GM stands for Generative Method, S-ACC, and U-ACC, respectively, for seen and unseen classes' Accuracies. H-mean amounts to the harmonic mean between the S-ACC and U-ACC.

				Metriics (%)		
FE	Split Class	Type Split	ZSL ACC	GZSL S-ACC	GZSL U-ACC	GZSL H-Mean
	nosplit	13-6	43.402	68.67	28.316	40.098
	nosplit	15-4	**59.662**	**70.439**	**42.325**	**52.877**
	ccs	13-6	36.631	55.512	23.887	33.401
	ccs	15-4	**51.177**	**70.124**	**31.286**	**43.268**
ResNet101	gcs	13-6	34.417	48.293	22.182	30.401
	gcs	15-4	**44.049**	**59.324**	**43.467**	**50.172**
	pas19	13-6	47.26	59.637	33.417	42.833
	pas19	15-4	**66.364**	**83.91**	**46.204**	**59.594**
	mas19	13-6	43.117	63.738	26.783	37.717
	mas19	15-4	**61.421**	**80.765**	**41.743**	**55.039**
	nosplit	13-6	55.367	**90.976**	45.209	60.402
	nosplit	15-4	**69.893**	87.664	**50.684**	**64.232**
	ccs	13-6	53.247	**94.592**	51.901	67.026
	ccs	15-4	**77.292**	94.254	**57.499**	**71.425**
CLIP-H14	gcs	13-6	41.208	**90.156**	40.769	56.147
	gcs	15-4	**61.074**	87.569	**47.656**	**61.722**
	pas19	13-6	59.422	89.408	50.395	64.458
	pas19	15-4	**71.635**	**89.507**	**63.809**	**74.504**
	mas19	13-6	66.899	92.489	53.909	68.115
	mas19	15-4	**77.305**	**93.521**	**58.388**	**71.892**
	nosplit	13-6	70.093	89.067	61.968	73.087
	nosplit	15-4	**80.093**	**94.331**	**67.016**	**78.361**
	ccs	13-6	57.351	**97.485**	44.135	60.761
	ccs	15-4	**63.478**	94.464	**57.008**	**71.105**
DINOv2-L14	gcs	13-6	42.066	**96.908**	36.717	53.256
	gcs	15-4	**61.721**	91.693	**50.615**	**65.225**
	pas19	13-6	58.924	**95.158**	56.051	70.548
	pas19	15-4	**83.698**	93.784	**76.562**	**84.302**
	mas19	13-6	63.211	93.291	59.551	72.697
	mas19	15-4	**89.471**	**96.184**	**74.77**	**84.136**
	nosplit	13-6	49.915	75.328	38.665	51.101
	nosplit	15-4	**62.269**	**56.064**	**59.688**	**57.820**
	ccs	13-6	52.642	60.840	42.067	49.741
	ccs	15-4	**64.776**	**72.914**	**45.312**	**55.891**
MambaV-B-1K	gcs	13-6	32.996	72.397	23.000	34.910
	gcs	15-4	**47.410**	**76.457**	**38.656**	**51.350**
	pas19	13-6	47.756	58.216	49.170	53.312
	pas19	15-4	**68.986**	**64.775**	**63.248**	**64.003**
	mas19	13-6	52.632	68.012	40.767	50.977
	mas19	15-4	**61.141**	**78.700**	**39.320**	**52.440**

According to the reported results, several key observations emerge. First, a consistent drop in accuracy (Performance Gap) is observed across all feature extractors and splitting strategies when moving from the 15-4 to the 13-6 split. This is expected, as fewer seen classes reduce the information available for transferring knowledge to unseen categories. The decline is particularly notable in U-ACC and H-mean. In the GZSL setting, S-ACC remains high across models, yet this often comes at the expense of U-ACC, reflecting the known bias towards seen categories. Among the models, DINOv2-L14 and CLIP-H14 achieve the best H-mean values and show a significantly stronger capacity for generalizing to unseen classes compared to Resnet101. Similarly, MambaV-B-1 K exhibits a performance drop when moving from the 15-4 to the 13-6 split, especially in U-ACC and H-Mean, highlighting its sensitivity to reduced class supervision. Again, attribute-based splits seem to be more robust than the class-based ones, with a lower performance drop. Furthermore, the magnitude of the Performance Gap varies across feature extractors, with ResNet101 and MambaV-B-1 K exhibiting larger drops than CLIP-H14 and DINOv2-L14, reinforcing the greater robustness of the latter two in the ZSL remote sensing context. Finally, attribute-based splitting strategies such as MAS and PAS prove more robust, whereas CCS and GCS cause sharper declines in U-ACC, suggesting that these constraints may impair model generalization.

6. Discussion and conclusions

Our framework introduces a novel dataset and methodological approach to evaluate generative attribute-based ZSL in RS. Unlike traditional ZSL methods that often rely on fixed benchmark splits, our framework systematically assesses performance across multiple attribute and class splits. Moreover, we also integrated a modular approach to evaluate different visual backbones, to understand which one provides the best features to be fed to the considered generative methods.

We conducted extensive experiments using various visual backbones and generative ZSL models across different split types.

In the following, we introduced a list of the main discoveries and observations made during our experimental setting:

- Considering UB analysis, used to have a controlled environment to isolate the effect of feature extractors and generative paradigms, results outline that Transformer-based architectures (e.g. ViT-B16, DINOv2-B14, DINOv2-L14) generally outperform CNN backbones and Mamba-based ones in the CZSL task, with DINOv2-B14 achieving the highest ZSL accuracy and GZSL accuracy;
- Specifically considering the adopted ZSL generative approach, TFVAEGAN consistently outperforms FREE and CLSWGAN in terms of both ZSL and GZSL, confirming TFVAEGAN's robustness in generating semantically rich representations that aim to better generalize to unseen classes;
- Additionally, our robustness analysis demonstrated that attribute-based splits (MAS, PAS) consistently yielded the highest performance, suggesting that semantic-based feature selection enhances knowledge transfer to unseen categories. Conversely, class-based splits (CCS, GCS) exhibited higher variability, with notable performance gaps arising due to semantic loss and class imbalance. In particular, PAS emerged as the most performant one.
- The ablation reveals that each architectural factor contributes additively to model performance. The feature extractor emerges as the dominant factor initially, while the class split exhibits the greatest absolute impact in the final configuration. These findings underscore the necessity of carefully controlling for all design choices in our ZSL framework, suggesting that optimal performance arises from the composition of robust feature extractors, generative models, and split strategies. This aligns with results observed in (Rossi et al. 2024).
- Finally, comparing a classical ZSL setting (with only 20% of unseen classes) with respect to a harder and real-world one (more than 30% of unseen classes), we analyzed which combination of feature extractor, generative model, and split type provided the less performance drop (and so higher degree of generalizability). The consistent performance observed with attribute-based splits and transformer-based backbones suggested that incorporating rich semantic information effectively enhances a model's ability to generalize to unseen classes. This finding aligns with recent studies emphasizing the importance of semantic alignment in ZSL (Liu et al. 2023).

Analyzing our initial research question (to what extent can ZSL methods generalize across different attribute splits in RS, and how does the choice of split and visual back *bone affect the model's performance?)*, we can confirm that attribute-based splitting, paired with express generative ZSL methods and backbones significantly influence the final generalization performance, providing a reasonable robustness capability also in hard ZSL setting.

With these learned lessons, we can design novel ways to re-adapt our framework to be applied to different kinds of use cases, also to face challenges like more complex, multi-source, and heterogeneous remote sensing data as the ones detailed in (Liu et al. 2024). For example, involving data heterogeneity and scalability, RS pipelines must operate across diverse geographic regions and sensor designs. Our framework could be easily deployed to include more classes and attribute to catch a higherer variance of both geographical places and capturing sensors. On this line, we envision feeding attributes into novel open-vocabulary, multimodal foundation models (e.g. CLIP-style frameworks extended to RS modalities [Liu et al. 2024]) to extract semantic embeddings even from heterogeneous sources. These embeddings could then be used within our attribute-based generative ZSL pipeline, enabling ZSL classification, but also providing informed features to support other tasks (like image segmentation pipelines [Liu et al. 2025]).

Despite the outlined results and possibility of extension, our work presents some limitations. First, while our evaluation strategy offers a comprehensive assessment of model performance, it may increase computational complexity due to the need for multiple training and testing cycles across different splits. Moreover, we adopted feature extractors from classical CNN to ViT architectures. This amounts to a limitation, considering the very recent success of Multimodal Large Language models in this field (Zhang et al. 2024). Finally, our reliance on attribute-based semantic spaces, which are manually engineered and collected, may introduce human biases that could affect model generalization. Future research should explore automated methods for semantic space construction to mitigate these biases. Additionally, while our framework focuses on inductive ZSL settings, extending the evaluation to transductive scenarios, where unlabeled data from unseen classes is available during training, could provide further insights. For these reasons, we plan to expand our experimental setting, including not only a broader attribute semantic space, but also re-label more complex datasets, integrating into our comparative benchmark additional generative ZSL methods.

To conclude, this work establishes a foundation for more reliable and interpretable ZSL frameworks in RS, paving the way for practical deployment in real-world Earth Observation tasks. Future research should focus on integrating multimodal embeddings from geospatial metadata, enhancing domain adaptation techniques, and exploring self-supervised pretraining strategies to further improve ZSL generalization in RS applications.

Note

1. https://aws.amazon.com/pm/sagemaker/?nc1=h_ls

Disclosure statement

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

Funding

This work was funded by Grant Agreement N. 101190021 AI4COPSEC – CUP I33C24005360005 - "Security enhancement through heterogeneous data fusion and improved AI/ML-powered Copernicus maritime and border surveillance services."

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Aleissaee, A. A., A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G. S. Xia, and F. S. Khan. 2023. "Transformers in Remote Sensing: A Survey." *Remote Sensing* 15 (7): 1860. https://doi.org/10.3390/rs15071860.

Al Rahhal, M. M., Y. Bazi, H. Elgibreen, and M. Zuair. 2023. "*Vision-language Models for Zero-Shot Classification of Remote Sensing Images*." Applied Sciences 13 (22): 12462. https://doi.org/10.3390/app132212462.

Bucher, M., S. Herbin, and F. Jurie. 2016. "Improving Semantic Embedding Consistency by Metric Learning for Zero-Shot Classiffication." In *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14. Springer, 730–746.

Chen, S., W. Wang, B. Xia, Q. Peng, X. You, F. Zheng, and L. Shao. 2021. Free: Feature Refinement for Generalized Zero-Shot Learning." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 122–131.

Damalla, R., R. Datla, C. Vishnu, and C. K. Mohan. 2023. "Self-supervised Embedding for Generalized Zero-Shot Learning in Remote Sensing Scene Classification." *Journal of Applied Remote Sensing* 17 (3): 032405–032405. https://doi.org/10.1117/1.JRS.17.032405.

Demirel, B., R. Gokberk Cinbis, and N. Ikizler-Cinbis. 2017. "Attributes2classname: A Discriminative Model for Attribute-Based Unsupervised Zero-Shot Learning. In *Proceedings of the IEEE International Conference On Computer Vision*, 1232–1241.

Ding, Z., M. Shao, and Y. Fu. 2017. "Low-rank Embedded Ensemble Semantic Dictionary for Zero-Shot Learning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2050–2058.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, et al. 2021. "An Image is Worth 16(16 Words: Transformers for Image Recognition at Scale". *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2010.11929.

Elhoseiny, M., B. Saleh, and A. Elgammal. 2013. "Write a Classifier: Zero-shot Learning using Purely Textual Descriptions." In *Proceedings of the IEEE International Conference on Computer Vision*, 2584–2591.

Fu, Z., T. Xiang, E. Kodirov, and S. Gong. 2017. "Zero-shot Learning on Semantic Class Prototype Graph." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (8): 2009–2022. https://doi.org/10.1109/TPAMI.2017.2737007.

Hatamizadeh, A., and J. Kautz. 2025. "Mambavision: A Hybrid Mamba-Transformer Vision Backbone." In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25261–25270.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Kenton, J. D. M. W. C., and L. K. Toutanova. 2019. "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of naacL-HLT, Vol. 1*. Minneapolis, Minnesota, 2.

Lampert, C. H., H. Nickisch, and S. Harmeling. 2009. "Learning to Detect Unseen Object Classes by between-class Attribute Transfer." in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2009.5206594.

Lei Ba, J., K. Swersky, and S. Fidler. 2015. "Predicting Deep Zero-shot Convolutional Neural Networks using Textual Descriptions." In *Proceedings of the IEEE International Conference on Computer Vision*, 4247–4255.

Li, A., Z. Lu, L. Wang, T. Xiang, and J. R. Wen. 2017. "Zero-shot Scene Classification for High Spatial Resolution Remote Sensing Images." *IEEE Transactions on Geoscience and Remote Sensing* 55 (7): 4157–4167. https://doi.org/10.1109/TGRS.2017.2689071.

Li, X., C. Wen, Y. Hu, and N. Zhou. 2023. "Empirical Evidence regarding Few-Shot Learning for Scene Classification in Remote Sensing Images." *Applied Sciences* 14:10776. https://www.mdpi.com/2076-3417/14/23/10776.

Li, Z., D. Zhang, Y. Wang, D. Lin, and J. Zhang. 2022. "Generative Adversarial Networks for Zero-Shot Remote Sensing Scene Classification." *Applied Sciences* 12 (8): 3760. https://doi.org/10.3390/app12083760.

Liu, C., S. Ma, Z. Li, W. Yang, and Z. Han. 2023. "Multi-level Cross-modal Feature Alignment via Contrastive Learning towards Zero-shot Classification of Remote Sensing Image Scenes." arXiv preprint arXiv:2306.06066.

Liu, C., Y. Sun, Y. Xu, Z. Sun, X. Zhang, L. Lei, and G. Kuang. 2024. "A Review of Optical and sar Image Deep Feature Fusion in Semantic Segmentation." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17: 12910–12930. https://doi.org/10.1109/JSTARS.2024.3424831.

Liu, C., Y. Sun, X. Zhang, Y. Xu, L. Lei, and G. Kuang. 2025. "Oshfnet: A Heterogeneous Dual-Branch Dynamic Fusion Network of Optical and sar Images for Land use Classification." *International Journal of Applied Earth Observation and Geoinformation* 141:104609. https://doi.org/10.1016/j.jag.2025.104609

Liu, F., D Chen, Z Guan, X Zhou, J Zhu, Q Ye, L Fu, and J Zhou 2024. "Remoteclip: A Vision Language Foundation Model for Remote Sensing." *IEEE Transactions on Geoscience and Remote Sensing* 62: 5622216. https://doi.org/10.1109/TGRS.2024.3390838.

Liu, S., and M. Ozay. 2023. "Task Guided Representation Learning Using Compositional Models for Zero-Shot Domain Adaptation." *Neural Networks* 165:370–380. https://doi.org/10.1016/j.neunet.2023.05.030

Ma, Y., E. Cambria, and S. Gao. 2016. "Label Embedding for Zero-shot Fine-grained Named Entity Typing." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 171–180.

Narayan, S., A. Gupta, F. S. Khan, C. G. Snoek, and L. Shao. 2020. "Latent Embedding Feedback and Discriminative Features for Zero-shot Classification." In *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. Springer, pp. 479–495.

Norouzi, M., T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. 2013. "Zero-shot Learning by Convex Combination of Semantic Embeddings." arXiv preprint arXiv:1312.5650.

Oquab, M., T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, et al. 2023. "Dinov2: Learning Robust Visual Features without Supervision." arXiv preprint arXiv:2304.07193.

Paheding, S., A. Saleem, M. F. H. Siddiqui, N. Rawashdeh, A. Essa, and A. A. Reyes. 2024. "Advancing Horizons in Remote Sensing: A Comprehensive Survey of Deep Learning Models and Applications in Image Classification and Beyond." *Neural Computing and Applications* 36 (27): 16727–16767. https://doi.org/10.1007/s00521-024-10165-7

Palatucci, M., D. Pomerleau, G. E. Hinton, and T. M. Mitchell. 2009. "Zero-shot Learning with Semantic Output Codes." *Advances in Neural Information Processing Systems* 22: 1410–1418.

Pan, C., J. Huang, J. Hao, and J. Gong. 2020. "Towards Zero-Shot Learning Generalization via a Cosine Distance Loss." *Neurocomputing* 381:167–176. https://doi.org/10.1016/j.neucom.2019.11.011

Pourpanah, F., M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X. Z. Wang, and Q. J. Wu. 2022. "A Review of Generalized Zero-Shot Learning Methods." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45:4051–4070.

Quan, J., C. Wu, H. Wang, and Z. Wang. 2018. "Structural Alignment Based Zero-Shot Classification for Remote Sensing Scenes." In *2018 IEEE International Conference on Electronics and Communication Engineering (ICECE)*, 17–21. https://doi.org/10.1109/ICECOME.2018.8645056.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, et al. 2021. "Learning Transferable Visual Models From Natural Language Supervision." In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html.

Rahman, S., S. Khan, and F. Porikli. 2018. "A Unified Approach for Conventional Zero-Shot, Generalized Zero-Shot, and Few-Shot Learning." *IEEE Transactions on Image Processing* 27 (11): 5652–5667. https://doi.org/10.1109/TIP.2018.2861573.

Rambabu, D., G. Swetha, R. Datla, V. Chalavadi, and C. K. Mohan. 2024. "RSZero-CSAT: Zero-Shot Scene Classification in Remote Sensing Imagery Using a Cross Semantic Attribute-Guided Transformer." In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN60899.2024.10650858.

Reddy, Y., P. Viswanath, and B. E. Reddy. 2018. "Semi-supervised Learning: A Brief Review." *Int. J. Eng. Technol* 7 (1.8): 81–85. https://doi.org/10.14419/ijet.v7i1.8.9977.

Rossi, L., M. C. Fiorentino, A. Mancini, M. Paolanti, R. Rosati, and P. Zingaretti. 2024. "Generalizability and Robustness Evaluation of Attribute-Based Zero-Shot Learning." *Neural Networks* 175:106278. https://doi.org/10.1016/j.neunet.2024.106278

Sammon, J. W. 1969. "A Nonlinear Mapping for Data Structure Analysis." *IEEE Transactions on Computers* C-18 (5): 401–409. https://doi.org/10.1109/T-C.1969.222678.

Socher, R., M. Ganjoo, C. D. Manning, and A. Ng. 2013. "Zero-shot Learning through Cross-Modal Transfer. Advances in Neural Information Processing Systems 26 (NIPS 2013) .

Song, X., H. Zeng, S. Zhang, L. Herranz, and S. Jiang. 2020. "Generalized Zero-shot Learning with Multi-source Semantic Embeddings for Scene Recognition." In *Proceedings of the 28th ACM International Conference on Multimedia*, 3976–3985.

Sumbul, G., R. G. Cinbis, and S. Aksoy. 2017. "Fine-grained Object Recognition and Zero-Shot Learning in Remote Sensing Imagery." *IEEE Transactions on Geoscience and Remote Sensing* 56 (2): 770–779. https://doi.org/10.1109/TGRS.2017.2754648.

Sun, X., J. Gu, and H. Sun. 2021. "Research Progress of Zero-Shot Learning." *Applied Intelligence* 51 (6): 3600–3614. https://doi.org/10.1007/s10489-020-02075-7

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. "Rethinking the Inception Architecture for Computer Vision." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826.

Tan, C., F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. 2018. "A Survey on Deep Transfer Learning." In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks*, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27. Springer, pp. 270–279.

Tan, X., B. Xi, J. Li, T. Zheng, Y. Li, C. Xue, and J. Chanussot. 2024. "Review of Zero-Shot Remote Sensing Image Scene Classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17: 11274–11289. 10.1109/JSTARS.2024.3410995.

Wang, C., G. Peng, and B. D. Baets. 2021. "A Distance-Constrained Semantic Autoencoder for Zero-Shot Remote Sensing Scene Classification." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14: 12545–12556. https://doi.org/10.1109/JSTARS.2021.3132189.

Wang, D., Y. Li, Y. Lin, and Y. Zhuang. 2016. "Relational Knowledge Transfer for Zero-shot Learning." In *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30*.

Wang, W., V. W. Zheng, H. Yu, and C. Miao. 2019. "A Survey of Zero-Shot Learning: Settings, Methods, and Applications." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10:1–37.

Wang, Y., Q. Yao, J. T. Kwok, and L. M. Ni. 2020. "Generalizing from a Few Examples: A Survey on Few-Shot Learning." *ACM Computing Surveys (Csur)* 53:1–34.

Xia, G. S., W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maitre. 2010. "Structural High-resolution Satellite Image Indexing. ISPRS TC VII Symposium - 100 Years ISPRS. https://hal.science/hal-00458685.

Xian, Y., C. H. Lampert, B. Schiele, and Z. Akata. 2018. "Zero-shot Learning – a Comprehensive Evaluation of the Good, the bad and the Ugly." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (9): 2251–2265. https://doi.org/10.1109/TPAMI.2018.2857768.

Xian, Y., T. Lorenz, B. Schiele, and Z. Akata. 2018. "Feature Generating Networks for Zero-Shot Learning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5542–5551.

Xu, W., J. Wang, Z. Wei, M. Peng, and Y. Wu. 2023. "Deep Semantic-Visual Alignment for Zero-Shot Remote Sensing Image Scene Classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 198:140–152. https://doi.org/10.1016/j.isprsjprs.2023.02.012

Ye, H. J., H. Hu, and D. C. Zhan. 2021. "Learning Adaptive Classifiers Synthesis for Generalized Few-Shot Learning." *International Journal of Computer Vision* 129 (6): 1930–1953. https://doi.org/10.1007/s11263-020-01381-4

Zhang, L., T. Xiang, and S. Gong. 2017. "Learning a Deep Embedding Model for Zero-Shot Learning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021–2030.

Zhang, W., M. Cai, T. Zhang, Y. Zhuang, and X. Mao. 2024. "Earthgpt: A Universal Multi-modal Large Language Model for Multi-sensor Image Comprehension in Remote Sensing Domain." *IEEE Transactions on Geoscience and Remote Sensing* 62: 1–20.