remote sensing

MDPI

Article

GRADE: A Generalization Robustness Assessment via Distributional Evaluation for Remote Sensing Object Detection

Decheng Wang [1,2], Yi Zhang [1], Baocun Bai [3], Xiao Yu [1,*], Xiangbo Shu [4] and Yimian Dai [2]

[1] Beijing Institute of Tracking and Telecommunication Technology, Beijing 100094, China
[2] Visual Computing and Intelligent Perception (VCIP) Lab, College of Computer Science, Nankai University, Tianjin 300350, China; yimian.dai@nankai.edu.cn
[3] Equipment Project Management Center, Beijing 100094, China
[4] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; shuxb@njust.edu.cn
* Correspondence: 2120240703@mail.nankai.edu.cn

Highlights

What are the main findings?

- We propose the **GRADE (Generalization Robustness Assessment via Distributional Evaluation) framework,** an evaluation paradigm that systematically links model performance degradation (the effect) to quantifiable data distribution shifts (the cause), moving beyond traditional "black-box" metrics like mAP.
- The framework introduces hierarchical divergence metrics (**Scene-level Fréchet Inception Distance (FID)** and **Instance-level FID**) to create an adaptively weighted **Generalization Score (GS)** that demonstrates high fidelity to empirical model rankings across diverse remote sensing datasets.

What are the implications of the main finding?

- The GRADE framework provides an analytical tool that allows researchers to **attribute generalization failure to specific sources** (e.g., failure to adapt to new background contexts vs. novel object appearances), guiding targeted model improvement.
- This work establishes a **standardized and interpretable protocol** for comparing the cross-domain robustness of object detectors, enabling fairer and more reliable model selection for real-world deployment scenarios.

Abstract

The performance of remote sensing object detectors often degrades severely when deployed in new operational environments due to covariate shift in the data distribution. Existing evaluation paradigms, which primarily rely on aggregate performance metrics such as mAP, generally lack the analytical depth to provide insights into the mechanisms behind such generalization failures. To fill this critical gap, we propose the **GRADE (Generalization Robustness Assessment via Distributional Evaluation)** framework, a multi-dimensional, systematic methodology for assessing model robustness. The framework quantifies shifts in background context and object-centric features through a hierarchical analysis of distributional divergence, utilizing **Scene-level Fréchet Inception Distance (FID)** and **Instance-level FID**, respectively. These divergence measures are systematically integrated with a standardized performance decay metric to form a unified, adaptively weighted **Generalization Score (GS)**. This composite score serves not only as an evaluation tool but also as a powerful analytical tool, enabling the fine-grained attribution of performance loss to specific sources of domain shift—whether originating from scene variations or anomalies in object appearance. Compared to conventional single-dimensional evaluation methods, the

check for updates

Academic Editor: Farid Melgani

Received: 19 October 2025
Revised: 14 November 2025
Accepted: 18 November 2025
Published: 20 November 2025

Citation: Wang, D.; Zhang, Y.; Bai, B.; Yu, X.; Shu, X.; Dai, Y. GRADE: A Generalization Robustness Assessment via Distributional Evaluation for Remote Sensing Object Detection. *Remote Sens.* **2025**, *17*, 3771. https://doi.org/10.3390/rs17223771

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

GRADE framework offers enhanced interpretability, a standardized evaluation protocol, and reliable cross-model comparability, establishing a principled theoretical foundation for cross-domain generalization assessment. Extensive empirical validation on six mainstream remote sensing benchmark datasets and multiple state-of-the-art detection models demonstrates that the model rankings produced by the GRADE framework exhibit high fidelity to real-world performance, thereby effectively quantifying and explaining the cross-domain generalization penalty.

Keywords: generalization robustness; distributional evaluation; covariate shift; remote sensing; object detection

1. Introduction

In the contemporary landscape of Earth observation, deep learning has emerged as a transformative force, catalyzing unprecedented breakthroughs in the automated interpretation of remote sensing imagery. This technological paradigm shift has demonstrated exceptional proficiency in a spectrum of crucial tasks, ranging from fine-grained object detection [1,2] to large-scale land-cover classification [3]. These advancements are not merely academic; they constitute the operational backbone for critical real-world applications, including national land resource management, rapid disaster response, environmental monitoring, and strategic reconnaissance [4]. The efficacy of these deep learning models, however, is predicated on a foundational, yet fragile, assumption: that the training (source) and testing (target) data are **independent and identically distributed (i.i.d.)**.

In practice, this i.i.d. assumption is frequently and profoundly violated when models are deployed in the wild. Remote sensing data is inherently dynamic and heterogeneous. Consequently, when a pre-trained detection model confronts novel operational environments, its performance can degrade substantially—and often unpredictably. This degradation is driven by systemic **covariate shifts** in the data distribution, which can arise from a multitude of factors: variations in atmospheric conditions (e.g., haze, cloud cover), changes in illumination due to diurnal or seasonal cycles, different sensor specifications (e.g., spectral bands, radiometric resolution), and diverse geographical topographies. For instance, a model meticulously trained on high-resolution aerial imagery of temperate urban landscapes may fail catastrophically when applied to satellite imagery of tropical regions, where phenomena like dense foliage, unique architectural styles, and distinct shadow patterns diverge sharply from its learned experience.

To counteract the performance decay stemming from such distribution shifts, the research community has explored various mitigation strategies. These include sophisticated data augmentation techniques, transfer learning from pre-trained foundation models, and more explicit domain adaptation methodologies [5–7]. Despite these concerted efforts, the efficacy of such methods remains constrained, particularly when models are deployed in entirely unseen domains for which no prior data, labeled or unlabeled, is available. A more profound and systemic issue is that the majority of research has gravitated towards proposing remedial techniques rather than investing in a foundational, systematic analysis of *why* cross-domain performance degrades. This critical gap in understanding stems from the conspicuous lack of a systematic theoretical framework for evaluating generalization. The absence of such a framework impedes the scientific, impartial, and reproducible quantification of a model's capacity to generalize across the diverse and unpredictable operational scenarios it will inevitably encounter.

The rigorous evaluation of generalization in the specific context of remote sensing object detection remains a nascent, yet critically important, field. Within the optical remote sensing modality, distributional divergence stands as the foremost barrier to achieving robust and reliable generalization. This is not a theoretical concern but an empirical reality. For example, state-of-the-art models trained on the widely-used DOTA dataset [8] consistently exhibit marked performance drops when evaluated on other benchmark datasets like FAIR1M [9] or DroneVehicle [10]. This performance degradation is directly attributable to the inherent heterogeneity across these datasets, which encompasses a wide array of variations in imaging resolution, sensor viewing angles, environmental conditions, object scales and densities, background complexity, and sensor-specific noise profiles. Thus, addressing the challenge of covariate shift is not merely an incremental improvement but represents the central challenge in advancing the frontier of generalization research.

Furthermore, conventional evaluation practices in the field predominantly rely on aggregate, task-specific metrics such as mean Average Precision (mAP) to compare and rank model performance [11,12]. While these metrics are indispensable for indicating overall detection quality within a given domain, they function as a ”black box,” offering no insight into the underlying causal mechanisms of performance failure when significant distribution shifts occur. A high mAP score provides confidence, but a low score offers no diagnosis. Although the broader computer vision community, particularly in image classification, has benefited from the development of comprehensive diagnostic benchmarks like PACS [13] and WILDS [14], the object detection field lacks similarly systematic and analytical studies. Commendable efforts like the COCO-O dataset [15] have been valuable in highlighting the severity of the problem. However, research in this vein often concludes with the descriptive act of reporting performance drops rather than providing a prescriptive, diagnostic analysis. A framework capable of characterizing the specific contributions of various heterogeneity factors—such as background context versus object appearance—to a model's degradation is conspicuously absent from the literature.

Therefore, a pivotal and pressing open question remains: *how can we systematically quantify distributional divergence and rigorously link it to observable performance degradation in order to forge reliable, interpretable, and actionable criteria for model generalization in complex, multi-domain remote sensing contexts?*

The ambition of this study is not merely to engineer another novel detection model that claims superior generalization. Instead, our primary objective is to construct a comprehensive and principled evaluation framework. This framework is architected to address a core scientific question that transcends any single model architecture: **what constitutes a principled, equitable, and comprehensive evaluation of a model's generalization capacity amidst significant and multifaceted distributional shifts?**

To this end, we introduce the **GRADE (Generalization Robustness Assessment via Distributional Evaluation)** framework. GRADE is founded on the dual pillars of data distribution divergence analysis (the "cause") and performance decay measurement (the "effect"). To quantify distribution shifts with semantic granularity, we employ a bi-level analysis using **Scene-level Fréchet Inception Distance (FID)** to capture shifts in background and environmental context, and **Instance-level FID** to capture shifts in object-centric features. To quantify the corresponding performance loss, we utilize a normalized relative performance drop metric that is more robust than absolute changes in mAP. These two dimensions—distributional divergence and performance decay—are then intelligently fused via an adaptive weighting scheme into a consolidated **Generalization Score (GS)**. This composite score holistically and equitably reflects a model's cross-domain robustness, penalizing models that fail on domains that are significantly different from the source.

This framework methodologically bridges the long-standing gap between "distribution shift" and "performance decline". It enables not only the quantification of overall degradation but, more importantly, an analysis of its primary sources (e.g., failure to adapt to new backgrounds vs. inability to recognize novel object appearances). By doing so, GRADE provides multi-dimensional interpretability and standardized comparability, establishing a unified and theoretically grounded foundation for evaluating cross-domain generalization.

Our primary contributions are threefold:

1. **A Principled Framework for Generalization Assessment:** We propose the GRADE framework, a novel paradigm that moves beyond conventional i.i.d. evaluation protocols by explicitly integrating distributional divergence analysis with performance degradation modeling. This protocol provides granular, analytical insights into *why* a model's performance declines, not just by how much, attributing failure to specific types of domain shift.
2. **Hierarchical Distributional Divergence Metrics:** We introduce Scene-level and Instance-level FID as specialized metrics to characterize and disentangle distribution shifts at distinct semantic levels (context vs. object). These metrics are integrated with performance measures into an adaptively weighted Generalization Score (GS) that facilitates a fine-grained, interpretable assessment of model vulnerabilities.
3. **Standardized and Empirically Validated Paradigm:** We establish a standardized and reproducible evaluation pipeline and rigorously validate it across multiple public benchmarks and state-of-the-art models. Our extensive empirical results demonstrate that the framework maintains high consistency and stability, providing a reliable and transferable methodology for assessing and comparing cross-domain robustness in a fair and insightful manner.

The remainder of this paper is organized as follows. Section 2 reviews related work in object detection, distributional distance measures, and generalization evaluation. Section 3 provides a detailed methodological exposition of the proposed GRADE framework. Section 4 presents the comprehensive experimental setup, results, and analysis. Finally, Section 5 provides concluding remarks and discusses promising directions for future research.

2. Related Work

Deep learning has undeniably propelled rapid and transformative advancements in remote sensing image analysis, establishing new state-of-the-art performance benchmarks across numerous tasks. Yet, a formidable challenge continues to cast a shadow over these successes: the generalization of models across disparate and unseen domains. The brittleness of models when faced with data distribution shifts remains a significant bottleneck for real-world deployment. Current research pertinent to this challenge can be broadly classified into three interconnected areas: advancements in the architectural design of detection methodologies, the theoretical and practical quantification of data distribution divergence, and the development of more sophisticated generalization evaluation paradigms.

2.1. Remote Sensing Object Detection

The dominant paradigm in remote sensing object detection has undergone a profound transition, moving from classical methods reliant on handcrafted features to the now-ubiquitous end-to-end deep learning architectures. Early techniques, which often depended on meticulously engineered features like SIFT, HOG, or spectral and textural indices, offered valuable initial solutions but demonstrated limited robustness and scalability. Their performance was intrinsically capped by the expressive power of the features themselves, often failing to generalize across variations in illumination, season, and sensor type.

The advent of Convolutional Neural Networks (CNNs) was truly transformative. By learning hierarchical feature representations directly from data, CNN-based detectors overcame many limitations of their predecessors. This era saw the rise of two principal architectural families: two-stage detectors, such as the R-CNN series [16–18], which prioritize accuracy through a region proposal mechanism followed by classification and refinement; and single-stage detectors, like SSD and YOLO [19–21], which achieve real-time performance by formulating detection as a direct regression problem. More recently, the field has been further advanced by Transformer-based models, including DETR and its variants [22]. These models leverage self-attention mechanisms to effectively model long-range dependencies and global context, a capability that has proven particularly advantageous for improving object discrimination in the cluttered and complex scenes typical of remote sensing imagery [23,24]. Concurrently, the fusion of vision and language processing has opened new frontiers in open-set and zero-shot detection through powerful vision-language models [25–27].

While these continuous architectural innovations have significantly boosted **in-domain** accuracy, their performance often degrades sharply and unpredictably in cross-domain scenarios where the foundational i.i.d. assumption is violated. This persistent gap between in-domain excellence and out-of-domain fragility underscores the urgent need for a robust evaluation framework—one that can not only measure performance but also systematically assess, diagnose, and guide the development of truly generalizable models.

2.2. Quantifying Data Distribution Differences

A rigorous, quantitative understanding of model generalization is impossible without the precise measurement of the **distributional divergence** between the source (training) and target (testing) domains. This measurement formalizes the degree to which the i.i.d. assumption is violated. Early statistical methods, such as the Kullback–Leibler (KL) divergence [28], while theoretically elegant, proved computationally intractable in high-dimensional pixel spaces due to the challenge of accurate non-parametric density estimation.

To overcome this, kernel-based methods operating in a Reproducing Kernel Hilbert Space (RKHS), like the Maximum Mean Discrepancy (MMD) [29], offered a more viable alternative by comparing the mean embeddings of distributions without explicit density estimation. More recently, metrics derived from the feature spaces of deep generative models have gained widespread prominence. Among these, the **Fréchet Inception Distance (FID)** [30] has been widely adopted. By modeling the distribution of deep features from a pre-trained Inception network as a multivariate Gaussian, FID computes the Wasserstein-2 distance between the feature distributions of two datasets. Its strong correlation with human perceptual judgment of image quality and diversity has made it a de facto standard. Variants such as the Kernel Inception Distance (KID) [31] have also been proposed to offer improved robustness and unbiasedness for smaller sample sizes.

However, a critical limitation persists: these metrics operate at the data-space level and are fundamentally **task-agnostic**. They quantify the overall statistical difference between two sets of images but do not inherently establish an interpretable, causal link to the performance degradation of a specific downstream task like object detection. A large FID score does not automatically imply a large performance drop for a robust model, nor does a small score guarantee good generalization. This disconnect highlights the necessity for a task-aware evaluation framework that explicitly connects measurable distributional shifts to tangible performance outcomes.

2.3. Generalization Evaluation of Object Detection Models

Most current evaluations of object detector generalization are conducted in an ad hoc manner, primarily depending on custom-built benchmark datasets and traditional performance metrics like mAP. These benchmarks typically fall into two categories. The first includes datasets with **synthetic corruptions**, such as COCO-C [32]. These are valuable for assessing model robustness to a predefined set of low-level perturbations like Gaussian noise, motion blur, or simulated weather effects. However, they often fail to capture the complexity, structure, and semantic nature of real-world distribution shifts encountered in remote sensing.

The second, more relevant category comprises **real-world datasets** collected from diverse geographical domains, sensors, and conditions, such as the COCO-O benchmark [15]. While these benchmarks are invaluable for effectively demonstrating the existence and severity of the generalization problem, the evaluation protocol itself often stops short. The analysis typically relies on standard metrics (e.g., mAP) that can only confirm that a performance drop occurred, without offering any diagnostic insight into *why* it occurred. Consequently, existing research in this area has been largely confined to the important but incomplete task of dataset construction, rather than the development of comprehensive, diagnostic evaluation protocols. There remains a clear and pressing need for an analytical framework that can systematically probe and uncover specific model failure modes under various, well-characterized domain shifts, moving the community from simply observing the problem to actively diagnosing it.

3. Methodology

The cornerstone of our investigation is the development of a novel evaluation framework, which we term the **GRADE** framework. The primary objective of GRADE is to move beyond the prevalent paradigm of single-metric, leaderboard-driven model comparison, which often fails to provide a deep, diagnostic understanding of a model's out-of-distribution (OOD) generalization capabilities. Instead, we propose a principled, multi-faceted methodology that systematically dissects the complex interplay between data distribution shifts and model performance degradation. The framework is architected to first independently quantify the causal factors of generalization failure—specifically, the statistical divergence between source (training) and target (deployment) data distributions. It then measures the observable effects of these shifts—the degradation in task-specific performance. Finally, it unifies these two dimensions into a coherent, interpretable, and comprehensive diagnostic score. This entire pipeline, which integrates a data-centric distribution analysis with a model-centric performance evaluation through a structured, multi-stage process, is depicted in Figure 1, ensuring that our assessment is both reproducible and standardized across different models and datasets.

3.1. The GRADE Framework: A Principled Overview

The conceptual philosophy of the GRADE framework is rooted in the principle that a truly robust model should not only perform well on unseen data but should also exhibit graceful degradation as the deployment domain increasingly diverges from its training domain. To capture this behavior, our framework is structured as a four-step sequential process, as detailed in Algorithm 1.

First, in the **Distributional Divergence Measurement** stage, we characterize the "distance" between the source domain and a collection of target domains. Crucially, we do this at multiple semantic levels to disentangle different sources of domain shift. Second, in the **Domain Structure Analysis** stage, we leverage these divergence measures to understand the relational structure of the target domains, using unsupervised clustering to group them

into coherent, challenging subsets. This prevents domains that are highly similar to each other from disproportionately influencing the final score. Third, in the **Performance Degradation Measurement** stage, we evaluate the target model's performance on each domain and normalize the performance drop relative to its baseline performance on the source domain. This ensures a fair comparison between models of varying capabilities. Finally, in the **Adaptive Weighting and Score Calculation** stage, we synthesize these components into a single **Generalization Score (GS)**. This score is a weighted aggregation of the model's performance drops, where the weights are adaptively determined by the magnitude of the distribution shift for each domain cluster. This ensures that a model's performance on more challenging, divergent domains is given greater importance, thereby rewarding true generalization robustness.

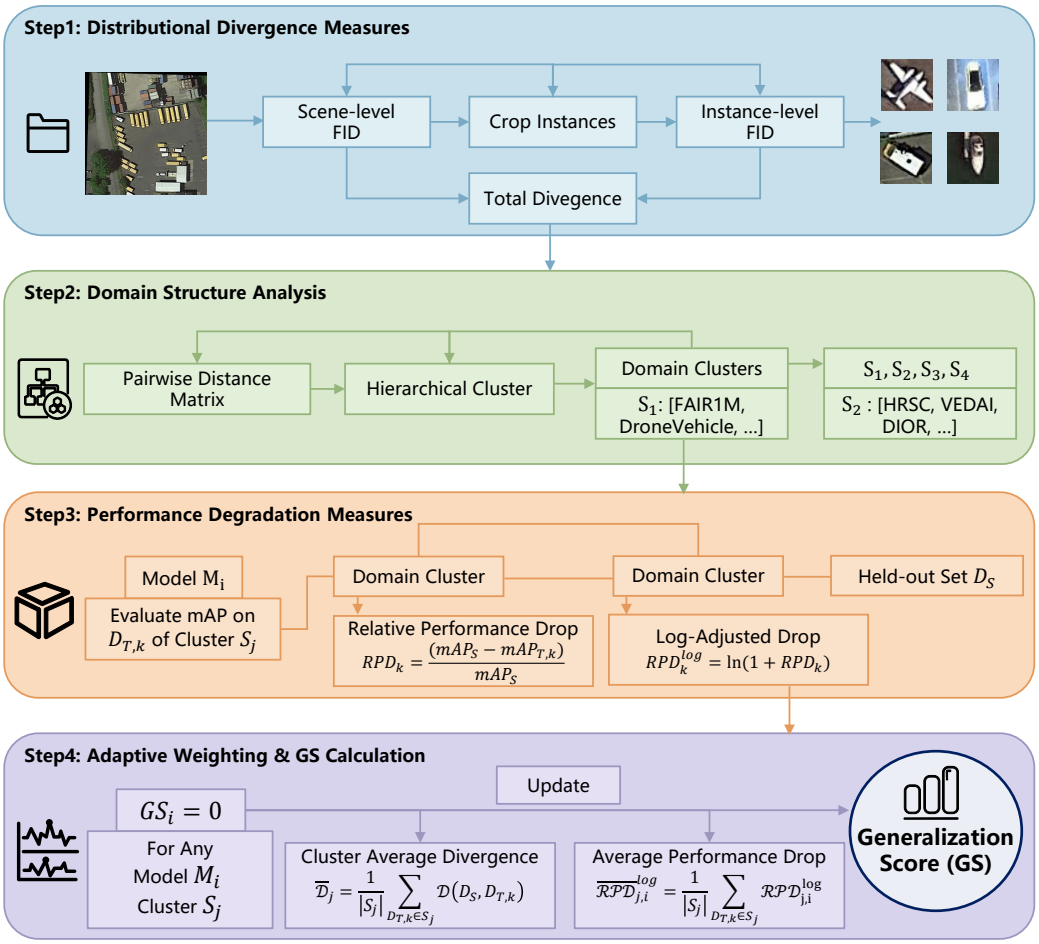


Figure 1. The four-step process of the GRADE framework. The process begins by (**1**) **Measuring Distributional Divergence** at both the scene and object instance levels to quantify domain shifts. Next, (**2**) **Domain Structure Analysis** uses these divergence scores to group target datasets into distinct clusters. Then, (**3**) **Performance Degradation** is calculated for each model as a normalized performance drop on these clusters. Finally, (**4**) **Adaptive Weighting & Score Calculation** integrates the performance drops and divergence measures into a single Generalization Score (GS), where performance on more challenging domains is given greater weight.

3.2. Distributional Divergence Measures

A rigorous quantification of the statistical shift between data distributions is the foundational first step of the GRADE framework. To this end, we employ a metric grounded in statistical theory that is also sensitive to the high-dimensional, semantic nature of image data.

Algorithm 1 GRADE: Generalization Robustness Assessment Framework

Input: Source dataset D_S; A set of K target datasets $\{D_{T,k}\}_{k=1}^{K}$; A set of N models to evaluate $\{M_i\}_{i=1}^{N}$
Output: Generalization Score (GS) for each model M_i.
Step 1: Distributional Divergence Measures
for each target dataset $D_{T,k}$ in $\{D_{T,k}\}_{k=1}^{K}$ **do**
    Let $\Phi(\cdot)$ be a pre-trained feature extractor (e.g., Inception-v3).
    Extract source features $\mathbf{F}_S = \{\Phi(x)|x \in D_S\}$.
    Extract target features $\mathbf{F}_{T,k} = \{\Phi(y)|y \in D_{T,k}\}$.
    Compute Scene-level FID: $FID_{scene}(D_S, D_{T,k})$ from features of full images.
    Compute Instance-level FID: $FID_{inst}(D_S, D_{T,k})$ from features of cropped object instances.
end for
Step 2: Domain Structure Analysis via Clustering
Construct a pairwise distance matrix $\mathbf{D} \in \mathbb{R}^{K\times K}$ where each element $\mathbf{D}_{ij}$ is the total divergence $\mathcal{D}_{total}(D_{T,i}, D_{T,j})$ between two target domains.
Apply Hierarchical Agglomerative Clustering on $\mathbf{D}$ to partition the K target domains into J disjoint subsets $\{S_j\}_{j=1}^{J}$.
Step 3: Performance Degradation Measures
for each model M_i and each domain cluster S_j **do**
    Evaluate baseline source accuracy mAP_S on a held-out validation set from D_S.
    for each target dataset $D_{T,k} \in S_j$ **do**
        Evaluate target accuracy $mAP_{T,k}$ on $D_{T,k}$.
        Compute the Relative Performance Drop: $\mathcal{RPD}_k = (mAP_S - mAP_{T,k})/mAP_S$.
        Compute the log-adjusted performance drop: $\mathcal{RPD}_k^{\log} = \ln(1 + \mathcal{RPD}_k)$.
    end for
end for
Step 4: Adaptive Weighting and Generalization Score (GS) Calculation
for each model M_i **do**
    Initialize $GS_i = 0$.
    for each domain cluster S_j **do**
        Compute the average divergence for the cluster: $\overline{\mathcal{D}}_{total,j} = \frac{1}{|S_j|}\sum_{D_{T,k}\in S_j}\mathcal{D}_{total}(D_S, D_{T,k})$.
        Compute the adaptive weight ω_j using the softmax function (Equation (6)).
        Compute the average performance drop for the cluster: $\overline{\mathcal{RPD}}_{j,i}^{\log} = \frac{1}{|S_j|}\sum_{D_{T,k}\in S_j}\mathcal{RPD}_{k,i}^{\log}$.
        Update the score: $GS_i = GS_i + \omega_j \cdot \overline{\mathcal{RPD}}_{j,i}^{\log}$.
    end for
end for
return The set of GS scores $\{GS_i\}_{i=1}^{N}$, representing a cross-domain generalization penalty (lower is better).

3.2.1. Theoretical Foundation and In-Depth Formulation of FID

The cornerstone of our distributional divergence measure is the **Fréchet Inception Distance (FID)** [30]. FID is a powerful and widely adopted metric that approximates the Wasserstein-2 distance between two probability distributions of features extracted from a deep neural network. Its theoretical appeal stems from the properties of the Wasserstein metric, which, unlike metrics such as Kullback-Leibler (KL) divergence, provides a meaningful distance even for distributions with non-overlapping supports. This makes it particularly suitable for the complex and high-dimensional manifolds of image features. FID's empirical success is demonstrated by its strong correlation with human perceptual judgment of image quality, realism, and diversity, making it a superior choice for our purposes.

The process begins by mapping high-dimensional raw image data into a more compact and semantically meaningful feature space using a pre-trained feature extractor, denoted by the function $\Phi(\cdot)$. For this role, we employ the Inception-v3 network pre-trained on the ImageNet dataset, a standard and deliberate choice. The features learned by Inception-v3, specifically from its final average pooling layer, have proven remarkably general-purpose, capturing a rich hierarchy of textures, shapes, and object parts that are relevant even for specialized domains like remote sensing imagery.

Let D_S be the source dataset and D_T be a target dataset. The extracted feature sets, $\mathbf{F}_S = \{\Phi(x_i)|x_i \in D_S\}$ and $\mathbf{F}_T = \{\Phi(y_j)|y_j \in D_T\}$, are assumed to be samples from continuous multivariate distributions. For tractability and following the standard FID methodology, we model these feature distributions as multivariate Gaussians, $\mathcal{P}_S \sim \mathcal{N}(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S)$ and $\mathcal{P}_T \sim \mathcal{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T)$. The parameters $(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S)$ and $(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T)$ are the sample mean vectors and covariance matrices estimated from the feature sets $\mathbf{F}_S$ and $\mathbf{F}_T$, respectively. The squared Wasserstein-2 distance between these two Gaussian distributions, which defines the FID, is then computed analytically as:

$$\mathrm{FID}(\mathcal{P}_S, \mathcal{P}_T) = \underbrace{\|\boldsymbol{\mu}_S - \boldsymbol{\mu}_T\|_2^2}_{\text{Term 1: Mean Discrepancy}} + \underbrace{\mathrm{Tr}\Big(\boldsymbol{\Sigma}_S + \boldsymbol{\Sigma}_T - 2(\boldsymbol{\Sigma}_S^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_S^{1/2})^{1/2}\Big)}_{\text{Term 2: Covariance Discrepancy}} \tag{1}$$

To fully appreciate its diagnostic power, we dissect this formulation in greater detail:

- **Term 1: Squared Euclidean Distance between Mean Vectors ($d^2(\mu_S, \mu_T)$):** This term quantifies the dissimilarity in the "center of mass" of the two feature distributions. It captures systematic, global shifts that affect the entire dataset. In remote sensing, a large value for this term might signify a bulk translation of the feature manifold due to a change in overall scene brightness (e.g., different sun angles or sensor exposure), a consistent color cast from seasonal variation (e.g., lush green summer landscapes vs. arid brown winter ones), or a fundamental difference in predominant land cover (e.g., urban vs. agricultural). It effectively measures the difference in the average feature activations across all samples.
- **Term 2: Trace of the Covariance Discrepancy:** This more intricate term measures the difference in the internal structure, or "shape", of the feature distributions as captured by their covariance matrices $\boldsymbol{\Sigma}_S$ and $\boldsymbol{\Sigma}_T$. It can be geometrically interpreted as the work required to optimally transform one distribution's covariance ellipsoid into the other's through scaling and rotation of its principal axes. A large value here indicates nuanced structural shifts, which can be further decomposed:
  - **Difference in Variance (related to $\mathrm{Tr}(\boldsymbol{\Sigma}_S)$ and $\mathrm{Tr}(\boldsymbol{\Sigma}_T)$):** The trace of a covariance matrix represents the total variance across all feature dimensions, which serves as a proxy for feature diversity or heterogeneity. For instance, a dataset spanning multiple continents and weather conditions will naturally exhibit higher total variance than a dataset from a single geographical region under clear skies. This term captures such differences in diversity.
  - **Difference in Correlation (Off-diagonal elements and matrix product):** The term $(\boldsymbol{\Sigma}_S^{1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_S^{1/2})^{1/2}$ is a matrix geometric mean that elegantly handles the interaction between the two covariance structures. It penalizes misalignment in the principal axes of variation, capturing changes in how features co-vary. For example, in a port dataset, features corresponding to ships might consistently co-occur with features for docks. If a target dataset consists of ships in the open sea, this correlation structure would be broken. This term is highly sensitive to such subtle but crucial changes in feature relationships.

3.2.2. Hierarchical FID Metrics for Diagnostic Insight

A core innovation of the GRADE framework is its application of FID not as a monolithic, single-value metric, but as a diagnostic instrument computed at two distinct semantic levels of the data hierarchy. This hierarchical approach facilitates the **diagnostic disentanglement** of domain shifts, allowing us to differentiate between shifts originating from the background context and those stemming from the foreground objects of interest.

- **Scene-level FID (FID_{scene}):** This metric is computed using feature vectors extracted from entire, uncropped images. It is therefore designed to measure shifts in the ambient visual domain. This global measure is highly sensitive to changes in background context (e.g., urban vs. rural, desert vs. forest), atmospheric conditions (e.g., clear, hazy, cloudy), global illumination models (e.g., time of day, shadows), and sensor-specific characteristics or artifacts. A high FID_{scene} value strongly suggests that a model's performance degradation may arise from its inability to adapt to novel environmental contexts.
- **Instance-level FID (FID_{inst}):** In contrast, this metric focuses exclusively on the objects of interest by isolating them from their contextual background. To compute it, we first use the ground-truth bounding boxes to crop tight patches around each object. Features are then extracted only from these cropped patches. By design, FID_{inst} quantifies the divergence in object-centric attributes. This includes fine-grained intra-class variations (e.g., different models of aircraft), object texture and material properties, object scale distribution, common occlusion patterns, and aspect ratios. A high FID_{inst} value indicates that the model's failure is likely attributable to an inability to recognize novel object appearances, even if the background context remains familiar.

These two metrics can be combined into a single, holistic measure of divergence, $\mathcal{D}_{total}$, defined as a weighted linear combination:

$$\mathcal{D}_{total}(D_S, D_T) = \alpha \cdot FID_{scene}(D_S, D_T) + \beta \cdot FID_{inst}(D_S, D_T) \quad (2)$$

The hyperparameters $\alpha \geq 0$ and $\beta \geq 0$ allow for task-specific emphasis, offering flexibility for bespoke domain analyses. For instance, a task focused on detecting small objects in highly cluttered scenes might warrant a higher β to emphasize object-level shifts. In the absence of prior task-specific knowledge, and for the purpose of a general-purpose and balanced evaluation, we adhere to the principle of indifference and set $\alpha = \beta = 1.0$.

3.3. Generalization Score (GS) Formulation

Having established a rigorous method for quantifying the "cause" (distributional shift), we now turn to formalizing the "effect" (performance degradation) and synthesizing these two components into our final Generalization Score (GS).

3.3.1. Relative Performance Drop ($\mathcal{RPD}$)

A direct comparison of raw mean Average Precision (mAP) scores across domains can be misleading. An absolute drop of 5% mAP is far more significant for a high-performing model on a relatively easy task (e.g., a drop from 95% to 90%) than for a moderate-performing model on a much harder one (e.g., from 50% to 45%). To account for this baseline dependency and establish a more equitable measure of robustness, we define the **Relative Performance Drop ($\mathcal{RPD}$)** for a given target domain $D_{T,k}$:

$$\mathcal{RPD}_k = \frac{mAP_S - mAP_{T,k}}{mAP_S} \quad (3)$$

This dimensionless metric normalizes the performance loss by the model's own source-domain capability (mAP_S), effectively measuring the fraction of its potential that is forfeited when generalizing to the target domain k. It thus provides a scale-invariant measure of performance degradation.

Furthermore, to mitigate the disproportionate influence of extreme outliers (e.g., a catastrophic failure where $mAP_{T,k} \to 0$, causing $\mathcal{RPD}_k \to 1$) and to improve numerical stability and gradient behavior in potential future extensions, we apply a monotonic logarithmic transformation:

$$\mathcal{RPD}_k^{\log} = \ln(1 + \mathcal{RPD}_k) \tag{4}$$

This transformation gracefully compresses large drops while preserving the rank order of performance, ensuring that the final score is robust to singular catastrophic failures on any one domain.

3.3.2. Adaptive Weighting and Final Score Formulation

The final **Generalization Score (GS)** is the culmination of the GRADE framework, designed as a penalty metric where a lower score indicates superior generalization robustness. The GS achieves a principled synthesis of domain shift magnitude with the resultant performance drop. It is calculated as an adaptively weighted sum of the average performance drops across the J domain clusters identified in Step 2 of our algorithm. For a given model M_i, the score is:

$$GS_i = \sum_{j=1}^{J} \omega_j \cdot \overline{\mathcal{RPD}}_{j,i}^{\log} \quad \text{where} \quad \overline{\mathcal{RPD}}_{j,i}^{\log} = \frac{1}{|S_j|} \sum_{D_{T,k} \in S_j} \mathcal{RPD}_{k,i}^{\log} \tag{5}$$

Here, $\overline{\overline{\mathcal{RPD}}}_{j,i}^{\log}$ is the average log-adjusted RPD for model M_i across all datasets within cluster S_j.

The adaptive weights ω_j are central to the fairness and diagnostic power of the GS. They formalize a critical principle: a model's robustness should be judged more heavily on its ability to generalize to domains that are substantially different from its training experience. A simple average would be naive, treating a small, near-domain shift as equal in importance to a large, far-domain shift. We therefore compute these weights using a softmax function, which transforms the raw cluster divergence scores into a normalized probability distribution of domain difficulty:

$$\omega_j = \frac{\exp(\overline{\mathcal{D}}_{total,j}/\tau)}{\sum_{l=1}^{J} \exp(\overline{\mathcal{D}}_{total,l}/\tau)} \quad \text{s.t.} \quad \sum_{j=1}^{J} \omega_j = 1 \tag{6}$$

In this equation, $\overline{\mathcal{D}}_{total,j}$ represents the average total divergence of all datasets within cluster S_j from the source domain D_S. The **temperature parameter** $\tau > 0$ is a hyperparameter that modulates the "sharpness" of the weighting distribution.

- As $\tau \to \infty$, the weights approach a uniform distribution ($\omega_j \to 1/J$), effectively resulting in a simple average of performance drops across clusters. This represents a "risk-neutral" evaluation.
- As $\tau \to 0^+$, the softmax function approaches an argmax, placing almost all weight ($\omega_j \to 1$) on the single most divergent cluster and ignoring all others. This corresponds to a "worst-case" or "risk-averse" evaluation.

By selecting an appropriate τ (e.g., $\tau = 1.0$ in our experiments), we can strike a balance, ensuring that more challenging domain clusters are given greater emphasis without

completely ignoring performance on easier ones. This mechanism renders the GRADE framework robust and adaptable to diverse evaluation scenarios, from targeted stress-testing to balanced, holistic assessment.

4. Experiments

To empirically validate the efficacy, robustness, and diagnostic utility of the GRADE framework, we designed a comprehensive suite of experiments. Our evaluation protocol is structured to first understand the landscape of existing datasets, then to rigorously assess the generalization capabilities of representative state-of-the-art models, and finally, to provide qualitative evidence that corroborates our quantitative findings.

4.1. Evaluation Models and Datasets

Models. We selected six representative and highly-cited object detection models to ensure a diverse and challenging evaluation. These models span various architectural paradigms: **ReDet** [33] (a rotation-equivariant detector designed for aerial objects), **RoITrans** [34] (a key innovator in learning rotation-invariant RoIs), **Oriented RepPoints** [35] (an anchor-free approach using adaptive point sets), **Oriented RCNN** [36] (a classic and strong two-stage rotated detector), **S^2ANet** [37] (which focuses on feature alignment), and **R3Det** [38] (an efficient single-stage refined detector). This selection covers both two-stage and single-stage, as well as anchor-based and anchor-free methodologies, providing a broad survey of modern object detection architectures.

Datasets. Our experimental testbed includes six diverse and publicly available benchmark datasets. The **DOTA** dataset [8], a large-scale and challenging benchmark with 15 categories and highly varied object scales, served as the single **source domain** for training all models. For out-of-domain evaluation, we utilized five distinct target datasets: **FAIR1M** [9], **DroneVehicle** [10], **DIOR** [39] , **HRSC2016** [40], and **VEDAI** [41]. These datasets were chosen specifically for their heterogeneity, collectively representing significant shifts in geographical location, sensor type, resolution, object scale, class distribution, and background complexity. For instance, HRSC2016 focuses on ship detection with arbitrary orientations, while DroneVehicle presents a nadir-view perspective of small, dense vehicles.

Experimental Setup. In adherence to a strict generalization protocol, all models were trained *exclusively* on the official training split of the DOTA dataset. No fine-tuning or any form of domain adaptation was performed on the target datasets. The models were then directly evaluated on the test splits of the target domains to measure their zero-shot, cross-domain generalization capability.

4.2. Dataset Clustering and Visualization Analysis

Before conducting model evaluations, it is crucial to first quantify the relationships between the domains themselves. To construct meaningful and non-redundant cross-domain evaluation scenarios, we first performed an in-depth analysis of the inter-dataset relationships. Visualizing intrinsic dataset properties, such as the relative average target sizes (Figure 2), immediately revealed significant heterogeneity. For example, HRSC2016 is dominated by large, high-aspect-ratio ship targets, whereas FAIR1M contains a multitude of very small object instances.

To move beyond simple heuristics, we computed our total divergence metric ($\mathcal{D}_{total}$) for all pairs of datasets. We then performed hierarchical agglomerative clustering based on these pairwise divergence scores. The resulting bar chart, shown in Figure 3, clearly partitioned the six datasets into three distinct clusters: {DIOR}, which stands alone due to its unique scene complexity; {FAIR1M, DroneVehicle}, which share characteristics of

having small, densely packed objects; and {HRSC2016, VEDAI}, which are grouped by their focus on specific object types (ships and vehicles, respectively) in less cluttered scenes. This principled, data-driven grouping allowed us to systematically construct diverse evaluation scenarios by selecting representative datasets from each cluster. This ensures our evaluation covers a wide gamut of distributional shifts efficiently, avoiding the redundancy of testing on highly similar domains.

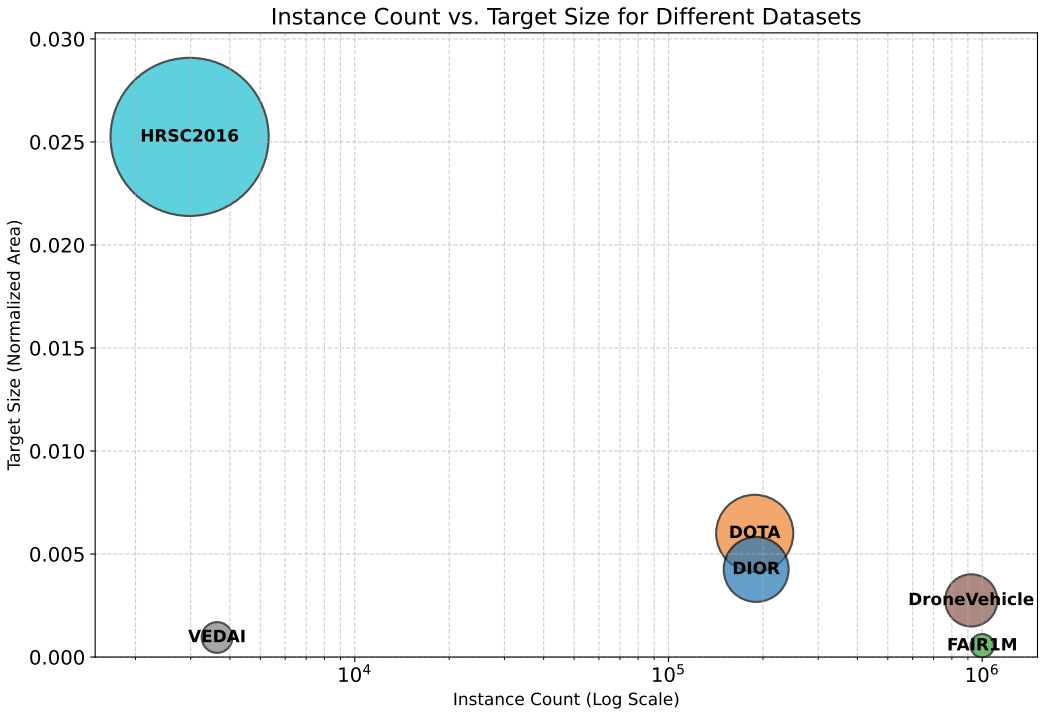


Figure 2. Comparison of datasets by instance count and target size. The x-axis shows the number of instances on a log scale, and the y-axis shows the normalized target area. The bubble size is proportional to the target area.

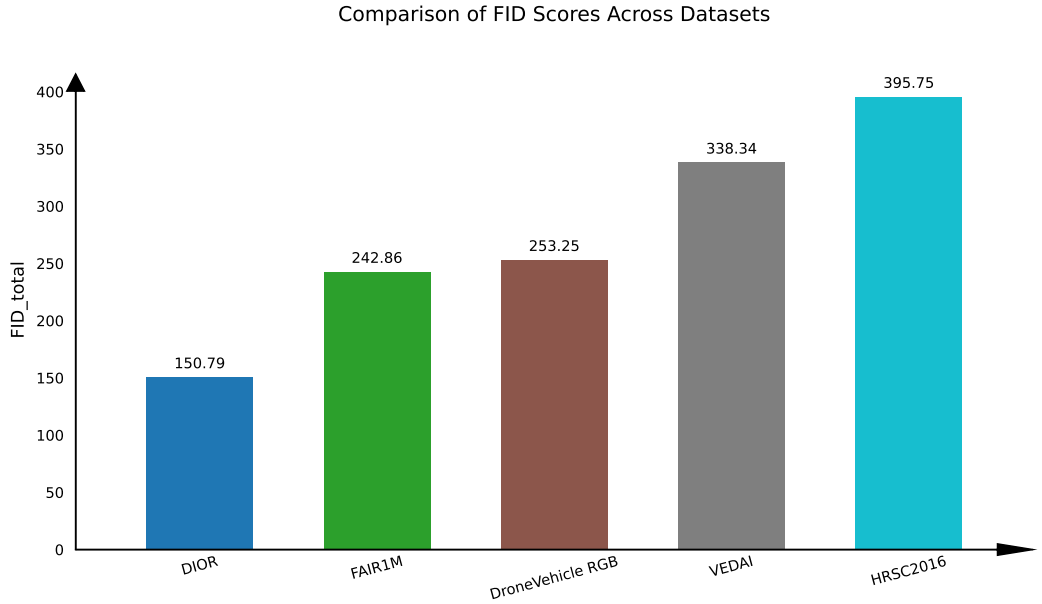


Figure 3. Visualization of dataset distribution distance based on total distributional divergence ($\mathcal{D}_{total}$). The plot quantifies the statistical shift of each target domain from the source domain, DOTA, revealing that DIOR has the smallest distributional shift while HRSC2016 has the most significant.

4.3. Generalization Experiment Results

We conducted extensive cross-domain generalization experiments, training six models on the DOTA source dataset and evaluating them on four different combinations of target datasets, sampled according to our clustering analysis as depicted in Figure 4. We assessed the models using our proposed **Generalization Score (GS)**, where a lower score indicates a smaller generalization penalty and thus better performance. To validate our metric, we compared its rankings to a new pseudo-ground truth metric we term the **Relative Generalization Index (RGI)**. A higher RGI score indicates better performance. The RGI is formulated to reward models that not only achieve high performance on target domains but also do so with minimal performance degradation relative to the source domain. It is defined as the average ratio of target domain performance to the performance drop across N target domains:

$$RGI = \frac{1}{N} \sum_{i=1}^{N} \frac{AP_i}{AP_{source} - AP_i} \tag{7}$$

where AP_i is the model's mAP on the i-th target domain, and AP_{source} is its performance on the source domain (DOTA) test set.

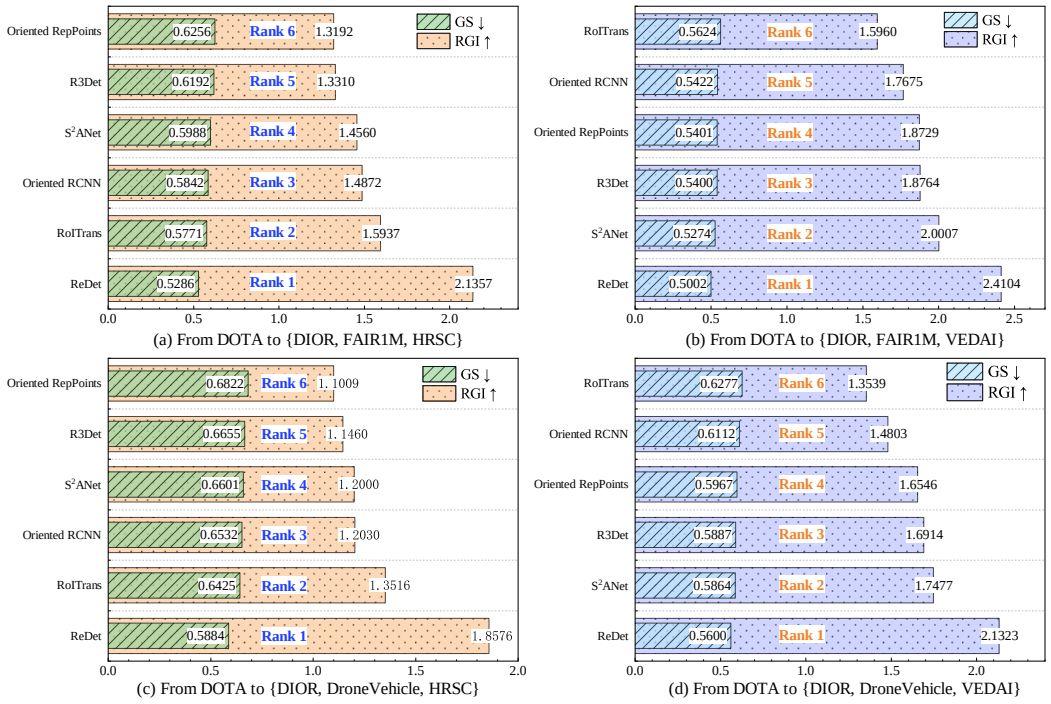


Figure 4. Cross-domain generalization results for six detectors trained on DOTA and evaluated on four target domain combinations (**a**–**d**). The proposed Generalization Score (GS ↓) consistently produces rankings identical to the Relative Generalization Index (RGI ↑), confirming its validity for assessing model robustness.

The results, presented in Tables 1–4, demonstrate a remarkable and consistent alignment between our proposed GS ranking and the empirical RGI ranking. **In every one of the four distinct and challenging experimental scenarios, the ranking order derived from our GS score perfectly matched the ranking order derived from the RGI score.** This high fidelity is a powerful validation, showing that the GRADE framework, despite being a more complex and diagnostic synthesis of distributional and performance metrics, successfully and accurately captures the true generalization hierarchy of the models.

Table 1. Model performance for cross-domain transfer from DOTA to {DIOR, FAIR1M, HRSC}. Lower GS is better; Higher RGI is better.

Model	RGI (Pseudo-GT)	RGI Rank	GS (Ours)	GS Rank
ReDet	2.1357	1	0.5286	1
RoITrans	1.5937	2	0.5771	2
Oriented RCNN	1.4872	3	0.5842	3
S^2ANet	1.4560	4	0.5988	4
R3Det	1.3310	5	0.6192	5
Oriented RepPoints	1.3192	6	0.6256	6

Table 2. Model performance for cross-domain transfer from DOTA to {DIOR, FAIR1M, VEDAI}. Lower GS is better; Higher RGI is better.

Model	RGI (Pseudo-GT)	RGI Rank	GS (Ours)	GS Rank
ReDet	2.4104	1	0.5002	1
S^2ANet	2.0007	2	0.5274	2
R3Det	1.8764	3	0.5400	3
Oriented RepPoints	1.8729	4	0.5401	4
Oriented RCNN	1.7675	5	0.5422	5
RoITrans	1.5960	6	0.5624	6

Table 3. Model performance for cross-domain transfer from DOTA to {DIOR, DroneVehicle, HRSC}. Lower GS is better; Higher RGI is better.

Model	RGI (Pseudo-GT)	RGI Rank	GS (Ours)	GS Rank
ReDet	1.8576	1	0.5884	1
RoITrans	1.3516	2	0.6425	2
Oriented RCNN	1.2030	3	0.6532	3
S^2ANet	1.2000	4	0.6601	4
R3Det	1.1460	5	0.6655	5
Oriented RepPoints	1.1009	6	0.6822	6

Table 4. Model performance for cross-domain transfer from DOTA to {DIOR, DroneVehicle, VEDAI}. Lower GS is better; Higher RGI is better.

Model	RGI (Pseudo-GT)	RGI Rank	GS (Ours)	GS Rank
ReDet	2.1323	1	0.5600	1
S^2ANet	1.7477	2	0.5864	2
R3Det	1.6914	3	0.5887	3
Oriented RepPoints	1.6546	4	0.5967	4
Oriented RCNN	1.4803	5	0.6112	5
RoITrans	1.3539	6	0.6277	6

In the first scenario presented in Table 1, our framework not only correctly identifies **ReDet** as the top performer and **Oriented RepPoints** as the weakest, but it also accurately resolves the subtle performance differences among the mid-tier models. For instance, **Oriented RCNN** and **S^2ANet** have very close RGI scores (1.4872 vs. 1.4560), and our GS metric mirrors this proximity with similarly close scores (0.5842 vs. 0.5988), correctly preserving their rank order. This initial result highlights the sensitivity and reliability of our proposed evaluation methodology.

As shown in Table 2, swapping the HRSC dataset for VEDAI significantly reshuffles the model hierarchy, demonstrating that generalization capability is highly dependent on the specific nature of the domain shift. Notably, **S^2ANet** shows a remarkable improvement,

jumping from 4th to 2nd place, while **RoITrans**, a strong performer in the previous scenario, drops to last place. This dynamic shift underscores the importance of a robust evaluation framework. Despite this significant reordering, the ranking from our GS score remains in perfect agreement with the RGI ranking, even correctly ordering the nearly identical scores of **R3Det** (1.8764) and **Oriented RepPoints** (1.8729).

The third scenario in Table 3, which introduces the challenging DroneVehicle dataset, continues to validate the consistency of our framework. Once again, the GS rankings perfectly align with the RGI rankings. This experiment particularly highlights the framework's precision in handling extremely close performance levels. For example, our method correctly resolves the minute performance difference between **Oriented RCNN** and **S^{2}ANet**, whose RGI scores differ by a mere 0.003. The corresponding GS scores (0.6532 vs. 0.6601) accurately capture this narrow margin while preserving the correct rank order, proving the high sensitivity of our GS metric.

In the final and arguably most diverse test scenario presented in Table 4, the perfect correlation between the GS and RGI rankings holds true yet again, solidifying the reliability of our proposed evaluation methodology. This combination of target domains reveals further insights into model-specific generalization behavior. For instance, **RoITrans**, which ranked 2nd in two other scenarios, consistently places last in scenarios involving VEDAI, suggesting a particular vulnerability to the characteristics of that dataset. In contrast, models like **S^{2}ANet** demonstrate more stable high-tier performance. The ability of our GS score to consistently mirror the empirical RGI rankings across all these varied and challenging scenarios validates its robustness and utility as a comprehensive measure of generalization.

5. Discussion

To provide qualitative and intuitive validation that complements our quantitative results, we visualized the detection outputs from all six models on representative samples from the target domains (Figure 5). And the figure's upper section displays the performance on sparse targets, whereas the lower section displays the performance on dense targets. These visualizations serve as a powerful sanity check, allowing for a direct, human-in-the-loop assessment of model behavior.

The visual results clearly illustrate the performance hierarchy that was precisely captured by our GS metric in the corresponding quantitative experiment (Table 3). The lower-ranked models, such as **Oriented RepPoints** (Rank 6), exhibit a high incidence of failure modes, including numerous false positives in cluttered backgrounds and poor localization with inaccurate bounding boxes that fail to tightly enclose the target objects. Intermediate models like **Oriented RCNN** and **S^{2}ANet** (Ranks 3 and 4) demonstrate improved performance but still struggle with challenging cases involving dense or occluded targets, often missing objects or producing duplicate detections. In stark contrast, the top-ranked model, **ReDet** (Rank 1), demonstrates superior generalization. It maintains high precision and recall, producing accurate, well-oriented bounding boxes and maintaining structural coherence even in the most challenging scenes. These visual results provide compelling, tangible evidence that our GRADE framework accurately captures meaningful and practically significant differences in model robustness.

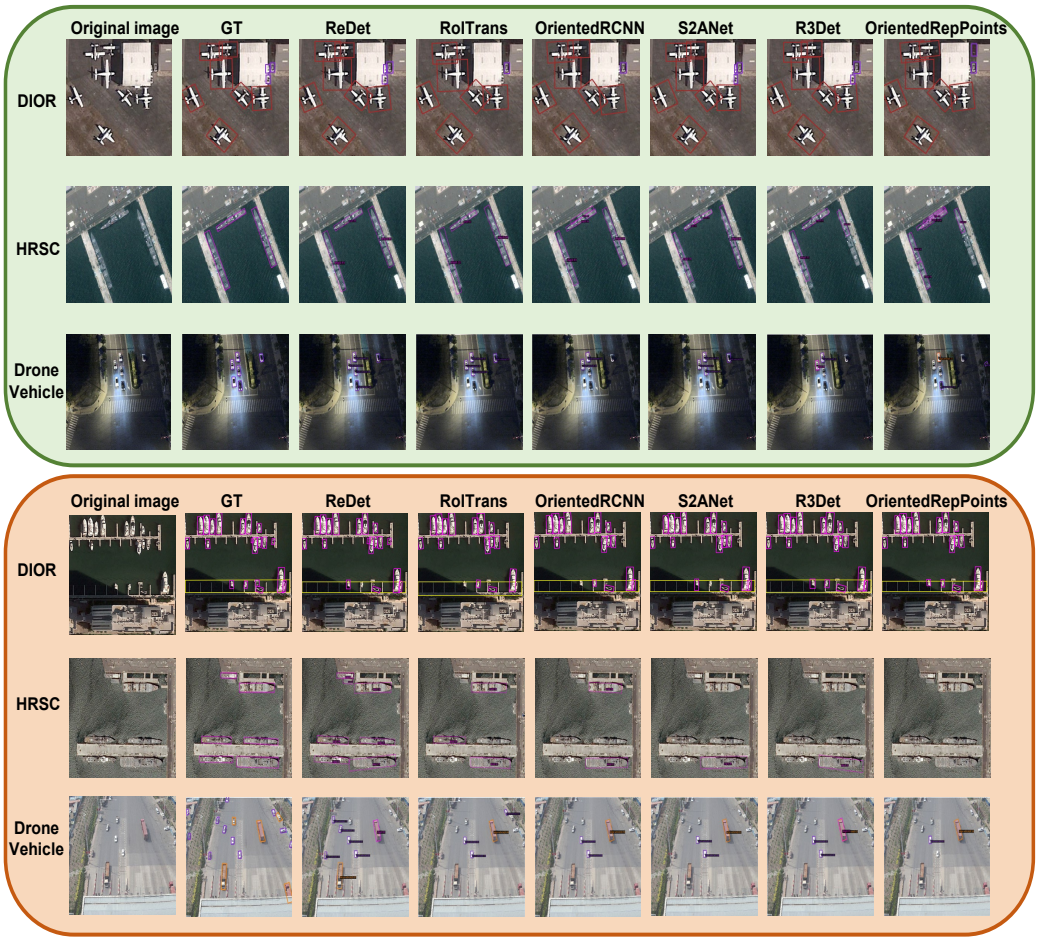


Figure 5. Detection results of different models on DIOR, DroneVehicle, and HRSC datasets.

6. Conclusions

This paper introduced the **GRADE** framework, a comprehensive framework for assessing the generalization robustness of remote sensing object detectors. By unifying the quantification of data distribution divergence and performance degradation, our framework provides a systematic evaluation of cross-domain robustness. We introduced **Scene-level** and **Instance-level FID** to dissect distributional shifts and integrated these with a relative performance drop metric into an adaptively weighted **Generalization Score (GS)**. This score not only quantifies robustness but also helps analyze the sources of performance decay. Extensive experiments on six major benchmarks confirmed that our framework's rankings perfectly match the rankings from the **Relative Generalization Index (RGI)**, a robust pseudo-ground truth metric we introduced, thus offering a reliable and interpretable basis for model selection, optimization, and deployment. Future work will extend this framework to more heterogeneous domains and other vision tasks. When combined with model optimization techniques like domain adaptation, it can create a closed-loop system to guide the development of more robust and generalizable models.

Author Contributions: Conceptualization, X.Y., D.W. and X.S.; methodology, D.W. and Y.Z.; software, Y.Z.; validation, Y.Z. and B.B.; formal analysis, D.W.; investigation, Y.Z. and B.B.; resources, B.B. and Y.D.; data curation, Y.Z. and X.S.; writing—original draft preparation, D.W.; writing—review and editing, X.Y. and Y.D.; visualization, Y.Z.; supervision, X.Y.; project administration, X.Y.; funding acquisition, X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (62301261, 62222207, and 62427808).

Data Availability Statement: The datasets used in this study are all publicly available.

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

References

1. Dai, Y.; Zou, M.; Li, Y.; Li, X.; Ni, K.; Yang, J. DenoDet: Attention as Deformable Multisubspace Feature Denoising for Target Detection in SAR Images. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 4729–4743. [CrossRef]
2. Pang, D.; Shan, T.; Ma, Y.; Ma, P.; Hu, T.; Tao, R. LRTA-SP: Low-Rank Tensor Approximation with Saliency Prior for Small Target Detection in Infrared Videos. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 2644–2658. [CrossRef]
3. Ni, K.; Yuan, C.; Zheng, Z.; Wang, P. SAR Image Time Series for Land Cover Mapping via Sparse Local–Global Temporal Transformer Network. *IEEE Trans. Aerosp. Electron. Syst.* **2025**, *61*, 12581–12597. [CrossRef]
4. Zhao, Z.; Wen, Z.; Xue, C.; Cui, Z.; Hou, X.; Zhu, H.; Mu, Y.; Liu, Z.; Xia, Z.; Liu, X. Improved Clutter Suppression and Detection of Moving Target with a Fully Polarimetric Radar. *Remote Sens.* **2025**, *17*, 2975. [CrossRef]
5. Ringwald, T.; Stiefelhagen, R. Adaptiope: A Modern Benchmark for Unsupervised Domain Adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2021; pp. 101–110.
6. Hou, F.; Zhang, Y.; Dong, J.; Fan, J. End-to-End Model Enabled GPR Hyperbolic Keypoint Detection for Automatic Localization of Underground Targets. *Remote Sens.* **2025**, *17*, 2791. [CrossRef]
7. Liu, Y.; Zou, Y.; Qiao, R.; Liu, F.; Lee, M.L.; Hsu, W. Cross-Domain Feature Augmentation for Domain Generalization. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 3–9 August 2024; pp. 1146–1154.
8. Xia, G.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
9. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A Benchmark Dataset for Fine-grained Object Recognition in High-Resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
10. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-based RGB-infrared Cross-Modality Vehicle Detection via Uncertainty-aware Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [CrossRef]
11. Al-Emadi, S.A.; Yang, Y.; Ofli, F. Benchmarking Object Detectors under Real-World Distribution Shifts in Satellite Imagery. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 8299–8309.
12. He, T.; Sun, K.; Duan, Y.; Cui, W.; Wang, Z.; Gao, S.; Yao, Y.; Chen, Z. ViTrans: Inter-Frame Alignment Enhancement for Moving Vehicle Detection in Satellite Videos with Stabilization Offsets. *Remote Sens.* **2025**, *17*, 2973. [CrossRef]
13. Li, D.; Yang, Y.; Song, Y.; Hospedales, T. Deeper, Broader and Artier Domain Generalization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
14. Koh, P.W.; Sagawa, S.; Marklund, H.; Xie, S.M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R.L.; Gao, I.; et al. WILDS: A Benchmark of In-the-Wild Distribution Shifts. In Proceedings of the International Conference on Machine Learning (ICML), Virtual Event, 18–24 July 2021; pp. 5637–5664.
15. Mao, X.; Chen, Y.; Zhu, Y.; Chen, D.; Su, H.; Zhang, R.; Xue, H. COCO-O: A Benchmark for Object Detectors under Natural Distribution Shifts. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 6339–6350.
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
17. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
18. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
19. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
20. Zhang, Z.; Lu, X.; Cao, G.; Yang, Y.; Jiao, L.; Liu, F. ViT-YOLO: Transformer-based YOLO for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 2799–2808.
21. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection. *arXiv* **2025**, arXiv:2509.25164. [CrossRef]

22. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
23. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. DN-DETR: Accelerate DETR Training by Introducing Query Denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13619–13627.
24. Chen, Y.; Liu, B.; Yuan, Y. PR-Deformable DETR: DETR for Remote Sensing Object Detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2506105. [CrossRef]
25. Zhang, S.; Song, F.; Liu, X.; Hao, X.; Liu, Y.; Lei, T.; Jiang, P. Text Semantic Fusion Relation Graph Reasoning for Few-Shot Object Detection on Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1187. [CrossRef]
26. Yao, L.; Han, J.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, H. DetCLIPv2: Scalable Open-Vocabulary Object Detection Pre-Training via Word-Region Alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 23497–23506.
27. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September –4 October 2024; pp. 38–55.
28. Van Erven, T.; Harremos, P. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]
29. Li, C.; Chang, W.; Cheng, Y.; Yang, Y.; Poczos, B. MMD GAN: Towards Deeper Understanding of Moment Matching Network. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 60, pp. 1–11.
30. YU, Y.; Zhang, W.; Deng, Y. Frechet Inception Distance (FID) for Evaluating GANs. *China Univ. Min. Technol. Beijing Grad. Sch.* **2021**, *3*, 1–7.
31. Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; Yao, Y. Poly Kernel Inception Network for Remote Sensing Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024.
32. Michaelis, C.; Mitzkus, B.; Geirhos, R.; Rusak, E.; Bringmann, O.; Ecker, A.S.; Bethge, M.; Brendel, W. Benchmarking Robustness in Object Detection: Autonomous Driving When Winter is Coming. *arXiv* **2019**, arXiv:1907.07484.
33. Han, J.; Ding, J.; Xue, N.; Xia, G. ReDet: A Rotation-Equivariant Detector for Aerial Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021.
34. Ding, J.; Xue, N.; Long, Y.; Xia, G.; Lu, Q. Learning ROI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
35. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented RepPoints for Aerial Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
36. Pu, C.; Yu, J.; Su, W.; Liu, T. Rotated R-CNN: A Two-Stage Object Detection Method Adapted To Oriented Bounding Boxes. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 27–30 October 2024.
37. Liu, Y.; Sun, X.; Shao, W.; Yuan, Y. S²ANet: Combining Local Spectral and Spatial Point Grouping for Point Cloud Processing. *Virtual Real. Intell. Hardw.* **2024**, *6*, 267–279. [CrossRef]
38. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021.
39. Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote. Sens.* **2020**, *159*, 296–307. [CrossRef]
40. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A High Resolution Optical Satellite Image Dataset for Ship Recognition and Some New Baselines. In Proceedings of the International Conference on Pattern Recognition Applications and Methods (CPRAM), Porto, Portugal, 24–26 February 2017.
41. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]

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