REOBench: Benchmarking Robustness of Earth Observation Foundation Models

Xiang Li[1]*, Yong Tao[2]*, Siyuan Zhang[3]*, Siwei Liu[4], Zhitong Xiong[5]
Chunbo Luo[2], Lu Liu[2], Mykola Pechenizkiy[6], Xiao Xiang Zhu[5], Tianjin Huang[2,6]†

[1] University of Bristol, UK [2] University of Exeter, UK
[3] South China Normal University, China [4] The University of Aberdeen, UK
[5] Technical University of Munich, Germany [6] Eindhoven University of Technology, NL

Abstract

Earth observation foundation models have shown strong generalization across multiple Earth observation tasks, but their robustness under real-world perturbations remains underexplored. To bridge this gap, we introduce REOBench, the first comprehensive benchmark for evaluating the robustness of Earth observation foundation models across six tasks and twelve types of image corruptions, including both appearance-based and geometric perturbations. To ensure realistic and fine-grained evaluation, our benchmark focuses on high-resolution optical remote sensing images, which are widely used in critical applications such as urban planning and disaster response. We conduct a systematic evaluation of a broad range of models trained using masked image modeling, contrastive learning, and vision-language pre-training paradigms. Our results reveal that ❶ existing Earth observation foundation models experience significant performance degradation when exposed to input corruptions. ❷ The severity of degradation varies across tasks, model architectures, backbone sizes, and types of corruption, with performance drop varying from less than 1% to over 25%. ❸ Vision-language models show enhanced robustness, particularly in multimodal tasks. REOBench underscores the vulnerability of current Earth observation foundation models to real-world corruptions and provides actionable insights for developing more robust and reliable models. Code and data are publicly available at https://github.com/lx709/REOBench.

1 Introduction

Recent studies have shown that foundation models pre-trained on large-scale datasets have demonstrated powerful capabilities across multiple domains. Models such as MAE [1], CLIP [2], MiniGPT-4 [3], and LLaVA [4] have achieved remarkable success in multiple vision and vision-language tasks. These models are capable of extracting representative features from images, and more importantly, can be quickly adapted to multiple downstream tasks with minimal fine-tuning, significantly improving the efficiency and effectiveness of task-solving.

In the field of remote sensing, the rapid growth in data volume has sparked significant interest in developing foundation models for remote sensing image analysis. Earth observation foundation models (EOFMs) aim to leverage supervised or self-supervised training to build large-scale pre-trained models that can be adapted to a wide range of downstream tasks, thereby improving the performance and efficiency of Earth observation applications. In recent years, a growing body of research has focused on constructing such foundation models tailored for remote sensing. Mainstream

[*] First three authors contributed equally.

[†] Corresponding author: `t.huang2@exeter.ac.uk`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

models can be divided into unimodal pre-trained foundation models (e.g., SatMAE [5], RingMo [6], ScaleMAE [7], SpectralGPT [8]) and vision-language foundation models (e.g., RemoteCLIP [9], GeoRSCLIP [10], RSGPT [11], GeoChat [12]). These EOFMs have demonstrated their powerful performance in numerous downstream tasks. A comprehensive review of EOFMs can be found in [13–19].

Despite advancements in EOFMs, there remains a significant gap in systematically benchmarking their robustness towards image perturbations. Remote sensing images are particularly susceptible to factors such as weather conditions or sensor discrepancies, which can introduce significant noise and variability [20–22], posing challenges to current EOFMs. Therefore, developing a comprehensive benchmark to evaluate and compare the robustness of these models holds great academic and practical significance. Such benchmarking efforts can guide the design of highly robust EOFMs that effectively adapt to noise and data variability, ensuring stable and reliable results under diverse conditions.

To achieve this, we introduce **REOBench**, a comprehensive **Bench**mark designed to evaluate the **R**obustness of **E**arth **O**bservation foundation models, covering state-of-the-art models based on masked image modeling, contrastive learning, and large language models. REOBench focuses on high-resolution optical remote sensing images, which are widely used in real-world applications such as urban planning and disaster response. We conducted experiments on **six** widely studied remote sensing image understanding tasks, covering both vision-centric and vision-language tasks, under **twelve** types of perturbations. These include both appearance-based corruptions (e.g., noise, blur, haze) and geometric distortions (e.g., rotation, scale, translation), applied at varying severity levels to simulate realistic environmental and sensor-induced challenges. Our evaluation yields three key findings:

- ⋆ Existing Earth observation foundation models suffer noticeable performance degradation under common image corruptions, with particularly sharp drops for the models based on masked image modeling.
- ⋆ The degree of vulnerability to image corruptions varies across tasks, model architectures, and types of perturbations, with performance drop varying from less than 1% to over 20%.
- ⋆ Vision-language foundation models exhibit greater robustness to visual perturbations compared to vision-centric foundation models, particularly in image-level scene classification tasks.

In summary, REOBench provides the first large-scale, task-diverse, and perturbation-rich benchmark for evaluating robustness in EOFMs. It offers actionable insights for the research community and serves as a stepping stone toward building more reliable, generalizable, and trustworthy AI systems for Earth observation.

2 REOBench Dataset

To systematically evaluate the robustness of EOFMs, we construct a benchmark dataset by incorporating widely used remote sensing datasets spanning diverse tasks. Specifically, we include AID [23] for scene classification, ISPRS Potsdam [24] for semantic segmentation, DIOR [25] for object detection, and three subsets from VRSBench [26] for image captioning, visual question answering (VQA), and visual grounding. These datasets are selected based on their popularity, diversity of content, and relevance to the tasks under evaluation.

2.1 Corruptions in Remote Sensing Images

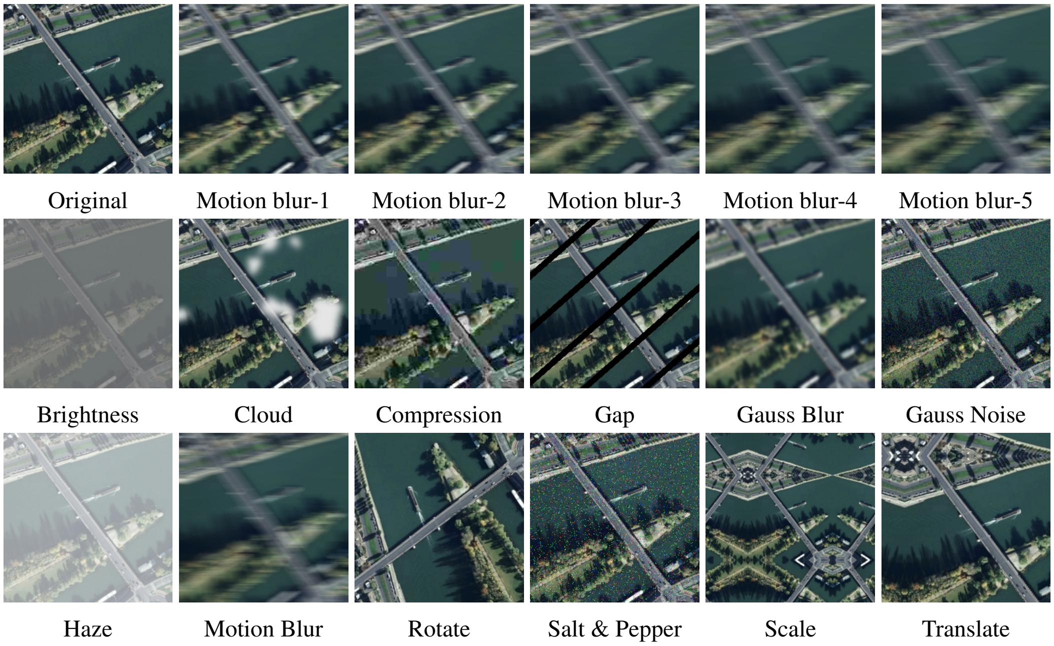
Remote sensing platforms are subject to a wide range of visual degradations that differ significantly from those encountered by ground-based cameras. To systematically evaluate the robustness of RSFMs, we construct a benchmark comprising **12 synthetic corruptions**, categorized into three types: *environmental*, *sensor-induced*, and *geometric*. Each corruption is generated using physically or statistically grounded procedures to ensure the resulting images remain photorealistic while faithfully reflecting failure modes commonly observed in satellite and UAV imagery.

Environmental Corruptions. Atmospheric and illumination variations constitute predominant environmental degradations. For instance, *Cloud* occlusions substantially obscure optical remote sensing data, severely impacting scene interpretability [27, 28]. Variations in *Brightness* resulting from shifting sun angles affect radiometric stability and degrade feature matching and object recognition performance [29, 30]. *Haze*, caused by aerosol scattering, significantly lowers image contrast and impairs detection and classification accuracy [31, 32]. Following established benchmarks [33], we simulate these environmental corruptions using physically motivated image augmentation techniques.

Sensor-induced Corruptions. Imperfections during sensor capture or data transmission introduce various degradations. *Gaussian Blur*, indicative of defocusing or modulation transfer function (MTF) degradation, compromises tie-point accuracy and feature localization [34, 35]. *Motion Blur*, arising from platform vibrations or rapid movements, negatively impacts object detection and tracking in aerial inspections [36]. *Gaussian Noise* and *Salt & Pepper Noise*, simulating electronic interference and bit-flip errors respectively, significantly decrease segmentation and classification accuracy [37, 38]. *Sensor Gap* degradations, exemplified by the Landsat-7 SLC-off issue, necessitate specialized gap-filling methodologies [39, 40]. Furthermore, *Compression* artifacts, such as those from JPEG/JPEG2000, substantially impair the quality of CNN feature extraction [41]. These sensor-induced corruptions are replicated through established augmentation and simulation protocols in line with existing research [33, 42].

Geometric Corruptions. Geometric distortions originate primarily from variations in sensor orientation, altitude, and registration accuracy. *Rotation* caused by platform roll or yaw introduces inconsistencies in orientation-sensitive feature extraction processes [43]. *Scale* alterations resulting from altitude fluctuations pose significant challenges for detectors lacking robust multi-scale adaptability [44]. *Translation*, modeling inaccuracies due to GPS drift, registration errors, or parallax, adversely affects pixel-aligned or patch-based analysis methods [45]. To effectively simulate these geometric degradations, we apply spatial transformations, including image *Rotation*, *Scaling*, and *Translation*, to remote sensing images.

In total, these corruption categories encompass twelve distinct types. Each type of corruption is applied consistently across all datasets at five severity levels. Fig. 1 illustrates one example of original and corrupted images.



Original Motion blur-1 Motion blur-2 Motion blur-3 Motion blur-4 Motion blur-5

Brightness Cloud Compression Gap Gauss Blur Gauss Noise

Haze Motion Blur Rotate Salt & Pepper Scale Translate

Figure 1: Example of perturbed images. In the first row, we present the original clean image alongside images perturbed by five levels of motion blur. The second and third rows illustrate examples of images corrupted by a range of perturbation types.

2.2 Definition of Corruption Robustness

We formalize the corruption robustness of EOFMs as its ability to maintain task performance in the presence of realistic geophysical and sensor-induced degradations frequently encountered in Earth observation. Let $f : \mathcal{X} \rightarrow \mathcal{Y}$ denote an EOFMs that maps an input image $x \in \mathcal{X}$ to a label $y \in \mathcal{Y}$, with (x, y) sampled from an underlying geospatial data-generating distribution $\mathcal{D}$. We define a set of corruptions $\mathcal{C} = \{c_1, \ldots, c_K\}$, where each $c_k \in \mathcal{C}$ represents a physically plausible corruption operator, such as haze, cloud occlusion, or sensor noise. Each corruption occurs with a non-zero prevalence $\mathbb{P}_{\mathcal{C}}(c_k) > 0$. To quantify robustness, we define the *relative task performance drop* ($\mathcal{R}_{\text{TP}}$), which measures the degradation in model performance under corrupted inputs relative to its accuracy on clean data:

$$\mathcal{R}_{\text{TP}} \;=\; \frac{\mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x) = y] - \mathbb{E}_{c\sim\mathcal{C}}\left[\mathbb{P}_{(x,y)\sim\mathcal{D}}[f(c(x)) = y]\right]}{\mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x) = y]}. \tag{1}$$

A *smaller* $\mathcal{R}_{\text{TP}}$ indicates greater robustness, as it reflects less relative degradation when encountering corrupted data.

3 Benchmark Robustness on EOFMs

We evaluate the robustness of EOFMs across six widely studied remote sensing image understanding tasks: scene classification, semantic segmentation, object detection, image captioning, visual question answering (VQA), and visual grounding. The evaluated models represent the current state-of-the-art in remote sensing and can be broadly categorized into the following three types.

MIM-based foundation models. Masked image modeling (MIM) has gained popularity through the pioneering work of MAE [1]. In the field of remote sensing, notable approaches include SatMAE [5], RVSA [46], ScaleMAE [7], and SatMAE++ [47].

CL-based foundation models. Building on the success of the pioneering work CLIP [2], multiple contrastive learning (CL) -based foundation models have been introduced in the field of remote sensing, such as RemoteCLIP [9] and GeoRSCLIP [10]. To investigate robustness with respect to different backbone sizes, we evaluate two commonly used architectures in our experiments: ViT-B/32 and ViT-L/14 [48].

LLM-based foundation models. Following the pioneering works of GPT-4 [49], MiniGPT-4 [3], and LLaVA [4], multimodal large language models (MLLMs) have attracted significant research attention in recent years. Notable approaches include GeoChat [12], LHRS-Bot [50], RS-LLaVA [51], VHM [52], SkySenseGPT [53], and Falcon [54]. In our experiments, we evaluate models with open-access code and pretrained weights for comparison.

3.1 Implementation Details

For MIM- and CLIP-based models, we take the vision backbones from pretrained foundation models and append a task-specific head (e.g., MLP, detectors, or segmentors) for each task. For these LLM-based models, since these generalist models usually freeze their vision backbones and can naturally handle multiple tasks, we directly evaluate *zero-shot* performance of these models to test their robustness.

For the scene classification task, we take the backbone from all pretrained foundation models and append a single linear layer after the backbone for classification. For the semantic segmentation (resp. detection) task, we follow RVSA [46] to use UpperNet [55] (resp. Oriented R-CNN [56]) as the segmentor (resp. detector) and replace its backbone with that from pretrained foundation models. Following RVSA [46], we build a feature pyramid from blocks 4, 6, 8, and 12 using up/down-sampling. All models are trained for 12 epochs with an initial learning rate of 1e-5, decayed by 0.1 at epochs 8 and 11.

For vision-language tasks, including image captioning, visual question answering, and visual grounding, we follow the original paper designs to craft task-specific prompts and evaluate the zero-shot performance of these foundation models to assess robustness. It should be noted that different pretrained foundation models are designed to accept images at specific resolutions. When the input image size differs from the pretrained backbone's expected resolution, we interpolate the position embeddings in the backbone to accommodate the new input dimensions.

3.2 Scene Classification

From Table 1, we can draw the following findings: 1) All benchmark methods suffer from serious performance under image corruptions for the scene classification tasks, especially for MIM-based methods. 2) CL- and LLM-based methods are more robust towards image corruptions than MIM-based methods. This is probably because CL- and LLM-based methods are trained by matching image-text pairs in a shared embedding space, learning high-level semantic features less sensitive to low-level corruptions. In contrast, MIM-based methods are trained by reconstructing pixel- or token-level details, making them sensitive to local corruptions. Specifically, VHM [52] achieves the least performance drop under image corruptions. 3) CL-based methods usually perform better than MIM and LLM-based methods on the scene classification task, for both clean and noisy images. Specifically, GeoRSCLIP [10] achieves the best scene classification performance under image corruptions.

Table 1: Scene classification performance on AID dataset across different image perturbations. zs denotes zero-shot evaluation.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Rotate	Salt Pepper	Scale	Translate	Avg	$\mathcal{R}_{\text{TP}}$
MIM-based																
SATLAS [57]	Swin-B	90.85	82.54	84.32	73.36	67.23	78.10	79.16	80.46	32.44	72.54	77.56	72.54	88.54	74.07	18.47
SatMAE [5]	ViT-L	72.05	44.82	59.58	67.26	46.49	71.33	71.25	28.31	63.85	69.15	70.45	59.74	66.12	59.86	16.92
Scale-MAE [7]	ViT-L	75.75	51.80	72.65	39.60	43.69	31.65	46.31	55.24	17.49	66.15	47.27	61.58	69.84	50.27	33.64
RVSA [46]	ViT-B	84.60	56.84	77.33	56.07	53.14	53.53	32.51	49.19	23.45	76.88	35.12	71.78	77.22	55.26	34.69
SatMAE++ [47]	ViT-L	91.35	64.62	82.64	62.69	60.70	48.23	76.98	62.56	29.43	85.49	73.22	75.79	87.61	67.50	26.11
CL-based																
RemoteCLIP$_{zs}$ [9]	ViT-L	81.10	78.32	80.64	73.91	79.43	76.83	76.72	80.10	57.02	82.80	70.90	68.39	80.72	75.48	6.93
RemoteCLIP [9]	ViT-B	96.85	90.80	95.36	91.13	88.96	89.18	94.25	87.46	63.75	96.22	91.43	83.62	**95.42**	88.97	8.15
RemoteCLIP [9]	ViT-L	95.45	93.11	93.80	88.77	94.21	92.47	94.20	93.37	**74.45**	95.01	86.99	83.37	94.06	90.32	5.38
GeoRSCLIP$_{zs}$ [10]	ViT-L	66.05	62.41	65.47	60.45	64.41	62.03	62.32	62.24	44.20	65.52	58.88	52.59	64.25	60.40	8.55
GeoRSCLIP [10]	ViT-B	96.90	93.59	96.04	91.01	93.34	92.60	92.99	92.91	57.78	95.70	88.22	75.70	93.87	88.65	8.51
GeoRSCLIP [10]	ViT-L	**97.40**	**96.27**	**96.45**	**92.28**	**95.62**	**96.09**	**95.68**	**96.00**	71.03	**97.20**	**92.75**	**77.16**	95.15	**91.80**	5.74
LLM-based																
GeoChat [12]	ViT-L	65.85	64.67	65.26	60.71	64.61	63.32	62.34	64.54	48.68	65.05	62.32	56.21	62.91	61.72	6.27
LHRS-Bot [50]	ViT-L	87.75	87.38	86.82	78.53	85.95	84.94	82.62	87.62	67.76	87.31	76.73	79.07	86.71	82.79	5.65
RS-LLaVA [51]	ViT-L	67.55	65.36	68.95	63.05	67.69	65.73	63.13	65.97	43.86	66.55	68.24	54.89	63.40	63.07	6.63
SkySenseGPT [53]	ViT-L	87.35	87.72	87.91	79.66	87.67	84.78	83.08	87.71	63.11	86.86	83.61	75.49	85.25	82.74	5.28
VHM [52]	ViT-L	80.60	79.81	80.67	76.10	80.82	78.81	76.92	79.55	59.74	80.47	75.43	72.57	79.73	76.72	**4.81**

3.3 Semantic Segmentation

The ISPRS Potsdam dataset provides both non-eroded and eroded labels, corresponding to annotations with and without object boundaries, respectively. In our experiments, we use the non-eroded labels for model evaluation and report the mean IoU for MIM- and CL-based methods. We omit LLM-based methods for semantic segmentation due to the lack of open-source MLLMs for this task. From Table 2, we can draw the following findings: 1) both MIM- and CL-based methods suffer from serious performance under image corruptions for the object detection task, with a mIoU drop of more than 10%. 2) MIM-based methods achieve much better performance than CL-based methods on clean and noisy images. This is probably because MIM-based methods can capture local details by pixel reconstruction, while CL-based methods align global visual embeddings with text features, thus losing fine-grained details. Specifically, ScaleMAE [7] achieves the best segmentation performance under image corruption, with an average mIoU of 60.02%. 3) MIM-based methods suffer from a more serious performance drop under image corruptions than CL-based methods. Specifically, GeoRSCLIP [10] achieves the best robustness across image corruptions, with a drop of 10.01% of mIoU under corruptions.

Table 2: Semantic segmentation performance (mIoU) on the ISPRS Potsdam dataset under different image perturbations.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Rotate	Salt Pepper	Scale	Translate	Avg	$\mathcal{R}_{\text{TP}}$
MIM-based																
SatMAE [5]	ViT-L	59.51	50.18	37.39	48.23	51.15	57.89	41.83	44.95	57.52	56.02	36.07	54.81	59.09	49.59	16.67
ScaleMAE [7]	ViT-L	68.92	64.37	65.43	**49.96**	41.84	64.86	**54.89**	**63.89**	64.49	64.51	**52.12**	65.45	68.38	**60.02**	12.91
RVSA [46]	ViT-B	**69.82**	**64.71**	**65.67**	47.99	45.34	**66.89**	48.89	61.52	**65.25**	**65.58**	45.26	**66.87**	**69.23**	59.43	14.88
SatMAE++ [47]	ViT-L	62.68	53.91	59.06	49.74	**53.94**	60.38	44.32	48.80	60.44	58.34	39.45	58.13	61.94	54.04	13.78
CL-based																
RemoteCLIP [9]	ViT-B	50.28	42.32	45.27	39.33	37.78	50.26	48.46	36.61	50.06	46.48	46.91	48.39	49.63	45.12	10.26
RemoteCLIP [9]	ViT-L	56.69	54.51	52.73	43.24	51.19	50.82	45.12	50.47	51.82	53.68	38.98	54.59	56.53	50.31	11.25
GeoRSCLIP [10]	ViT-B	51.44	42.89	46.41	40.37	38.64	51.35	49.79	38.56	51.15	47.89	48.24	49.28	50.87	46.29	**10.01**
GeoRSCLIP [10]	ViT-L	56.81	54.97	52.53	43.28	41.37	50.49	42.7	49.41	51.36	53.54	36.98	54.66	56.64	48.99	13.77

3.4 Object Detection

We evaluate robustness on the corrupted images from the DIOR dataset. We report mAP for MIM- and CL-based methods. We omit LLM-based methods for the object detection task due to the lack of open-source LLMs for this task. From Table 3, we can draw the following findings: 1) both MIM- and CL-based methods suffer from serious performance under image corruptions, with a mAP drop of more than 8%. 2) MIM- and CL-based methods achieve comparable performance for object detection on clean and noisy images. RVSA [46] attains the highest mAP of 70.96% on clean images but experiences the most severe performance decline under image corruptions. 3) MIM-based and CL-based methods exhibit a similar degree of performance degradation when subjected to image corruptions. Among them, SatMAE++ [47] demonstrates the most robust detection performance under noisy conditions.

Table 3: Object detection performance (mAP) on the DIOR dataset across different image perturbations.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Rotate	Salt Pepper	Scale	Translate	Avg	$\mathcal{R}_{\text{TP}}$
MIM-based																
SatMAE [5]	ViT-L	62.30	56.84	57.86	55.80	58.36	55.38	58.44	59.34	56.92	56.60	53.76	51.58	60.90	56.82	8.81
ScaleMAE [7]	ViT-L	70.20	64.80	65.98	62.50	64.46	62.58	**63.82**	66.10	**63.08**	63.44	**60.50**	53.08	68.26	63.22	9.94
RVSA [46]	ViT-B	**70.96**	60.59	65.02	61.58	64.60	62.35	62.87	63.98	62.88	**64.04**	56.61	55.97	**69.69**	62.51	11.91
SatMAE++ [47]	ViT-L	65.20	59.44	61.02	60.30	59.88	59.66	61.06	61.72	59.56	59.14	58.64	48.48	64.70	59.47	**8.79**
CL-based																
RemoteCLIP [9]	ViT-B	60.40	56.72	56.28	56.78	54.56	53.68	57.36	55.90	53.42	54.54	54.40	44.92	59.72	54.86	9.17
RemoteCLIP [9]	ViT-L	70.20	**66.52**	**66.62**	63.84	**65.40**	**63.62**	63.68	**66.76**	62.66	63.52	59.16	**57.42**	68.64	63.99	8.85
GeoRSCLIP [10]	ViT-B	60.20	56.28	56.04	56.08	55.46	53.38	56.92	55.50	53.38	53.98	53.48	46.98	59.32	54.73	9.09
GeoRSCLIP [10]	ViT-L	69.80	66.12	65.34	**65.34**	64.96	**63.62**	62.90	66.04	62.02	62.68	56.04	57.40	68.10	**63.38**	9.20

3.5 Image Captioning

Table 4 presents the zero-shot image captioning performance of GeoChat [12], SkySenseGPT [53], VHM [52], RS-LLaVA [51], and the recently introduced Falcon model [54]. Following the VRSBench protocol [26], caption quality is evaluated using the GPT-4-based CLAIR metric [58][3]. Given that geometric distortions—such as rotation, scaling, and translation—can substantially alter image content, we exclude performance measurements under these corruption conditions for image captioning, VQA, and visual grounding tasks.

As shown in the upper part of Table 4, all models experience performance degradation under noisy conditions. Among them, the Falcon [54] model achieves the best overall performance, significantly outperforming other methods on both clean and corrupted images. However, it also suffers the largest performance drop of 6.28%. In contrast, RS-LLaVA [51] demonstrates the strongest robustness to image corruptions, exhibiting the smallest decrease in CLAIR score, with only a 2.03% drop. Additionally, we present results for GeoChat [12], fine-tuned on the VRSBench training set, as shown in the lower part of Table 4. The fine-tuned GeoChat model on the target dataset exhibits significantly improved performance compared to its zero-shot counterpart, with similar performance drop under corruptions.

Table 4: Image captioning performance (CLAIR) on the VRSBench-Cap dataset across different image perturbations. *ft* denotes models trained on the VRSBench training set.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Salt Pepper	Avg	$\mathcal{R}_{\text{TP}}$
GeoChat [12]	ViT-L	41.39	40.06	40.45	37.65	40.20	39.76	38.48	40.38	39.92	37.61	39.59	4.35
SkySenseGPT [53]	ViT-L	48.29	47.21	46.64	44.22	46.25	45.52	44.97	46.14	45.13	44.36	45.60	5.57
VHM [52]	ViT-L	52.02	50.19	50.82	50.26	50.57	51.22	50.46	50.39	50.72	49.48	50.46	3.00
RS-LLaVA [51]	ViT-L	51.30	51.15	50.43	51.78	50.54	52.01	47.84	50.57	49.88	48.12	50.26	**2.03**
Falcon [54]	DaViT-B	**61.90**	**59.98**	**60.09**	**57.13**	**59.48**	**57.43**	**56.31**	**59.85**	**59.94**	**51.83**	**58.01**	6.28
GeoChat$_{ft}$ [12]	ViT-L	71.26	69.00	68.93	66.60	69.45	68.63	67.83	69.98	69.02	63.87	68.15	4.36

3.6 Visual Question Anaswering

Table 5 reports VQA performance across various image perturbations. Following the VRSBench protocol [26], VQA performance is evaluated using the GPT-4-based matching accuracy[4]. From

[3] We use the `gpt-4o-mini-2024-07-18` model to compute the CLAIR scores.

[4] We use the `gpt-4o-mini-2024-07-18` model to compute the matching accuracy for VQA.

Table 5, it is evident that all LLM-based models experience a moderate decline in performance under image perturbations. Overall, VHM [52] achieves the best accuracy across both clean and noisy images. LHRS-Bot [50], RS-LLaVA [51], and Falcon [54], despite showing relatively lower overall accuracy, exhibit less sensitivity to image corruptions. Additionally, the GeoChat [12] fine-tuned on the VRSBench training set surpasses zero-shot models in terms of absolute performance, while exhibiting a slightly smaller performance drop under perturbations, indicating improved robustness.

Table 5: VQA performance (Accuracy) on the VRSBench-VQA dataset across different image perturbations. *ft* indicates models fine-tuned on the VRSBench training set.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Salt Pepper	Avg	$\mathcal{R}_{\text{TP}}$
GeoChat [12]	ViT-L	56.63	53.89	54.82	55.14	55.99	55.88	55.44	56.22	54.08	54.04	55.06	2.77
LHRS-Bot [50]	ViT-L	35.72	35.72	35.69	35.72	35.72	35.72	35.72	35.72	35.34	35.72	35.56	**0.45**
SkySenseGPT [53]	ViT-L	60.21	59.26	59.73	57.93	59.64	59.21	58.27	59.63	59.17	57.27	58.90	2.18
VHM [52]	ViT-L	**61.72**	**60.91**	**61.07**	**60.40**	**61.49**	**60.91**	**60.91**	**61.12**	**59.97**	**60.39**	**60.90**	1.33
RS-LLaVA [51]	ViT-L	57.25	57.04	57.14	55.45	57.25	57.14	55.97	57.21	55.25	55.82	56.47	1.36
Falcon [54]	DaViT-B	33.27	32.83	32.70	32.19	33.30	33.43	32.85	32.76	32.97	31.55	32.73	1.59
GeoChat$_{ft}$ [12]	ViT-L	75.79	75.13	74.97	73.84	75.63	74.89	74.46	75.43	74.76	72.77	74.65	1.50

3.7 Visual Grounding

Table 6 presents the zero-shot visual grounding performance of comparing methods. We report grounding accuracy at an IoU threshold of 0.5. As shown in the upper part of Table 6, all methods experience noticeable declines in performance under image perturbations. The GeoGround [59] model achieves the best performance on both clean and perturbed images, with a grounding accuracy of 75.93% and the smallest drop of 4.48%. In comparison, the Falcon [54] model, despite not being trained on VRSBench, demonstrates competitive visual grounding capability, but experiences a more significant degradation in performance when exposed to image corruptions. The fine-tuned GeoChat [12] model shows substantial improvements over its zero-shot counterpart, with a substantially reduced performance drop under noisy conditions.

Table 6: Visual grounding performance on the VRSBench-Ref dataset across different image perturbations. We report grounding accuracy at an IoU threshold of 0.5. * indicates the GeoGround model includes VRSBench in its training data. *ft* indicates models fine-tuned on the VRSBench training set.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Salt Pepper	Avg	$\mathcal{R}_{\text{TP}}$
GeoChat [12]	ViT-L	18.96	17.09	16.54	16.52	16.19	16.61	16.93	17.09	16.91	16.57	16.72	11.81
VHM [52]	ViT-L	37.20	34.66	35.29	34.18	35.48	35.01	35.78	35.54	32.21	34.58	34.74	6.61
GeoGround* [59]	ViT-L	**75.93**	**73.57**	**71.57**	**71.30**	**72.23**	**73.23**	**72.92**	**74.06**	**72.11**	**71.77**	**72.53**	**4.48**
Falcon [54]	DaViT-B	73.30	71.31	69.92	65.83	68.61	70.79	64.28	71.04	68.17	59.53	67.72	7.61
GeoChat$_{ft}$ [12]	ViT-L	55.50	53.79	52.20	50.51	53.06	53.11	51.57	54.23	52.99	49.82	52.36	5.66

4 Discussion

In this section, we further analyze the robustness of EOFMs across model architectures, tasks, corruption categories, and backbone sizes.

4.1 Vision-Centric vs. Vision-Language Foundation Models

As shown in Fig. 2, vision-centric foundation models (MIM-based) tend to suffer greater performance degradation under visual perturbations compared to vision-language models (CL- and VLM-based). This difference is especially pronounced in image-level scene classification tasks, where MIM-based models exhibit an average performance drop exceeding 25%. In contrast, vision-language models consistently demonstrate stronger robustness across tasks, maintaining performance drops below 10% in most cases. This is probably due to the complementary grounding effect of language supervision. We also note that the robustness gap between vision-centric and vision-language models is less significant for segmentation and detection tasks.

4.2 Robustness Across Different Tasks

Fig. 2 further highlights that vulnerability to perturbations varies substantially across tasks. MIM-based models are particularly sensitive in classification tasks, while CL-based models maintain greater stability across classification, segmentation, and detection tasks. This can be attributed to the contrastive objective, which encourages learning of invariant and robust representations. LLM-based models, on the other hand, show the smallest performance degradation in vision-language tasks such as image captioning and visual question answering (VQA)—typically below 5%. These results suggest that LLM-based methods excel in corruption-robust generalization, particularly in tasks that benefit from multimodal alignment.

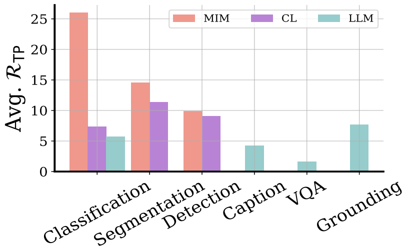


Figure 2: Robustness across different tasks and model architectures. We report the average $\mathcal{R}_{\text{TP}}$ across models.

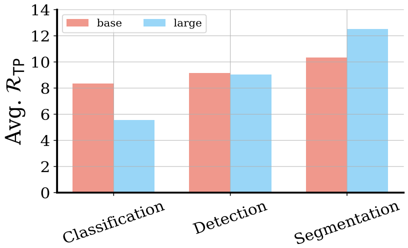


Figure 3: Robustness across different backbone sizes. We report the average $\mathcal{R}_{\text{TP}}$ for RemoteCLIP and GeoRSCLIP.

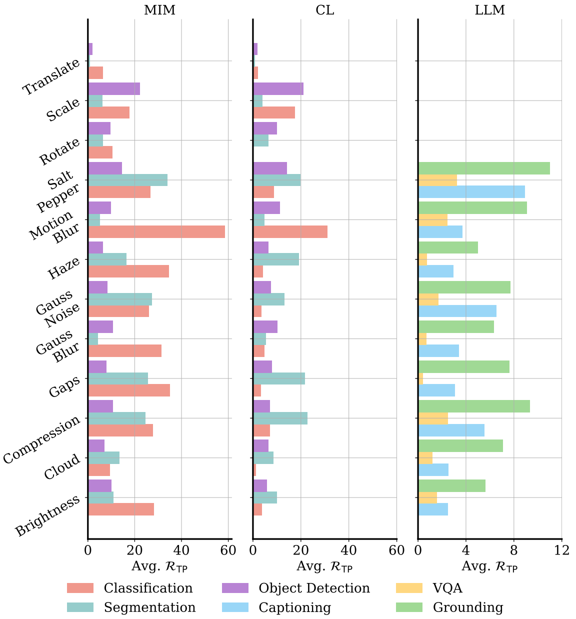


Figure 4: Robustness across different types of corruptions. We report the $\mathcal{R}_{\text{TP}}$ across models.

4.3 Robustness Across Different Backbone Sizes

For scene classification, the larger backbone (ViT-L) demonstrates greater robustness, showing less performance degradation under image corruptions. However, for fine-grained tasks such as semantic segmentation and object detection, ViT-L suffers larger performance drops compared to ViT-B. This suggests that while a larger backbone may enhance robustness in high-level recognition tasks, it may also amplify sensitivity to image corruptions in pixel- or region-level tasks.

4.4 Robustness Across Perturbation Types

As shown in Fig. 4, the performance degradation of MIM-, CL-, and LLM-based models varies notably across different visual perturbations. Motion blur causes the most severe drop, especially for MIM, which loses around 60% in performance, indicating a high sensitivity to spatial distortions. In contrast, translation has the least impact, suggesting minimal disruption to pattern recognition. Across perturbation types, LLM-based models consistently exhibit the strongest robustness, maintaining performance drops below 10% in most cases. This reinforces the value of language supervision in promoting the learning of more semantic and perturbation-invariant features.

4.5 Robustness Across Compound Perturbations

Real-world images often suffer from multiple simultaneous degradations (e.g., haze combined with noise). To investigate this, we evaluate object detection performance under selected compound perturbations on the DIOR-R dataset. The results from Table 7 reveal that models exhibit substantially lower accuracy under compound perturbations compared to single perturbations. For example, the detection performance of RemoteCLIP drops from 56.72% under Brightness to 40.58% under all three combined perturbations. Moreover, certain combinations of perturbations, such as *Brightness + Compression*, produce synergistic effects, causing performance declines that exceed the sum of individual perturbation effects.

Table 7: Object Detection performance (mAP) on DIOR-R under compound corruptions

Model	Clean	Brightness	Clouds	Compression	Brightness + Clouds	Brightness + Compression	Clouds + Compression	All Three
Brightness	✗	✓	✗	✗	✓	✓	✗	✓
Clouds	✗	✗	✓	✗	✓	✗	✓	✓
Compression	✗	✗	✗	✓	✗	✓	✓	✓
RemoteCLIP	60.40	56.72	56.28	56.78	49.98	45.36	52.57	40.58
GeoRSCLIP	60.20	56.28	56.04	56.08	50.27	44.90	52.39	40.94

4.6 Robustness in Multispectral Remote Sensing

In addition to high-resolution RGB imagery, multispectral data provide complementary spectral information that can improve scene understanding and robustness in Earth Observation. To assess the robustness of EO foundation models beyond optical imagery, we conduct preliminary experiments on two representative multispectral datasets: fMoW-Sentinel2 [60] and BigEarthNet [61], using two recently proposed foundation models, SatMAE [5] and EarthDial [62]. The results, summarized in Table 8, show substantial performance drops under image perturbations, underscoring the brittleness of current EO foundation models when applied to multispectral data.

Table 8: Scene classification performance on the multispectral dataset across different image perturbations.

Method	Backbone	Clean	Brightness Contrast	Cloud	Compression Artifacts	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Rotate	Salt Pepper	Scale	Translate	Avg	$\mathcal{R}_{\text{TP}}$
SatMAE [5]	fMoW-S2 [60]	59.75	37.46	58.69	33.58	50.01	29.35	36.64	41.57	40.12	38.93	51.79	43.35	59.68	43.43	27.31
EarthDial [62]	BigEarthNet [61]	46.521	31.47	46.48	45.40	34.97	36.37	38.69	33.25	39.84	29.51	22.71	39.18	45.30	36.93	20.62

5 Related Works

Robustness Research in Remote Sensing. Deep learning (DL)-based methods have achieved significant success in remote sensing image processing; however, their black-box nature raises concerns regarding interpretability, transparency, and vulnerability to adversarial examples. Recent studies have begun to address the robustness of DL models in this domain. Kazmi et al.[20] present a comprehensive literature review on adversarial attacks in aerial imagery processing, but do not provide an in-depth analysis of model robustness. Mei et al.[21] examine the robustness of DL-based methods for remote sensing image understanding, with a focus on image classification and object detection tasks. Lian et al. [63, 22] propose techniques to enhance adversarial robustness specifically for object detection in aerial imagery. [64] aims to improve the adversarial robustness of scene classification models in remote sensing via CAM-guided feature learning. These works only study the robustness of task-specific models. In contrast, our work for the first time investigates the robustness of foundation models in remote sensing.

Foundation Models in Remote Sensing. In general, there are four types of foundation models in remote sensing: MIM-based, CL-based, LLM-based, and diffusion-based methods. (1) Masked Image Modeling (MIM) has gained popularity through the pioneering work of MAE [1]. These methods typically employ an encoder network to learn feature representations by masking a portion of the image tokens, followed by a decoder network that reconstructs the masked image pixels in a self-supervised manner. In the field of remote sensing, notable approaches include SatMAE [5], RingMo [6], RVSA [46], ScaleMAE [7], SatMAE++ [47], and DOFA [65]. (2) Contrastive Learning (CL) employs separate encoders to project images and texts into a shared embedding space, using a contrastive objective to align the resulting embeddings. Building on the success of the pioneering work CLIP [2], several contrastive learning-based foundation models have been introduced in the field of remote sensing, including RS-CLIP [10], RemoteCLIP [9], GeoRSCLIP [10], SkyCLIP [66], S-CLIP [67], SatCLIP [68], and GeoCLIP [69]. (3) Following the pioneering works of MiniGPT-4 [3] and LLaVA [4], multimodal large language models (MLLMs) have attracted significant research attention in recent years. For instance, RSGPT [11] introduces the first GPT-based MLLM tailored for remote sensing image understanding. Other notable approaches include GeoChat [12], EarthGPT [70], EarthMarker [71], Popeye [72], RS-LLaVA [51], VHM [52], LHRS-Bot [50], SkyEyeGPT [73], SkySenseGPT [53], RSUniVLM [74], and Falcon [54]. (4) Diffusion-based foundation models learn the joint distribution between text prompts and images through a forward noising process followed by a reverse denoising process. Recent studies have applied these models to synthesize satellite [75, 76], aerial [77], hyperspectral [78], and multi-resolution imagery [79].

6 Conclusion and Future Work

In this work, we present REOBench, the first comprehensive benchmark for evaluating the robustness of EOFMs across six core tasks and twelve perturbation types. Our evaluation reveals that existing EOFMs experience noticeable performance degradation under image corruptions. We also observe significant variations in robustness across model types, task categories, and backbone sizes, offering valuable insights for future development of robust models. We hope REOBench will serve as a standard benchmark to drive the creation of more robust and reliable models for Earth observation.

Despite its contributions, this work has several limitations. First, the evaluation is limited to high-resolution optical imagery, excluding other key modalities such as multispectral (e.g., Sentinel-2), hyperspectral, and SAR data. Second, the benchmark's dataset and task coverage are not exhaustive. While it includes widely used datasets (AID, Potsdam, DIOR, VRSBench), they may not fully reflect global variation in geography, resolution, or sensor types. Additionally, important tasks such as change detection, region captioning, and object counting are currently not included.

7 Broader Impact

REOBench aims to improve the reliability of Earth observation foundation models by systematically evaluating their robustness to real-world noise and perturbations. This is critical for high-stakes applications such as disaster response and environmental monitoring. By identifying vulnerability patterns across tasks and models, our benchmark can guide the development of future robust models.

Acknowledgments and Disclosure of Funding

This work is supported by the EDF-KOIOS project and used the Dutch national e-infrastructure with the support of the SURF Cooperative using the funding of the projects EINF-12538, EINF-10925 and NWO-2023.027. We would like to express our deepest gratitude to the anonymous reviewers whose insightful comments and suggestions significantly improved the quality of this paper.

References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[3] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[5] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

[6] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2022.

[7] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

[8] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. SpectralGPT: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[9] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[10] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. RS5M and GeoRSCLIP: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[11] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, Yu Liu, and Xiang Li. RSGPT: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224:272–286, 2025.

[12] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024.

[13] Licheng Jiao, Zhongjian Huang, Xiaoqiang Lu, Xu Liu, Yuting Yang, Jiaxuan Zhao, Jinyue Zhang, Biao Hou, Shuyuan Yang, Fang Liu, et al. Brain-inspired remote sensing foundation models and open problems: A comprehensive survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:10084–10120, 2023.

[14] Xiang Li, Congcong Wen, Yuan Hu, Zhenghang Yuan, and Xiao Xiang Zhu. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine*, 2024.

[15] Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of earth and climate foundation models. *arXiv preprint arXiv:2405.04285*, 2024.

[16] Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey. *arXiv preprint arXiv:2410.16602*, 2024.

[17] Siqi Lu, Junlin Guo, James R Zimmer-Dauphinee, Jordan M Nieusma, Xiao Wang, Steven A Wernke, Yuankai Huo, et al. Vision foundation models in remote sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

[18] Chunlei Huo, Keming Chen, Shuaihao Zhang, Zeyu Wang, Heyu Yan, Jing Shen, Yuyang Hong, Geqi Qi, Hongmei Fang, and Zihan Wang. When remote sensing meets foundation model: A survey and beyond. *remote sensing*, 17(2), 2025.

[19] Ziyue Huang, Hongxi Yan, Qiqi Zhan, Shuai Yang, Mingming Zhang, Chenkai Zhang, YiMing Lei, Zeming Liu, Qingjie Liu, and Yunhong Wang. A survey on remote sensing foundation models: From vision to multimodality. *arXiv preprint arXiv:2503.22081*, 2025.

[20] Syed M. Kazam Abbas Kazmi, Nayyer Aafaq, Mansoor Ahmad Khan, Ammar Saleem, and Zahid Ali. Adversarial attacks on aerial imagery: The state-of-the-art and perspective. *2023 3rd International Conference on Artificial Intelligence (ICAI)*, pages 95–102, 2023.

[21] Shaohui Mei, Jiawei Lian, Xiaofei Wang, Yuru Su, Mingyang Ma, and Lap-Pui Chau. A comprehensive study on the robustness of image classification and object detection in remote sensing: Surveying and benchmarking. *ArXiv*, abs/2306.12111, 2023.

[22] Jiawei Lian, Shaohui Mei, Shun Zhang, and Mingyang Ma. Benchmarking adversarial patch against aerial detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[23] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

[24] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences; I-3*, 1(1):293–298, 2012.

[25] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020.

[26] Xiang Li, Jian Ding, and Mohamed Elhoseiny. VRSBench: A versatile vision-language benchmark dataset for remote sensing image understanding. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[27] Jacob Høxbroe Jeppesen, Rune Hylsberg Jacobsen, Fadil Inceoglu, and Thomas Skjødeberg Toftegaard. A cloud detection algorithm for satellite imagery based on deep learning. *Remote sensing of environment*, 229:247–259, 2019.

[28] Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal from satellite images using spatiotemporal generator networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1796–1805, 2020.

[29] Liangpei Zhang, Gui-Song Xia, Tianfu Wu, Liang Lin, Xue-Cheng Tai, et al. Deep learning for remote sensing image understanding. *J. Sensors*, 2016(2):1–2, 2016.

[30] Thomas Müller and Bastian Erdnüeß. Brightness correction and shadow removal for video change detection with uavs. In *Autonomous systems: sensors, processing, and security for vehicles and infrastructure 2019*, volume 11009, page 1100906. SPIE, 2019.

[31] Aliaksei Makarau, Rudolf Richter, Rupert Müller, and Peter Reinartz. Haze detection and removal in remotely sensed multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(9):5895–5905, 2014.

[32] Yuanyuan Li, Qiying Ling, Yiyao An, Hongpeng Yin, Xinbo Gao, Zhiqin Zhu, and Peng Han. DHC-Net: A remote sensing object detection under haze and class imbalance. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[33] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[34] Till Sieberth, Rene Wackrow, and JH Chandler. UAV image blur–its influence and ways to correct it. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40:33–39, 2015.

[35] Till Sieberth, Rene Wackrow, and JH Chandler. Influence of blur on feature matching and a geometric approach for photogrammetric deblurring. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40:321–326, 2014.

[36] T Sieberth, R Wackrow, and JH Chandler. Motion blur disturbs–the influence of motion-blurred images in photogrammetry. *The Photogrammetric Record*, 29(148):434–453, 2014.

[37] Sornkitja Boonprong, Chunxiang Cao, Wei Chen, Xiliang Ni, Min Xu, and Bipin Kumar Acharya. The classification of noise-afflicted remotely sensed data using three machine-learning techniques: effect of different levels and types of noise on accuracy. *ISPRS International Journal of Geo-Information*, 7(7):274, 2018.

[38] Ram M Narayanan, Sudhir K Ponnappan, and Stephen E Reichenbach. Effects of noise on the information content of remote sensing images. *Geocarto International*, 18(2):15–26, 2003.

[39] Jin Chen, Xiaolin Zhu, James E Vogelmann, Feng Gao, and Suming Jin. A simple and effective method for filling gaps in landsat ETM+ SLC-off images. *Remote sensing of environment*, 115(4):1053–1064, 2011.

[40] Feng Chen, Xiaofeng Zhao, Hong Ye, and Zdravko Karakehayov. Making use of the landsat 7 SLC-off etm+ image through different recovering approaches. *Data Acquisition Applications*, pages 317–342, 2012.

[41] Tajeddine Benbarrad, Lamiae Eloutouate, Mounir Arioua, Fatiha Elouaai, and My Driss Laanaoui. Impact of image compression on the performance of steel surface defect classification with a CNN. *Journal of Sensor and Actuator Networks*, 10(4):73, 2021.

[42] Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3D object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1022–1032, 2023.

[43] Jian Kang, Ruben Fernandez-Beltran, Zhirui Wang, Xian Sun, Jingen Ni, and Antonio Plaza. Rotation-invariant deep embedding for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.

[44] Wei Han, Jun Li, Sheng Wang, Yi Wang, Jining Yan, Runyu Fan, Xiaohan Zhang, and Lizhe Wang. A context-scale-aware detector and a new benchmark for remote sensing small weak object detection in unmanned aerial vehicle images. *International Journal of Applied Earth Observation and Geoinformation*, 112:102966, 2022.

[45] Wooju Lee, Donggyu Sim, and Seoung-Jun Oh. A CNN-based high-accuracy registration for remote sensing images. *Remote Sensing*, 13(8):1482, 2021.

[46] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022.

[47] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27811–27819, 2024.

[48] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[49] OpenAI. GPT-4 technical report, 2023.

[50] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. LHRS-Bot: Empowering remote sensing with VGI-enhanced large multimodal language model. In *European Conference on Computer Vision*, pages 440–457. Springer, 2024.

[51] Yakoub Bazi, Laila Bashmal, Mohamad Mahmoud Al Rahhal, Riccardo Ricci, and Farid Melgani. RS-LLaVA: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9):1477, 2024.

[52] Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, et al. VHM: Versatile and honest vision language model for remote sensing image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6381–6388, 2025.

[53] Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. SkySenseGPT: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

[54] Kelu Yao, Nuo Xu, Rong Yang, Yingying Xu, Zhuoyan Gao, Titinunt Kitrungrotsakul, Yi Ren, Pu Zhang, Jin Wang, Ning Wei, et al. Falcon: A remote sensing vision-language foundation model. *arXiv preprint arXiv:2503.11070*, 2025.

[55] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[56] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3520–3529, 2021.

[57] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023.

[58] David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. CLAIR: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646, 2023.

[59] Yue Zhou, Mengcheng Lan, Xiang Li, Yiping Ke, Xue Jiang, Litong Feng, and Wayne Zhang. GeoGround: A unified large vision-language model. for remote sensing visual grounding. *arXiv preprint arXiv:2411.11904*, 2024.

[60] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[61] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*, pages 5901–5904. IEEE, 2019.

[62] Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14303–14313, 2025.

[63] Jiawei Lian, Xiaofei Wang, Yuru Su, Mingyang Ma, and Shaohui Mei. CBA: Contextual background attack against optical aerial detection in the physical world. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

[64] Sumaiya Tasneem and Kazi Aminul Islam. Improve adversarial robustness of ai models in remote sensing via data-augmentation and explainable-ai methods. *Remote. Sens.*, 16:3210, 2024.

[65] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

[66] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. SkyScript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5805–5813, 2024.

[67] Sangwoo Mo, Minkyu Kim, Kyungmin Lee, and Jinwoo Shin. S-CLIP: Semi-supervised vision-language learning using few specialist captions. In *Neural Information Processing Systems*, 2023.

[68] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. SatCLIP: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4347–4355, 2025.

[69] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023.

[70] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. EarthGPT: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[71] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, Jun Li, and Xuerui Mao. EarthMarker: A visual prompting multi-modal large language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[72] Wei Zhang, Miaoxin Cai, Tong Zhang, Guoqiang Lei, Yin Zhuang, and Xuerui Mao. Popeye: A unified visual-language model for multi-source ship detection from remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[73] Yang Zhan, Zhitong Xiong, and Yuan Yuan. SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77, 2025.

[74] Xu Liu and Zhouhui Lian. RSUniVLM: A unified vision language model for remote sensing via granularity-oriented mixture of experts. *arXiv preprint arXiv:2412.05679*, 2024.

[75] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. DiffusionSat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024.

[76] Chenyang Liu, Ke-Yu Chen, Ruiyun Zhao, Zhengxia Zou, and Zhen Xia Shi. Text2Earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. *ArXiv*, abs/2501.00895, 2025.

[77] Ahmad Arrabi, Xiaohan Zhang, Waqas Sultani, Chen Chen, and Safwan Wshah. Cross-view meets diffusion: Aerial image synthesis with geometry and text guidance. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5356–5366, 2024.

[78] Li Pang, Datao Tang, Shuang Xu, Deyu Meng, and Xiangyong Cao. HSIGene: A foundation model for hyperspectral image generation. *ArXiv*, abs/2409.12470, 2024.

[79] Zhiping Yu, Chenyang Liu, Liqin Liu, Zhen Xia Shi, and Zhengxia Zou. MetaEarth: A generative foundation model for global-scale remote sensing image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:1764–1781, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See the Abstract and Introduction sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the Conclusion and Future Work section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical results in this study.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details on how to use foundation models for downstream tasks in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release our dataset at Huggingface and code at GitHub.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 3.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report the average performance drop under image corruptions in our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include details about computing resources in the supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics in this study.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Broader Impact section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We require users to adhere to the usage guidelines of our dataset.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The source datasets used in this study, including AID, ISPRS Potsdam 2D Labeling, and VRSBench, are properly cited and available online as open-access resources and are free for academic use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide detailed and well-structured documentation of our dataset in the supplementary and the project page.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs are not used as an important, original, or non-standard component of the core methods.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.