Visual Instruction Pretraining for Domain-Specific Foundation Models

Yuxuan Li[1] Yicheng Zhang[1] Wenhao Tang[1]

Yimian Dai[1] Ming-Ming Cheng[1,2] Xiang Li[1,2†] Jian Yang[1†]

[1] PCA Lab, VCIP, Computer Science, NKU [2]NKIARI, Futian, Shenzhen
yuxuan.li.17@ucl.ac.uk, {whtang, zhangyc}@mail.nankai.edu.cn
{yimian.dai, cmm, xiang.li.implus, csjyang}@nankai.edu.cn

Abstract

*Modern computer vision is converging on a closed loop in which perception, reasoning and generation mutually reinforce each other. However, this loop remains incomplete: the top-down influence of high-level reasoning on the foundational learning of low-level perceptual features is not yet underexplored. This paper addresses this gap by proposing a new paradigm for pretraining foundation models in downstream domains. We introduce **V**isual **insT**ruction **P**retraining (ViTP), a novel approach that directly leverages reasoning to enhance perception. ViTP embeds a Vision Transformer (ViT) backbone within a Vision-Language Model and pretrains it end-to-end using a rich corpus of visual instruction data curated from target downstream domains. ViTP is powered by our proposed Visual Robustness Learning (VRL), which compels the ViT to learn robust and domain-relevant features from a sparse set of visual tokens. Extensive experiments on 16 challenging remote sensing and medical imaging benchmarks demonstrate that ViTP establishes new state-of-the-art performance across a diverse range of downstream tasks. The code is available at* `github.com/zcablii/ViTP`.

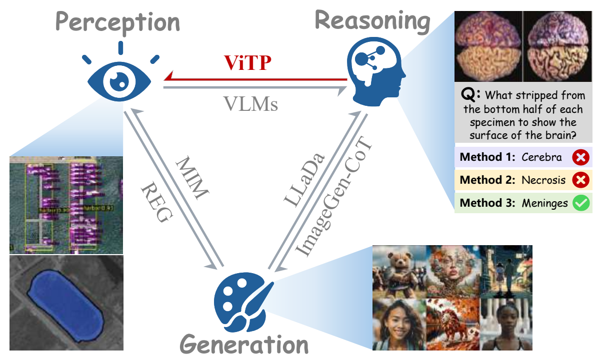


Figure 1. The synergistic relationship between perception, generation, and reasoning in modern CV. Our proposed ViTP forges a novel link from high-level reasoning to low-level perception, a previously underexplored connection.

1. Introduction

Recent years have witnessed a leap forward in computer vision (CV), largely catalyzed by the Transformer architecture [129]. This architecture has emerged as a unifying framework across the three cornerstone areas of CV: perception, generation, and reasoning. A synergistic relationship has become evident, where advancements in one area often precipitate progress in the others, as illustrated in Figure 1. For instance, generative pretraining methods like Masked Image Modeling (MIM) [6, 50] have significantly boosted the performance of downstream perception tasks. Conversely, perception-centric features from models like DINOv2 [107] have been leveraged to guide and accelerate generative diffusion model training [142]. Such interplay is also demonstrated by the modern Vision-Language Models (VLMs) [3, 21, 59], ImageGen-CoT [80] and LLaDa [105]. However, a crucial link remains underexplored: the top-down influence of high-level reasoning on the foundational learning of low-level perceptual features. This raises a compelling question: can abstract reasoning, which often demands highly discriminative visual features, be harnessed to guide a perception model toward learning more precise and efficient representations?

Answering this question is particularly critical for pretraining powerful foundation models in specialized domains like remote sensing and medical imaging, where prevailing paradigms face formidable challenges. Prevailing pretraining paradigms include supervised classification, MIM [6, 50, 111], image-image contrastive learning [41, 49, 107], and image-text contrastive learning [87, 109, 168]. Each

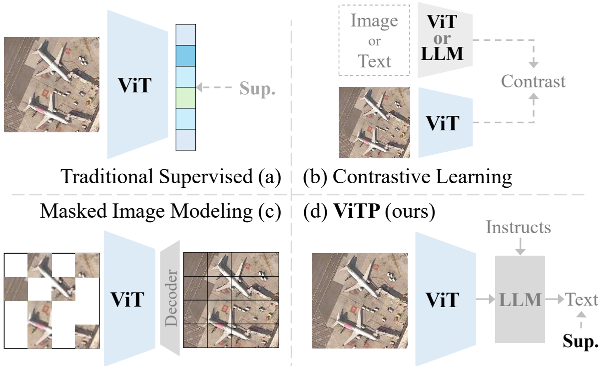


Figure 2. Comparison of pretraining paradigms for ViT foundation models. ViTP employs an instruction-following objective to directly instil domain-specific perception capabilities into the vision backbone.

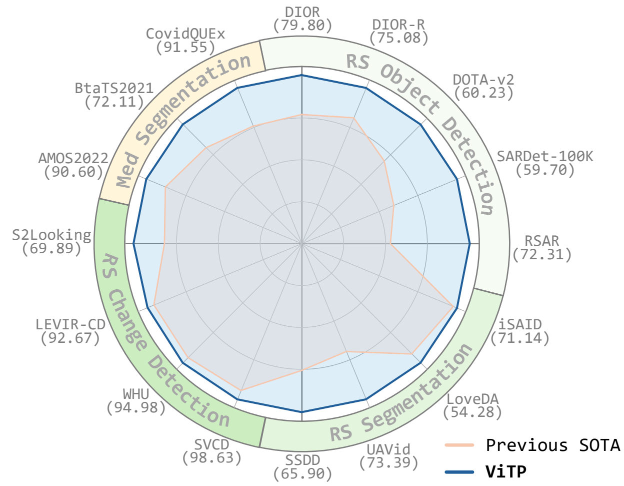


Figure 3. ViTP sets new SOTA performance across a diverse range of downstream tasks in medical imaging and remote sensing.

approach, however, presents notable drawbacks. Supervised pretraining risks overfitting to narrow label distributions, thereby limiting generalization. MIM, by focusing on pixel-level reconstruction, may inadvertently neglect the fine-grained details of small but critical objects, a common scenario in remote sensing and medical scans. Image-image contrastive methods, while effective, are often notoriously difficult to optimize and demand substantial computational resources. These paradigms are all fundamentally "bottom-up": they operate on the premise that robust low-level perception is a prerequisite for high-level reasoning [83, 92]. While this hierarchy is well-established, the reciprocal pathway (i.e. whether high-level reasoning can directly enhance low-level feature learning) remains largely unexplored.

Image-to-text contrastive learning, exemplified by CLIP [109], represents an initial step toward reasoning-guided pretraining. However, these methods primarily aim for general-purpose image-text alignment rather than optimizing a vision backbone for specialized, fine-grained downstream tasks. Consequently, the global, image-level features learned by models like CLIP's ViT often prove suboptimal for dense prediction tasks such as semantic segmentation or object detection, which demand precise, pixel-level understanding [78]. This limitation underscores that merely aligning global image and text representations is an insufficient strategy for forging a powerful, domain-specific perception backbone.

To bridge this gap, we introduce **V**isual **i**ns**T**ruction **P**retraining (ViTP), a novel, top-down pretraining paradigm that directly integrates high-level, instruction-based reasoning into the perceptual feature learning process of a ViT backbone. As illustrated in Figure 2, ViTP diverges from prior methods by embedding the ViT within a larger Vision-Language Model (VLM) and pretraining it via a visual instruction-following objective. We construct our training data with image-text pairs derived directly from the target downstream tasks. During training, image tokens from the ViT and text tokens from a user query are fed jointly into a Large Language Model (LLM) [88]. The LLM's generated response serves as the supervisory signal, compelling the ViT backbone to learn the complex data distributions of the downstream domain in an end-to-end fashion. This process endows the ViT with highly relevant, domain-specific perceptual capabilities. Furthermore, we introduce **V**isual **R**obustness **L**earning (VRL) by randomly dropping a large fraction of the ViT's output image tokens before they are passed to the LLM. This constraint implicitly forces the ViT's attention mechanism to encode more comprehensive and robust information within each of the remaining tokens, thereby enhancing the robustness and semantic richness of the learned visual features.

Extensive experiments on 16 challenging remote sensing and medical imaging benchmarks validate the effectiveness and efficiency of our approach. As shown in Figure 3, ViTP achieves new state-of-the-art results across several tasks. Notably, the ViTP pretraining process is computationally efficient, requiring only one day on 8 A40 GPUs. This work not only presents a novel pretraining paradigm but also offers a promising solution for creating powerful, domain-adapted foundation models.

Our contributions are summarized as follows:

- We introduce ViTP, a top-down pretraining paradigm that leverages instruction following objectives to imbue a ViT backbone with high-level semantic perception tailored for specific downstream domains.
- We propose Visual Robustness Learning, a regularization method that encourages the ViT to learn more comprehensive and robust feature representations by operating on a sparse set of visual tokens.

- Extensive experiments demonstrate the pretraining efficiency and state-of-the-art performance of ViTP on downstream tasks in remote sensing and medical imaging.

2. Related Work

2.1. Perception Foundation Model

2.1.1. General Domains

Early approaches predominantly relied on supervised pretraining over large-scale labeled datasets such as ImageNet [34, 47]. While effective in learning semantically meaningful representations, these methods risk overfitting to the label space of the pretraining task, thereby limiting generalization to diverse downstream applications. Consequently, the field has increasingly shifted toward unsupervised learning paradigms that leverage unlabeled data. Among these, contrastive learning and masked image modeling (MIM) emerged as two dominant strategies. Contrastive learning methods, including MoCo [49], BYOL [39], and DINOv2 [107], learn discriminative representations by minimizing the distance between positive pairs (e.g., augmented views of the same image) while maximizing the distance between negative pairs in the embedding space. These methods emphasize semantic invariance and shown strong generalization capabilities across tasks. On the other hand, MIM-based approaches such as MAE [50], SimMIM [147] and UM-MAE [70] adopt a generative pretext task: a large portion of the input image is masked, and the model is trained to reconstruct the missing content. This encourages the model to learn holistic, contextually rich visual features by capturing spatial dependencies and fine-grained details.

While both paradigms achieved impressive results on natural images, their direct transfer to specialized domains such as remote sensing or medical imagery often results in suboptimal performance [27, 42, 87]. This underscores the necessity for domain-specific adaptations in pretraining strategies.

2.1.2. Remote sensing

In the remote sensing domain, pretraining strategies are tailored to address unique challenges such as large scale variations, multi-spectral imagery, and domain-specific object categories. Supervised methods like SAMRS [133] and MSFA [76] achieve strong performance by leveraging annotated remote sensing datasets. To exploit the abundance of unlabeled data, MIM-based approaches such as RingMo [119], SatMAE [27], and Scale-MAE [111] are proposed, specifically designed to model the dense small objects, multi-modal or multi-scale characteristics of satellite imagery. Meanwhile, contrastive learning has been explored through both image-text and image-image formulations. Models like GeoRSCLIP [168] and RemoteCLIP [87] leverage large-scale image-text pairs to enable zero-shot classification capabilities. Image-based contrastive methods such as CACo [100] and Skysense [41] focus on learning discriminative features from unlabeled remote sensing imagery. Hybrid approaches like CMID [103] and GFM [101] further integrate the strengths of both MIM and contrastive learning to learn more comprehensive and robust representations.

2.1.3. Medical Images

Pretraining vision foundation models is particularly crucial in medical imaging, where labeled data is scarce and expensive to obtain. A prominent line of research focuses on adapting the general-purpose Segment Anything Model (SAM) [60] to the medical domain. Adaptations range from finetuning SAM on large-scale medical datasets, as in MedSAM [99], to pretraining SAM-like architectures from scratch on medical images, as demonstrated by SAM-Med2D [24]. Other frameworks, such as IMIS-SAM [26], aim to enhance the interactive capabilities of SAM for clinical workflows. Concurrently, specialized MIM pretraining methods like MedMAE [42] and S3D [131] are developed to better capture the unique characteristics of medical data.

A common limitation across these domain-specific foundation models is their heavy reliance on tailored architectural or training designs, which restricts their transferability to other domains. In contrast, our proposed ViTP framework offers a more generalizable pretraining strategy that can be seamlessly adapted to various downstream domains by simply curating the corresponding visual instruction datasets.

2.2. Continual Pretraining

Continual pretraining was first popularized in NLP, where Gururangan et al. [43] demonstrated that continuing pretraining on in-domain data significantly improves model performance. This concept is also successfully adapted to computer vision. For instance, CSPT [165] first pretrains on ImageNet and then continues pretraining on a target remote sensing dataset using an MIM objective. TOV [123] adopts a curriculum strategy, freezing early layers pretrained on natural images while finetuning deeper layers on specialized data. To reduce computational overhead, RemoteCLIP [87] initializes from pretrained CLIP [109] weights and continues pretraining on domain-specific data. Similarly, MSFA [76] adopts a multi-stage strategy: pretraining on ImageNet, followed by training on optical remote sensing detection datasets, and finally finetuning on SAR detection data.

Inspired by these successes, we adopt a continual pretraining approach, initializing our VLM from the well-established InternVL [19] model. However, our method differs from prior work in two key aspects: (1) to mitigate catastrophic forgetting of general visual knowledge, we incorporate a small proportion of general-domain data into the specialized training corpus; (2) we introduce Visual Robustness Learning as an additional regularizer to enhance the model's robustness during continual pretraining.

2.3. Visual Instruction Tuning

The rise of powerful VLMs such as Gemini [124], InternVL [20], and Qwen-VL [3] are largely driven by visual instruction tuning [88]. This paradigm aims to align visual representations with a pretrained Large Language Model (LLM) by projecting image features into the LLM's embedding space. The model is then finetuned on a corpus of image-text instruction pairs, enabling the LLM to interpret visual inputs and respond to textual instructions. In this setup, the vision encoder typically remains frozen, serving solely as a feature extractor, while all learning occurs in the LLM and the projection module.

While this paradigm effectively leverages perception to enhance reasoning, the reverse direction, i.e. using reasoning to improve perception, remains largely underexplored. Our proposed ViTP addresses this gap by inverting the traditional visual instruction tuning pipeline. Instead of using images to tune an LLM, we utilize an LLM to guide the pretraining of the vision encoder.

3. Method

In this section, we present our visual instruction pretraining (ViTP) framework, a novel paradigm designed to pretrain a ViT backbone. As illustrated in Figure 2, ViTP fundamentally differs from traditional pretraining by leveraging the reasoning capabilities of modern VLMs. The entire process is driven by a "visual instruction following objective", where the ViT learns to extract features that help an LLM answer questions about an image. To tailor the model for specific downstream applications, we outline a "data recipe" for curating a domain-specific pretraining dataset. During pretraining, we employ "Visual Robustness Learning (VRL)", a regularization technique that drops image tokens to enhance the semantic robustness of the learned representations. The final pretrained ViT serves as a powerful foundation model, readily adaptable to a variety of downstream tasks.

3.1. Visual Instruction Following Objective

The central hypothesis of ViTP is that a ViT can learn more potent and relevant features if its training is guided by high-level reasoning. To achieve this, we frame the pretraining as a visual instruction-following task. The framework, depicted in Figure 4, processes a domain-specific image through a ViT encoder to produce a sequence of image tokens. These tokens are projected into the LLM's embedding space and concatenated with the tokenized text of an instruction. A Large Language Model (LLM) then processes this combined sequence to generate a response. The entire model is trained end-to-end, allowing the supervisory signal from the LLM's response to directly optimize the ViT's feature extraction process.

Our approach employs a continual pretraining strategy [43], starting with a well-trained, general-purpose VLM. We then continue its training on our curated domain-specific datasets. This offers two primary advantages: 1) It leverages the VLM's vast pre-existing knowledge of general visual and linguistic patterns, providing a robust initialization. 2) This strong starting point significantly accelerates convergence during domain-specific pretraining, enhancing computational efficiency. Let the raw dataset be $\mathcal{D}_{\text{raw}} = \{(I, Q, R)\}$, where I is an image, Q is a text query (instruction), and R is the ground-truth text response. This is processed into the final training set $\mathcal{D} = \{(x_i, x_t, y^\star)\}$, where x_i, x_t, and $y^\star$ represent the processed image tokens, text tokens, and target response tokens, respectively.

3.1.1. ViT Feature Extraction and Projection

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the ViT backbone partitions it into a grid of non-overlapping patches. Each patch is linearly embedded, and these patch embeddings are prepended with a '[CLS]' token and processed through a series of Transformer blocks [129]. This yields a sequence of output image tokens $x_i' = \{t_1, t_2, \ldots, t_N\}$, where N is the sequence length. To align these visual tokens with the LLM's embedding space, a lightweight projection layer (e.g., a two-layer MLP) maps these visual tokens x_i' into the final image tokens x_i.

3.1.2. Instruction-following Token Concatenation

For each image, the corresponding text query Q is converted into a sequence of text tokens x_t using the LLM's tokenizer. These tokens represent the task-specific instruction. The projected image tokens x_i and text tokens x_t are then concatenated to form a unified input sequence. Crucially, learnable positional encodings are added to the embeddings of the image and text tokens to provide the LLM with spatial and sequential context. The final input sequence for the LLM is formed as:

$$S_{llm} = [\text{PE}(x_i)\,;\,\text{PE}(x_t)] \tag{1}$$

where $\text{PE}(\cdot)$ denotes the addition of positional encodings to the token embeddings and '[;]' signifies sequence concatenation.

3.1.3. LLM-based Supervision

The combined sequence S_{llm} is processed by the LLM, which acts as a reasoning engine to interpret the visual information from x_i in the context of the instruction x_t. The model then auto-regressively generates an output sequence

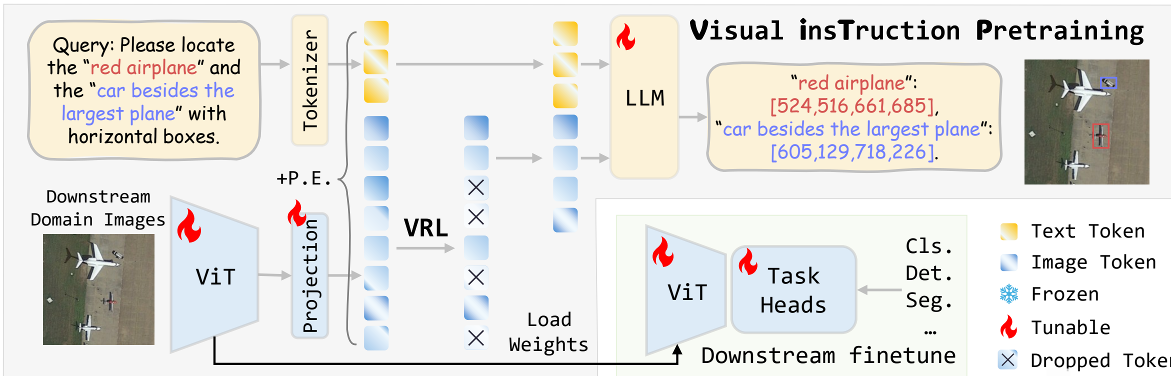


Figure 4. A conceptual illustration of the ViTP framework. A ViT backbone is embedded within a large VLM and then pretrained with domain-specific instruction following objective and Visual Robustness Learning (VRL). This process instils high-level semantic understanding into the ViT. The resulting weights are then used to initialize models for various downstream perception tasks.

O. The entire system is optimized by minimizing the discrepancy between the generated output O and the ground-truth target response $y^\star$. For instance, if the instruction asks to identify an object, $y^\star$ could be a textual description including its location in a structured format. During pretraining, the gradients from the output loss propagate back through the entire model, including the projection layer and the ViT backbone. We allow the weights of the ViT, the projection layer, and the LLM to be trainable. The optimization follows a standard supervised finetuning (SFT) objective, which minimizes the negative log-likelihood of the target sequence:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = \mathbb{E}\Big[-\log P_\theta\big(y^\star \,\big|\, S_{llm}\big)\Big], \tag{2}$$

where $(x_i, x_t, y^\star) \sim \mathcal{D}$ and P_θ is the probability distribution over the text sequences parameterized by the entire model θ.

3.2. Visual Robustness Learning

To foster the learning of more robust and semantically rich features, we introduce Visual Robustness Learning (VRL), a simple yet effective regularization technique applied during pretraining. As shown in Figure 4, VRL randomly drops a significant fraction of the projected image tokens x_i before they are concatenated with the text tokens x_t. This operation is performed after positional encodings are associated with the tokens, ensuring the LLM retains knowledge of the original spatial positions of the surviving tokens. The VRL objective is thus:

$$\mathcal{L}_{\mathrm{VRL}}(\theta) = \mathbb{E}\Big[-\log P_\theta\big(y^\star \,\big|\, [\mathcal{C}_r(\mathrm{PE}(x_i)); \mathrm{PE}(x_t)]\big)\Big], \tag{3}$$

where $\mathcal{C}_r$ is a random sampling operation that drops a proportion r of the tokens from a sequence. Formally, for a sequence $\mathcal{S}$:

$$\mathcal{C}_r(\mathcal{S}) \sim \big\{X \subseteq \mathcal{S} \,\big|\, |X| = \lceil (1-r)\cdot|\mathcal{S}|\rceil\big\}, \tag{4}$$

where the subset X is selected uniformly at random while preserving the original ordering, and $r \in [0, 1)$. This "torture" mechanism forces the ViT to encode more comprehensive information in each token, as the model must infer the full visual context from a partial input. It encourages the ViT's attention mechanism to learn robust, distributed and less redundant representations. As a practical benefit, dropping a large portion of tokens (e.g., $r = 0.75$) significantly reduces memory usage and accelerates computation, enhancing ViTP's scalability.

3.3. Pretraining Dataset Recipe

With the rapid development of VLMs, numerous image-text paired instruction datasets tailored for downstream domains have been released. The efficacy of ViTP is heavily dependent on the quality and composition of the pretraining dataset. We establish four key principles for constructing our data mixture:

1. **Scale and Diversity:** The dataset must be large and diverse, containing a wide array of visual concepts, scenes, and objects representative of the target domain.
2. **Modality Coverage:** The data must encompass all imaging modalities expected in downstream tasks. For instance, if a downstream task uses remote sensing Synthetic Aperture Radar (SAR) imagery, the pretraining mix should include such modality data to ensure the model learns modality-specific features.
3. **Task Capability Alignment:** The instruction-following tasks in the pretraining data should foster abilities required downstream. For instance, for downstream object detection tasks, including visual grounding and fine-grained VQA during pretraining empowers the backbone with localization and spatial reasoning capabilities.
4. **Preservation of Generality:** A certain fraction of general-domain natural images (e.g., from public VLM datasets) should be included. Domain-specific data can

Table 1. Overview of downstream task datasets for remote sensing, including object detection (horizontal and oriented), segmentation (semantic and instance), and change detection.

Task	Dataset	Modal	Box Format	Images	Classes
Object Detection	DIOR [63]	RGB	Hori. Box	23,463	20
	DIOR-R [25]	RGB	Ori. Box	23,463	20
	DOTA-v2.0 [143]	RGB	Ori. Box	11,268	18
	SARDet-100K [76]	SAR	Hori. Box	116,598	6
	SSDD [164]	SAR	Hori. Box	1,160	1
	RSAR [167]	SAR	Ori. Box	95,842	6
Semantic and Instance Segmentation	iSAID [140]	RGB	Mask	2,806	15
	LoveDA [136]	RGB	Mask	5,987	7
	UAVid [98]	RGB	Mask	5,510	8
	SSDD [164]	SAR	Polygons	1,160	1
Change Detection	SVCD [62]	RGB	Mask	16,000	2
	WHU [57]	RGB	Mask	11,456	2
	LEVIR-CD [11]	RGB	Mask	10,192	2
	S2Looking [114]	RGB	Mask	5,000	2
Medical Segmentation	AMOS2022 [58]	CT	Mask	19,310	14
	BraTS2021 [30]	MRI	Mask	76,467	3
	CovidQUEx [121]	X-ray	Mask	5,826	2

be limited in diversity, adding general data mitigates overfitting and prevents the model from losing its foundational ability to understand broad visual patterns.

3.4. Downstream Finetuning

Once ViTP is complete, the pretrained ViT backbone is then extracted to serve as a powerful backbone network. For downstream applications, this pretrained ViT is integrated into a standard ViT-Adapter [18] and combined with task-specific heads for detection, segmentation, or other tasks. The entire model is then finetuned on the target downstream dataset. This transfer learning approach leverages the rich, instruction-aware representations learned during ViTP, enabling faster adaptation and superior performance on specialized tasks.

3.5. Implementation Details

Training a VLM from scratch conventionally requires a multi-stage pipeline: vision–language contrastive pretraining, projector alignment, and large-scale instruction tuning. To bypass the prohibitive cost of these initial phases, we instead bootstrap from a publicly available, high-capacity VLM. Specifically, we initialise our weights from InternVL-2.5 [19], which uses a custmized ViT-Large backbone [20] and a Qwen2 [2] language model. All pretraining are performed on 8× NVIDIA A40 (48 GB) GPUs with a global batch size of 128, and all downstream task finetunings are conducted on $8\times$ NVIDIA RTX3090 (24 GB) GPUs. For ViTP pretraining, we use the AdamW optimizer with a learning rate of 2e-5 and a cosine decay schedule over 8,000 training steps. To accelerate both pre-training and downstream finetuning, Flash-Attention [29] is integrated into every ViT self-attention layer, yielding extra throughput gain without altering convergence behaviour.

Table 2. Composition of the pretraining dataset for remote sensing.

Dataset	Size	Sample Rate	Tasks
Mini-InternVL [38]	1394k	0.03	Caption,VQA,OCR
RSVQA [93]	100k	0.1	VQA
FIT_RS [97]	100k	0.1	VQA
GeoChat [61]	64k	2	VG
VRSBench [71]	38k	5	VG
RSVG [120]	5.5k	10	VG
DIOR-RSVG [159]	27k	8	VG
ISPRS_SAR [56]	1.5k	1	CLS
SAR_Sentinel-1&2 [128]	16k	1	CLS
VHM [108]	223k	1	Caption,VQA,CLS
LevirCCcaptions [86]	50k	0.5	Caption
GAIA [157]	33k	1	Caption
Million-AID [94]	920k	0.05	Caption,CLS

4. Experiments

This section offers a rigorous assessment of ViTP's fine-tuning capabilities on various downstream tasks from both remote-sensing and medical imaging domains. Dataset statistics and task specifications are summarized in Table 1. This section presents our experimental setup, a thorough analysis of the results, and an ablation study. We demonstrate ViTP's superior performance compared to existing state-of-the-art methods and analyze its efficiency and robustness. The best score is indicated in **bold**, while the second-best score is <u>underlined</u>.

4.1. Remote Sensing

For the ViTP pretraining phase, we leverage a diverse set of publicly available image-text paired instruction datasets, adhering to the principles outlined in Section 3.3. The pretraining datasets are listed in Table 2. In the dataset recipe, Million-AID [94], GAIA [157], LevirCCcaptions [86], VHM [108], RSVQA [93], and FIT_RS [97] are large-scale remote sensing visual instruction datasets, providing diverse remote sensing scenes and visual tasks. ISPRS_SAR [56] and SAR Sentinel-1&2 [128] datasets specifically address the unique challenges of Synthetic Aperture Radar (SAR) data. We construct question-answer pairs related to object identification, scene interpretation, and attribute recognition in SAR images using these existing SAR classification datasets. This ensures modality alignment and task relevance for SAR-specific downstream applications. GeoChat [61], DIOR-RSVG [159], RSVG [120], and VRSBench [71] datasets primarily focus on visual grounding tasks within the remote sensing domain, providing rich instructional guidance for spatial reasoning, object attribute recognition, and target localization. The mini-InternVL [38] dataset contributes to gen-

Table 3. Object detection performance (mAP %) on optical remote sensing datasets.

Model	DIOR	DIOR-R	Model	DOTA-v2
GASSL [1]	67.40	65.65	RetinaNet [84]	46.68
SatMAE [27]	70.89	62.30	F-RCNN [112]	47.31
RingMo [119]	75.90	-	FCOS [125]	48.51
CACO [100]	66.91	64.10	ATSS [163]	49.57
SSL4EO [139]	64.82	61.23	SASM [51]	44.53
CMID [103]	75.11	66.37	S2ANet [45]	49.86
RVSA [134]	73.22	70.96	KLD [153]	47.26
SatLas [7]	74.10	67.59	O-RepPoints [67]	48.95
GFM [101]	72.84	67.67	RoT Trans. [32]	52.81
ScaleMAE [111]	73.81	70.20	O-RCNN [146]	53.28
MA3E [79]	-	71.82	GGHF [53]	57.17
Sel-MAE [135]	78.70	71.75	DCFL [148]	57.66
SkySense [41]	<u>78.73</u>	<u>74.27</u>	BillionFM [10]	<u>58.69</u>
ViTP	**79.80**	**75.08**	**ViTP**	**60.23**

Table 4. Object detection performance (mAP %) on SAR datasets.

Model	SARDet-100K	Model	RSAR
DETR [9]	31.8	Def. DETR [174]	46.62
Sparse RCNN [118]	38.1	RetinaNet [84]	57.67
Dab-DETR [90]	45.9	ARS-DETR [158]	61.14
FCOS [125]	46.5	R3Det [152]	63.94
Grid RCNN [96]	48.8	LLMRotate [65]	64.1
GFL [69]	49.8	ReDet [46]	64.71
Defor. DETR [174]	50.0	O-RCNN [146]	64.82
MSFA [76]	53.7	S2ANet [45]	66.47
DenoDet [28]	55.4	RoI-Trans. [32]	66.95
DenoDetv2 [104]	56.4	SatMAE [27]	67.99
SARATR-X [68]	<u>57.3</u>	RemoteCLIP [87]	<u>69.18</u>
ViTP	**59.7**	**ViTP**	**72.31**

eral vision-language understanding capabilities, ensuring that ViTP retains broad applicability and does not overfit to highly specialized remote sensing patterns. The comprehensive pretraining dataset recipe enables ViTP to learn robust, instruction-aware visual representations that are highly relevant to diverse remote sensing tasks.

4.1.1. Object detection

Object detection in remote sensing [77, 143, 146] involves identifying and precisely localizing objects of interest (e.g., vehicles, ships, and bridges) within aerial or satellite imagery. This task is crucial for applications such as urban planning and disaster monitoring. Challenges include vast scale variations of objects, arbitrary orientations, dense object distributions, and complex backgrounds. To evaluate the applicability of our proposed model for remote sensing object detection tasks under various scenarios, we conducted experiments on the following datasets: DIOR [63], DIOR-R [25], DOTA-v2.0 [143], SARDet-100K [76], RSAR [167], and SSDD [164]. We use the standard COCO evaluation metrics (mean Average Precision, mAP) to evaluate the performance of the models. Following previous practice [41, 76, 167], we use the oriented R-CNN [146] as the default detector for oriented object detection tasks and Cascade R-CNN [8] for horizontal object detection tasks. For training and testing, we resized all optical datasets (DIOR, DIOR-R, and DOTA-v2) to a standard size of 1024$\times$1024 pixels. For the DOTA-v2 dataset specifically, we adopted a single-scale approach, cropping images into 1024$\times$1024 patches with a 200-pixel overlap to handle its original large size. The SAR datasets, SARDet-100K and RSAR, were resized to 800$\times$800 pixels.

Table 3 presents the object detection results on RGB remote sensing datasets. ViTP consistently achieves state-of-the-art performance across DIOR, DIOR-R, and DOTA-v2.0. For DOTA-v2.0, which is particularly challenging due to a large quantity of small objects and dense scenes, ViTP achieves a new state-of-the-art of 60.23 mAP. This significantly surpasses previous top performers like BillionFM (58.69), highlighting ViTP's superior ability to handle complex spatial relationships and arbitrary object orientations, likely due to its instruction-following pretraining that fosters fine-grained localization capabilities. To be noticed, SkySense [41] pretraining requires significantly more computational resources, over 17$\times$ the GPU hours of ViTP, ViTP still outperforms SkySense. These results underscore the effectiveness of ViTP's instruction-following pretraining in learning robust and task-relevant visual representations for optical remote sensing imagery.

Table 4 presents the object detection results on Synthetic Aperture Radar (SAR) datasets. SAR imagery poses unique challenges due to speckle noise, different scattering mechanisms, and a lack of visual texture compared to optical images. On SARDet-100K, ViTP achieves 59.7 mAP, significantly outperforming the previous state-of-the-art SARATR-X (57.3) and other detectors. For RSAR, ViTP sets a new state-of-the-art with 72.31 mAP. These remarkable improvements on SAR datasets demonstrate ViTP's strong generalization capabilities and its effectiveness in handling challenging modalities.

4.1.2. Semantic Segmentation

Semantic segmentation in remote sensing [138, 140, 164] involves classifying each pixel in an image into a predefined category, such as land cover types (e.g., forest, water, urban area) or specific objects. This pixel-level understanding is vital for environmental monitoring, urban planning, and resource management. We evaluated ViTP's performance on segmentation tasks using the following widely-used remote sensing datasets: iSAID [140], LoveDA [136], UAVid [98], and SSDD [164]. Following previous practice [41, 119], we use the UperNet [144] as the default segmentor for semantic segmentation tasks (iSAID, LoveDA, and UAVid) and Mask-RCNN [48] for instance segmentation tasks (SSDD). We report the mean Intersection over Union (mIoU) as the

Table 5. Semantic segmentation performance (mIoU %) on optical datasets.

model	iSAID	LoveDA	model	UAVid
SeCo [154]	57.20	43.63	CANet [151]	63.50
DenseCLIP [110]	59.23	49.58	MP-Former [161]	63.67
SatMAE [27]	62.97	-	ABCNet [66]	63.80
CACo [100]	64.32	48.89	DecoupleNet [95]	65.80
RVSA [134]	64.49	52.44	CoaT [150]	65.80
RSSFormer [149]	65.55	52.43	UNetFormer [138]	67.80
ScaleMAE [111]	65.77	-	MaskFormer [23]	68.54
GASSL [1]	65.95	48.76	LSKNet [75]	70.00
CMID [103]	66.21	-	Segmenter [116]	70.20
TOV [123]	66.24	49.70	RSSFormer [149]	70.69
RingMo [119]	67.20	-	DeepLabv3+ [17]	71.33
SatLas [7]	68.71	-	SegFormer [145]	71.44
Sel-MAE [135]	-	53.92	DenseCLIP [110]	71.54
LSKNet [74]	-	54.00	PSPNet [169]	71.71
SkySense [41]	70.91	-	OCRNet [156]	71.84
ViTP	**71.14**	**54.28**	**ViTP**	**73.39**

Table 6. Detailed object detection and instance segmentation performance on the SSDD Dataset (SAR modality).

Model	AP_{box}	AP_{box}^{75}	AP_{box}^{75}	AP_{mask}	AP_{mask}^{50}	AP_{mask}^{75}
BoxInst [127]	44.76	83.75	44.11	34.10	71.16	27.27
Mask2Former [22]	53.40	78.45	67.02	56.52	85.10	69.48
InstaBoost [35]	54.77	87.85	58.54	58.95	89.05	71.57
CondInst [126]	57.89	92.53	67.40	50.31	90.46	54.80
SAM-Seg [60]	62.41	94.32	75.38	59.46	92.79	72.17
CATNet [91]	64.66	96.46	79.81	64.11	96.35	77.87
HQ-ISNet [117]	65.58	95.48	80.76	64.75	95.26	81.70
RSP-Query [16]	66.50	95.80	81.81	64.57	95.97	81.67
SCNet [130]	67.25	95.75	83.38	62.66	94.75	76.53
ViTP	**70.80**	**97.80**	**86.60**	**65.90**	**96.80**	**81.80**

evaluation metric for semantic segmentation and Average Precision (AP) for instance segmentation tasks.

Table 5 and Table 6 demonstrate ViTP's strong performance in remote sensing semantic segmentation. It sets new state-of-the-art performance on iSAID, LoveDA, UAVid, and SSDD datasets. Especially for UAVid and SSDD datasets, ViTP surpasses previous state-of-the-art by a large margin. The consistent improvements across these semantic segmentation benchmarks confirm that ViTP's instruction-following pretraining on complex fine-grained regional understanding questions enables the ViT backbone to capture rich semantic information, which is crucial for accurate pixel-level classification.

4.1.3. Change Detection

Change detection in remote sensing focuses on identifying and characterizing differences in the state of an object or phenomenon by observing it at different times. This bitemporal analysis is fundamental for applications such as urban sprawl monitoring, deforestation tracking, and urbanization analysis. The primary challenges in this task include handling variations in illumination and atmospheric conditions between image acquisitions, precise image co-registration, and distinguishing meaningful semantic changes from irrelevant ones. To assess ViTP's capabilities in this domain, we conduct experiments on three widely-recognized public datasets: SVCD [62], LEVIR-CD [11] and WHU-CD [57]. Following the common practice, we employ a simple Siamese UperNet-based framework [144] as the default detector and report the F1-Score as our primary evaluation metric.

Table 7. Change detection performance (F1-Score %) on SVCD, WHU, LEVER-CD and S2Looking datasets.

Model	**SVCD**	**WHU**	**LEVIR**	**S2Looking**
Scale-MAE [111]	-	-	86.60	50.20
SeCo [154]	-	-	88.40	66.00
CACo [100]	-	-	89.20	65.90
GASSL [1]	-	-	89.60	66.30
SatMAE [27]	-	-	90.00	65.00
SatMAE++ [106]	-	-	90.70	56.40
CGNet [44]	-	-	92.01	64.33
Changer [37]	-	-	92.06	67.08
DiFormer [81]	-	-	92.15	66.31
Changen2 [173]	-	-	92.20	69.10
SkySense [41]	-	-	92.58	-
CLNet [172]	92.10	-	90.00	-
SRCDNet [89]	92.94	87.40	-	-
ESCNet [160]	93.54	-	-	-
DSAMNet [115]	93.69	-	-	-
GCD-DDPM [141]	94.93	92.54	90.96	-
CDContrast [137]	95.11	-	-	-
DDPM-CD [5]	95.62	92.65	90.91	-
DMNet [72]	95.93	-	-	-
SNUnet [36]	96.20	83.49	88.59	63.19
BIT [12]	-	83.98	89.31	63.76
BiFA [36]	-	94.37	90.69	-
SGSLN [170]	96.24	94.67	91.93	-
RSP [132]	96.81	-	90.93	-
SAAN [40]	97.03	-	91.41	-
SiamixFormer [102]	97.13	-	91.58	-
TransUNetCD [64]	97.17	93.59	91.11	-
RDPNet [13]	97.20	-	91.20	-
WNet [122]	97.56	91.25	90.67	-
ChangeMamba [14]	-	92.55	90.16	-
RS-Mamba [171]	-	92.79	89.77	-
ChangeFormer [4]	-	93.04	91.11	63.39
CDMamba [162]	-	93.76	90.75	67.08
LSKNet [75]	-	92.06	92.27	67.52
RVSA [134]	97.78	94.07	92.52	-
ChangeCLIP [33]	97.89	94.82	92.01	-
P2V-CD [82]	98.42	92.38	91.94	-
ViTP	**98.63**	**94.98**	**92.67**	**69.89**

The comprehensive results, presented in Table 7, show that ViTP achieves state-of-the-art performance across all of SVCD, WHU, LEVIR-CD and S2Looking datasets. For the SVCD dataset, which contains more diverse change types, ViTP shows a clear performance advantage. On the

Table 8. Composition of the pretraining dataset for medical imagery.

Dataset	Size	Sample Rate	Tasks
Mini-InternVL [38]	1394k	0.02	Caption,VQA,OCR
GMAI-MMBench [155]	5k	20	VQA
Open-i [31]	7k	5	Caption
Huatuo-OA [15]	647k	0.1	Caption
Huatuo-VQA [15]	647k	0.5	VQA
PMC-OA [85]	1647k	0.05	Caption
PMC-VQA [166]	227k	0.5	VQA
OmniMedical [52]	89k	1	VQA
Quilt-1M [54]	723k	0.1	Caption
Quilt-Instruct [113]	147	1	VQA

Table 9. Semantic segmentation performance (mDice %) on Medical Imagery Datasets.

model	AMOS2022	BraTS2021	CovidQUEx
5-points Prompt			
MedSAM [99]	81.36	69.44	78.18
SAMMed2D [24]	87.81	70.70	83.39
IMIS-SAM [26]	87.42	64.77	86.28
Box Prompt			
MedSAM [99]	86.74	71.23	78.87
SAMMed2D [24]	88.67	<u>71.26</u>	77.81
IMIS-SAM [26]	88.71	70.59	82.91
nnUNet [55]	87.28	71.03	<u>90.41</u>
MedMAE [42]	<u>90.11</u>	69.59	90.18
ViTP	**90.60**	**72.11**	**91.55**

LEVIR-CD dataset, which focuses on building changes, ViTP achieves a new state-of-the-art F1 of 92.67. This surpasses previous methods, demonstrating its proficiency in identifying fine-grained changes in complex suburban environments. Similarly, on the WHU-CD dataset, ViTP sets a new benchmark with an F1 of 94.98, effectively handling variations in building scale and appearance. These significant improvements underscore the rich semantic understanding power of ViTP for temporal analysis.

4.2. Medical Imaging

Similar to the approach for remote sensing, we collected several open-source medical image-text paired instruction datasets. The pretraining datasets are listed in Table 8. Among them, Huatuo [15], PMC [85], and OmniMedical [52] are large-scale, general medical visual instruction datasets covering various medical imaging modalities, including CT, MRI, ultrasound, and others. These datasets reframe various downstream tasks into a question-answering format, augmented with millions of descriptive captions, enabling the model to build comprehensive knowledge across diverse medical imaging modalities. Moreover, since medical image segmentation is a key downstream task, we also include the GAMI-MMBench [155] dataset. It provides VQA based on ROIs and segmentation masks, which aids in aligning the model with dense prediction tasks. The Open-i [31] and Quilt [54] datasets are included to compensate for the relative scarcity of X-ray and pathology images in other general medical collections. Finally, as with the remote sensing domain, mini-InternVL [38] is incorporated to enhance general vision-language understanding.

4.2.1. Semantic Segmentation

Semantic segmentation in medical image analysis involves classifying each pixel in an image into a predefined category, such as an organ (e.g., lung, kidney, heart) or a specific lesion. This pixel-level understanding is vital for clinical diagnosis, disease analysis, and patient prognosis. We evaluated ViTP's performance on segmentation tasks using three challenging medical datasets: AMOS2022 [58], BraTS2021 [30], and ConvidQUEx [121]. These datasets encompass three primary clinical modalities: CT (AMOS2022), MRI (BraTS2021), and chest X-ray (ConvidQUEx), and covering both organ and lesion segmentation scenarios. We use UperNet [144] as the default segmentation head. We report the mean Dice Score (mDice) as the evaluation metric for all segmentation tasks.

Table 9 demonstrates the strong performance of ViTP in medical semantic segmentation. ViTP achieves new state-of-the-art results on the AMOS, BraTS, and ConvidQUEx datasets. Specifically, similar to ViTP, SAM-based approaches (MedSAM [99], SAMMed2D [24] and IMIS-SAM [26]) are also finetuned on the downstream datasets. However, ViTP achieves substantial performance gains, which can be attributed to its more effective pretraining strategy. When compared to the classic medical segmentation model (nnU-Net [55]), ViTP not only exhibits superior generalization but also surpasses its performance, despite nnU-Net being a highly specialized model. Finally, a comparison with MedMAE [42], a recent 2D medical pretraining method, highlights the effectiveness of our proposed pretraining paradigm.

5. Analysis and Ablation Study

To comprehensively understand the impact of various components and hyperparameters on ViTP's performance, we conduct a series of rigorous ablation studies and analytical experiments. This section details the ablation studies, including: (1) validating the principles of our data recipe, (2) assessing the impact of pre-training steps, (3) evaluating the effect of the VRL image token dropping ratio, and (4) analyzing the influence of the ViTP language model's size.

For these studies, we utilize the ViTP model pre-trained on remote sensing data and evaluate its performance on the RSAR dataset, a challenging benchmark for oriented object detection in SAR imagery. We selected this specific downstream task for several reasons. Firstly, the SAR modal-

Table 10. Validation of the data recipe. Performance on the RSAR Benchmark confirms the necessity of our comprehensive data curation strategy.

Pretrain Paradigm	RSAR mAP
w/o Diversity	52.6
w/o SAR	52.5
w/o Grounding	53.0
w/o General data	52.3
full data	54.6

ity presents unique challenges not found in optical imagery, making it an excellent testbed for a model's ability to generalize to specialized visual domains. Secondly, oriented object detection demands precise localization, which stringently tests the model's capacity to capture fine-grained spatial information and recognize weak patterns amidst dense objects. Therefore, performance on the RSAR dataset serves as a fair and insightful proxy for the model's overall representational power. For experimental efficiency, unless otherwise specified, models are finetuned on the RSAR validation set and evaluated on its test set.

Furthermore, we investigate the unique advantages of the ViTP paradigm over other pre-training methods during the finetuning stage, specifically focusing on its data efficiency and robustness to various data corruptions.

5.1. Data Recipe

Our ViTP pretraining dataset was carefully curated according to the principles stated in Section 3.3. Table 10 ablates the contribution of different data types, where "w/o Diversity" refers to the exclusion of diverse remote sensing visual instruction datasets (Million-AID, GAIA, Levir-CC, VHM, VRSBench, and RSVG) during pre-training. Removing any single component: diverse datasets, SAR-specific datasets, grounding datasets, or general-domain datasets, leads to a significant drop in downstream performance on the RSAR dataset. Excluding SAR data prevents the model from learning modality-specific features, while removing grounding data impairs its localization ability. Omitting general data leads to overfitting on specialized patterns, which leads to the most severe performance drop. This confirms that a diverse mixture of diverse, general, domain-specific, and task-specific instruction data is crucial for optimal performance.

5.2. Impact of Pretraining Steps

The duration of pre-training is a significant factor influencing downstream task performance. As depicted in Figure 5, the model's mAP on the RSAR dataset generally improves with an increasing number of pre-training steps. This trend indicates that more extensive instruction tuning enables the model to learn progressively better feature representations. However, the performance curve begins to plateau around

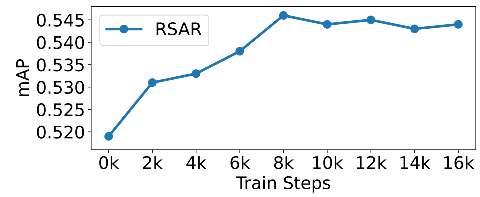
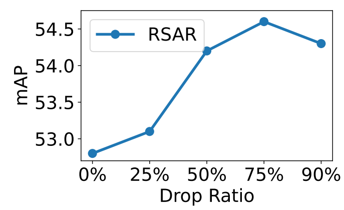


Figure 5. Effect of Pretraining Duration. RSAR mAP improves with more pretraining steps before saturating at approximately 8k steps.



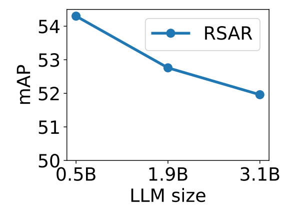


Figure 6. Impact of the VRL Drop Ratio. ViTP performance on RSAR reaches a peak with a 75% token drop ratio.

Figure 7. Impact of LLM size. A larger LLM hinders the performance of ViTP on the RSAR Benchmark.

8k steps, with no further improvements observed thereafter. Consequently, we set the number of ViTP pre-training steps to 8k by default for our main model, balancing performance with computational cost. This pre-training phase takes approximately 23 hours to complete 8k steps training on 8$\times$ A40 GPUs, demonstrating the efficiency of our proposed method.

5.3. Impact of VRL Token drop Ratio

The Visual Robustness Learning (VRL), introduced in Section 3.2, involves randomly dropping a portion of image tokens during pre-training. This study investigates the optimal dropping ratio for this process. An appropriate ratio compels the ViT to learn more robust and semantically rich representations by inferring missing visual context. Conversely, an excessively high ratio risks obscuring essential image information, thereby posing an ill-posed training target and hindering effective learning.

As shown in Figure 6, the model's performance on the RSAR dataset varies significantly with different dropping ratios. The performance peaks at a ratio of 75%, which improves the mAP from 52.8 (without VRL) to 54.6. This suggests that a 75% masking ratio effectively regularizes the model without excessively impeding information flow. Ratios from 0% to 50% appear to provide insufficient regularization. Surprisingly, even with a very high masking rate of 90%, the model still benefits from this strong regularization, achieving a notable 54.3 mAP.

Table 11. Pretraining efficiency and performance. Time is estimated for pretraining on 8× A40 GPUs. ViTP achieves SOTA or competitive performance with a fraction of the computational cost.

	Methods	Hours	DIOR-R	iSAID
MIM	RVSA [134]	250	70.96	64.49
	Scale-MAE [111]	<u>60</u>	70.20	65.77
CL	RemoteCLIP [87]	100	70.20	62.53
	SkySense [41]	400	<u>74.27</u>	**70.91**
	ViTP	**23**	**75.08**	<u>70.54</u>

Table 12. Data efficiency on the RSAR Benchmark. ViTP demonstrates significantly stronger performance in low-data regimes compared to MIM and contrastive methods.

Model	100%	20%	10%	5%	2%
RemoteCLIP [87]	69.18	63.14 ↓6.04	57.10 ↓12.08	47.25 ↓21.93	34.78 ↓34.40
SatMAE [27]	67.99	61.36 ↓6.63	55.24 ↓12.75	50.76 ↓17.23	37.90 ↓30.09
ViTP	**72.31**	**67.07** **↓5.24**	**61.68** **↓10.63**	**56.42** **↓15.89**	**46.98** **↓25.33**

5.4. Impact of Language Model Size

This study investigates the influence of the Large Language Model's (LLM) capacity on ViTP's overall performance. As shown in Figure 7, as the LLM size increases, the downstream performance on the RSAR dataset steadily degrades. This phenomenon is analogous to the findings in Masked Autoencoders [50] pretrain paradigm, which highlight the importance of a lightweight decoder. It indicates that for the challenging SAR modality, a more powerful LLM might compensate for suboptimal visual features extracted by the Vision Transformer (ViT). This could reduce the optimization pressure on the ViT, preventing it from learning the most effective modality-specific representations.

5.5. Training Efficiency

A key advantage of the ViTP paradigm is its exceptional pretraining efficiency, which lowers the barrier for developing specialized foundation models. As detailed in Table 11, ViTP pretraining completes in one day on 8 A40 GPUs. This is dramatically faster than competing paradigms. For instance, ViTP is **2.6x** faster than the efficient MIM-based method (Scale-MAE) and over **17x** faster than the high-performing but computationally intensive contrastive learning method, SkySense.

5.6. Data Efficiency

We investigate ViTP's performance in low-data regimes by finetuning it on subsets of the RSAR training dataset, ranging from 2% to 20%. To fully assess the modeling potential of each method, we extend the finetuning schedule to 36 epochs for this specific study. As shown in Table 12, ViTP consistently outperforms SatMAE[27] (an MIM method) and RemoteCLIP [87] (a contrastive method) across all data fractions. ViTP's advantage becomes more pronounced as the amount of training data decreases. For instance, with only 2% of the training data, ViTP achieves 46.98 mAP, substantially outperforming SatMAE (37.90 mAP) and RemoteCLIP (34.78 mAP). Notably, ViTP finetuned on just 20% of the data (67.07 mAP) achieves performance comparable to or exceeding many state-of-the-art methods trained on the full dataset (as shown in Table 4). This superior data efficiency suggests that the rich semantic priors acquired during instruction tuning enable the model to generalize more effectively from a limited number of examples.

5.7. Model Robustness

While most research datasets contain curated clean imagery, real-world remote sensing images are often degraded by atmospheric conditions (e.g., clouds, haze), sensor noise, or processing artifacts. To evaluate ViTP's resilience to such degradations, we assess its robustness against 12 common image corruptions from the REOBench [73] benchmark on the DIOR-R dataset. We retrained our ViTP model on the DIOR-R dataset using an image size of 800×800 pixels and without any test-time augmentation, matching the settings of the REOBench baseline models. As shown in Table 13, the standard ViTP model is significantly more robust than all Masked Image Modeling (MIM) and Contrastive Learning (CL) models. It achieves the highest average mAP of 69.00 across all corruptions and the lowest performance drop (Δ_{TP}) of 4.37 from its clean-image performance. We also ablate the effect of our proposed Visual Robustness Learning (VRL). The model trained without VRL already demonstrates superior robustness over the baselines. The inclusion of VRL further boosts the average performance from 67.13 to 69.00 mAP and reduces the performance degradation (Δ_{TP}) from 4.78 to 4.37. This confirms that our instruction tuning paradigm, enhanced with VRL, is a highly effective technique for learning robust and comprehensive feature representations.

6. Limitation and Future Work

Despite the strong performance and efficiency of ViTP, we acknowledge several limitations that open promising avenues for future research. A primary limitation is that ViTP's success is intrinsically linked to the quality and diversity of the instruction-following dataset. As our experiments suggest, a carefully curated data recipe is important for achieving optimal performance. The current process of collecting, filtering, and balancing datasets for specialized domains like remote sensing or medical imaging requires considerable manual effort and deep domain expertise. Future work could focus on developing more automated or semi-automated strategies for generating high-

Table 13. Model robustness to common image corruptions. ViTP demonstrates superior resilience compared to MIM and CL baselines on the REOBench benchmark (DIOR-R, mAP). It maintains the highest average performance across all corruptions and suffers the smallest performance drop (Δ_{TP}) from the clean baseline. The inclusion of Visual Robustness Learning (VRL) further enhances this robustness.

	Model	Clean	Bright Contrast	Cloud	Comp. Artifact	Data Gaps	Gauss Blur	Gauss Noise	Haze	Motion Blur	Rotate	Salt Pepper	Scale	Trans	**Avg**	$\Delta_{\text{TP}} \downarrow$
MIM	SatMAE [27]	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	5.49
	ScaleMAE [111]	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	6.98
	RVSA [134]	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	8.45
	SatMAE++ [106]	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	5.73
CL	RemoteCLIP [87]	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	6.21
	GeoRSCLIP [168]	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	6.42
	ViTP w/o VRL	<u>71.91</u>	69.11	67.25	66.20	66.87	67.67	67.47	67.12	65.94	65.54	66.87	65.23	70.34	<u>67.13</u>	<u>4.78</u>
	ViTP	**73.37**	**70.56**	**69.54**	**67.74**	**70.58**	**69.05**	**68.97**	**70.85**	**67.86**	**67.10**	**67.23**	**66.64**	**71.87**	**69.00**	**4.37**

quality, domain-specific instruction data. Leveraging large language models to synthesize diverse question-answer or description-grounding pairs from unlabeled images could be a promising direction to enhance the scalability and accessibility of our pretraining paradigm. Moreover, our current work focuses on pretraining 2D image-based Vision Transformers. Extending the ViTP framework to other data modalities, such as video and 3D point clouds, represents another exciting frontier. Adapting the instruction-following objective to these domains could unlock new capabilities in temporal reasoning and 3D spatial understanding.

7. Conclusion

In this paper, we addressed the underexplored, top-down pathway from high-level reasoning to low-level perception in vision foundation models. We introduced Visual insTruction Pretraining (ViTP), a novel paradigm that leverages the high-level semantic reasoning of a Vision-Language Model to directly guide the feature learning of a ViT backbone. Complemented by our Visual Robustness Learning (VRL), ViTP compels the vision model to learn rich, robust, and task-relevant representations from downstream-specific instructional data. Our extensive experiments on 16 challenging remote sensing and medical imaging benchmarks validated the effectiveness of the proposed ViTP method. ViTP not only established new state-of-the-art performance across a diverse range of downstream tasks but also proved to be remarkably computationally efficient. We believe this work opens a promising new avenue for a deeper integration of high-level reasoning into the core of visual feature learning.

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 7, 8

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 6

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 4

[4] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IEEE International Geoscience and Remote Sensing Symposium*, 2022. 8

[5] Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M Patel. Ddpm-cd: Denoising diffusion probabilistic models as feature extractors for remote sensing change detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5250–5262. IEEE, 2025. 8

[6] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[7] 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. 7, 8

[8] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018. 7

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 7

[10] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023. 7

[11] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote sensing*, 12(10):1662, 2020. 6, 8

[12] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *TGRS*, 2021. 8

[13] H Chen, F Pu, R Yang, R Tang, and X Xu. Rdp-net: Region detail preserving network for change detection. arxiv 2022. *arXiv preprint arXiv:2202.09745*, 2022. 8

[14] Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–20, 2024. 8

[15] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370, 2024. 9

[16] Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. 8

[17] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 8

[18] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 6

[19] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3, 6

[20] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 4, 6

[21] Zhenyuan Chen, Lingfeng Yang, Shuo Chen, Zhaowei Chen, Jiajun Liang, and Xiang Li. Revisiting prompt pretraining of vision-language models. *arXiv preprint arXiv:2409.06166*, 2024. 1

[22] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 8

[23] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 8

[24] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*, 2023. 3, 9

[25] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *TGRS*, 2022. 6, 7

[26] Junlong Cheng, Bin Fu, Jin Ye, Guoan Wang, Tianbin Li, Haoyu Wang, Ruoyu Li, He Yao, Junren Cheng, JingWen Li, et al. Interactive medical image segmentation: A benchmark dataset and baseline. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20841–20851, 2025. 3, 9

[27] 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. 3, 7, 8, 11, 12

[28] Yimian Dai, Minrui Zou, Yuxuan Li, Xiang Li, Kang Ni, and Jian Yang. Denodet: Attention as deformable multi-subspace feature denoising for target detection in sar images. *IEEE Transactions on Aerospace and Electronic Systems*, 2024. 7

[29] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6

[30] Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *arXiv preprint arXiv:2405.18368*, 2024. 6, 9

[31] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 9

[32] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *CVPR*, 2019. 7

[33] Sijun Dong, Libo Wang, Bo Du, and Xiaoliang Meng. Changeclip: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:53–69, 2024. 8

[34] 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. *arXiv preprint arXiv:2010.11929*, 2020. 3

[35] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 682–691, 2019. 8

[36] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. Snunet-cd: A densely connected siamese network for change detec-

tion of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 2021. 8

[37] Sheng Fang, Kaiyu Li, and Zhe Li. Changer: Feature interaction is what you need for change detection. *TGRS*, 2023. 8

[38] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):32, 2024. 6, 9

[39] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 2020. 3

[40] Haonan Guo, Xin Su, Chen Wu, Bo Du, and Liangpei Zhang. Saan: Similarity-aware attention flow network for change detection with vhr remote sensing images. *IEEE Transactions on Image Processing*, 33:2599–2613, 2024. 8

[41] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. *arXiv*, 2023. 1, 3, 7, 8, 11

[42] Anubhav Gupta, Islam Osman, Mohamed S Shehata, and John W Braun. Medmae: A self-supervised backbone for medical imaging tasks. *arXiv preprint arXiv:2407.14784*, 2024. 3, 9

[43] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020. 3, 4

[44] Chengxi Han, Chen Wu, Haonan Guo, Meiqi Hu, Jiepan Li, and Hongruixuan Chen. Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023. 8

[45] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *TGRS*, 2021. 7

[46] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *CVPR*, 2021. 7

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[48] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7

[49] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 3

[50] 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. 1, 3, 11

[51] Liping Hou, Ke Lu, Jian Xue, and Yuqiu Li. Shape-adaptive selection and measurement for oriented object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 923–932, 2022. 7

[52] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024. 9

[53] Zhanchao Huang, Wei Li, Xiang-Gen Xia, and Ran Tao. A general gaussian heatmap label assignment for arbitrary-oriented object detection. *IEEE Transactions on Image Processing*, 31:1895–1910, 2022. 7

[54] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017, 2023. 9

[55] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 9

[56] ISPRS. Classification of multiscale marine phenomenon in sar images. https://www.isprs2024tc1.net/contest/, 2025. Accessed: 2025-08-22. 6

[57] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018. 6, 8

[58] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. 6, 9

[59] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025. 1

[60] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3, 8

[61] 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. 6

[62] MA Lebedev, Yu V Vizilter, OV Vygolov, Vladimir A Knyaz, and A Yu Rubis. Change detection in remote sensing images using conditional adversarial networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:565–571, 2018. 6, 8

[63] 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*, 2020. 6, 7

[64] Qingyang Li, Ruofei Zhong, Xin Du, and Yu Du. Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 8

[65] Qingyun Li, Yushi Chen, Xinya Shu, Dong Chen, Xin He, Yi Yu, and Xue Yang. A simple aerial detection baseline of multimodal language models. *arXiv preprint arXiv:2501.09720*, 2025. 7

[66] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Libo Wang, and Peter M. Atkinson. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021. 8

[67] Wentong Li, Yijie Chen, Kaixuan Hu, and Jianke Zhu. Oriented reppoints for aerial object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1829–1838, 2022. 7

[68] Weijie Li, Wei Yang, Yuenan Hou, Li Liu, Yongxiang Liu, and Xiang Li. Saratr-x: Towards building a foundation model for sar target recognition. *IEEE Transactions on Image Processing*, 2025. 7

[69] Xiang Li, Chengqi Lv, Wenhai Wang, Gang Li, Lingfeng Yang, and Jian Yang. Generalized focal loss: Towards efficient representation learning for dense object detection. *TPAMI*, 2022. 7

[70] Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. *arXiv preprint arXiv:2205.10063*, 2022. 3

[71] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. *Advances in Neural Information Processing Systems*, 37:3229–3242, 2024. 6

[72] Xi Li, Li Yan, Yi Zhang, and Huaien Zeng. Esr-dmnet: Enhanced super-resolution-based dual-path metric change detection network for remote sensing images with different resolutions. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. 8

[73] Xiang Li, Yong Tao, Siyuan Zhang, Siwei Liu, Zhitong Xiong, Chunbo Luo, Lu Liu, Mykola Pechenizkiy, Xiao Xiang Zhu, and Tianjin Huang. Reobench: Benchmarking robustness of earth observation foundation models. *arXiv preprint arXiv:2505.16793*, 2025. 11

[74] Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, and Xiang Li. Large selective kernel network for remote sensing object detection. In *ICCV*, 2023. 8

[75] Yuxuan Li, Xiang Li, Yimain Dai, Qibin Hou, Li Liu, Yongxiang Liu, Ming-Ming Cheng, and Jian Yang. Lsknet: A foundation lightweight backbone for remote sensing. *IJCV*, 2024. 8

[76] Yuxuan Li, Xiang Li, Weijie Li, Qibin Hou, Li Liu, Ming-Ming Cheng, and Jian Yang. Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. *NeurIPS*, 2024. 3, 6, 7

[77] Yuxuan Li, Xiang Li, Yunheng Li, Yicheng Zhang, Yimian Dai, Qibin Hou, Ming-Ming Cheng, and Jian Yang. Sm3det: A unified model for multi-modal remote sensing object detection. *arXiv preprint arXiv:2412.20665*, 2024. 7

[78] Yunheng Li, Yuxuan Li, Quansheng Zeng, Wenhai Wang, Qibin Hou, and Ming-Ming Cheng. Unbiased region-language alignment for open-vocabulary dense prediction. *ICCV*, 2025. 2

[79] Zhihao Li, Biao Hou, Siteng Ma, Zitong Wu, Xianpeng Guo, Bo Ren, and Licheng Jiao. Masked angle-aware autoencoder for remote sensing images. In *European Conference on Computer Vision*. Springer, 2024. 7

[80] Jiaqi Liao, Zhengyuan Yang, Linjie Li, Dianqi Li, Kevin Lin, Yu Cheng, and Lijuan Wang. Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning. *arXiv preprint arXiv:2503.19312*, 2025. 1

[81] Hui Lin, Renlong Hang, Shanmin Wang, and Qingshan Liu. Diformer: A difference transformer network for remote sensing change detection. *IEEE Geoscience and Remote Sensing Letters*, 2024. 8

[82] Manhui Lin, Guangyi Yang, and Hongyan Zhang. Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images. *IEEE Transactions on Image Processing*, 32:57–71, 2022. 8

[83] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[84] Tsung-Yi Lin, Priya Goyal, Ross Girshick, and Kaiming He. Focal loss for dense object detection. In *ICCV*, 2017. 7

[85] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *MICCAI*, 2023. 9

[86] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. 6

[87] 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*, 62:1–16, 2024. 1, 3, 7, 11, 12

[88] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 2, 4

[89] Mengxi Liu, Qian Shi, Andrea Marinoni, Da He, Xiaoping Liu, and Liangpei Zhang. Super-resolution-based change detection network with stacked attention module for images with different resolutions. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021. 8

[90] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *ICLR*, 2022. 7

[91] Ye Liu, Huifang Li, Chao Hu, Shuang Luo, Huanfeng Shen, and C Chen. Catnet: context aggregation network for instance segmentation in remote sensing images. *arXiv preprint arXiv:2111.11057*, 2021. 8

[92] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *CVPR*, 2021. 2

[93] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020. 6

[94] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14:4205–4230, 2021. 6

[95] Wei Lu, Si-Bao Chen, Qing-Ling Shu, Jin Tang, and Bin Luo. Decouplenet: A lightweight backbone network with efficient feature decoupling for remote sensing visual tasks. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 8

[96] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *CVPR*, pages 7363–7372, 2019. 7

[97] 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. 6

[98] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. UAVid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020. 6, 7

[99] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024. 3, 9

[100] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5270, 2023. 3, 7, 8

[101] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3, 7

[102] Amir Mohammadian and Foad Ghaderi. Siamixformer: A fully-transformer siamese network with temporal fusion for accurate building detection and change detection in bi-temporal remote sensing images. *International Journal of Remote Sensing*, 44(12):3660–3678, 2023. 8

[103] Dilxat Muhtar, Xueliang Zhang, Pengfeng Xiao, Zhenshi Li, and Feng Gu. Cmid: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023. 3, 7, 8

[104] Kang Ni, Minrui Zou, Yuxuan Li, Xiang Li, Kehua Guo, Ming-Ming Cheng, and Yimian Dai. Denodet v2: Phase-amplitude cross denoising for sar object detection. *arXiv preprint arXiv:2508.09392*, 2025. 7

[105] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. 1

[106] 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. 8, 12

[107] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 3

[108] 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*, pages 6381–6388, 2025. 6

[109] 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. 1, 2, 3

[110] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 8

[111] 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. 1, 3, 7, 8, 11, 12

[112] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 7

[113] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava:

Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192, 2024. 9

[114] Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 2021. 6

[115] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE transactions on geoscience and remote sensing*, 60:1–16, 2021. 8

[116] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 8

[117] Hao Su, Shunjun Wei, Shan Liu, Jiadian Liang, Chen Wang, Jun Shi, and Xiaoling Zhang. Hq-isnet: High-quality instance segmentation for remote sensing imagery. *Remote Sensing*, 12(6):989, 2020. 8

[118] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, 2021. 7

[119] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. Ringmo: A remote sensing foundation model with masked image modeling. *TGRS*, 2023. 3, 7, 8

[120] Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International conference on Multimedia*, pages 404–412, 2022. 6

[121] Anas M. Tahir, Muhammad E. H. Chowdhury, Yazan Qiblawey, Amith Khandakar, Tawsifur Rahman, Serkan Kiranyaz, Uzair Khurshid, Nabil Ibtehaz, Sakib Mahmud, and Maymouna Ezeddin. Covid-qu-ex dataset, 2022. 6, 9

[122] Xu Tang, Tianxiang Zhang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Wnet: W-shaped hierarchical network for remote-sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. 8

[123] Chao Tao, Ji Qi, Guo Zhang, Qing Zhu, Weipeng Lu, and Haifeng Li. Tov: The original vision model for optical remote sensing image understanding via self-supervised learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:4916–4930, 2023. 3, 8

[124] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 4

[125] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 7

[126] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European conference on computer vision*, pages 282–298. Springer, 2020. 8

[127] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5443–5452, 2021. 8

[128] Paritosh Tiwari. Sentinel-1&2 image pairs (sar & optical). `https://www.kaggle.com/datasets/requiemonk/sentinel12-image-pairs-segregated-by-terrain/code?datasetId=1201791`, 2025. Accessed: 2025-08-22. 6

[129] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 4

[130] Thang Vu, Haeyong Kang, and Chang D Yoo. Scnet: Training inference sample consistency for instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2701–2709, 2021. 8

[131] Tassilo Wald, Constantin Ulrich, Stanislav Lukyanenko, Andrei Goncharov, Alberto Paderno, Maximilian Miller, Leander Maerkisch, Paul Jaeger, and Klaus Maier-Hein. Revisiting mae pre-training for 3d medical image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5186–5196, 2025. 3

[132] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–20, 2022. 8

[133] Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36:8815–8827, 2023. 3

[134] 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*, 2023. 7, 8, 11, 12

[135] F Wang, H Wang, D Wang, Z Guo, Z Zhong, L Lan, W Yang, and J Zhang. Harnessing massive satellite imagery with efficient masked image modeling. *arXiv preprint arXiv:2406.11933*, 2025. 7, 8

[136] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv*, 2021. 6, 7

[137] Jue Wang, Yanfei Zhong, and Liangpei Zhang. Change detection based on supervised contrastive learning for high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. 8

[138] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M. Atkinson. UNet-Former: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS*

Journal of Photogrammetry and Remote Sensing, 2022. 7, 8

[139] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 2023. 7

[140] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 28–37, 2019. 6, 7

[141] Yihan Wen, Xianping Ma, Xiaokang Zhang, and Man-On Pun. Gcd-ddpm: A generative change detection model based on difference-feature-guided ddpm. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 8

[142] Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *arXiv preprint arXiv:2507.01467*, 2025. 1

[143] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018. 6, 7

[144] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 7, 8, 9

[145] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 8

[146] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *ICCV*, 2021. 7

[147] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022. 3

[148] Chang Xu, Jian Ding, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Dynamic coarse-to-fine learning for oriented tiny object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7318–7328, 2023. 7

[149] Rongtao Xu, Changwei Wang, Jiguang Zhang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation. *TIP*, 2023. 8

[150] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *ICCV*, 2021. 8

[151] Michael Ying Yang, Saumya Kumaar, Ye Lyu, and Francesco Nex. Real-time semantic segmentation with context aggregation network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021. 8

[152] Xue Yang, Junchi Yan, Ziming Feng, and Tao He. R3det: Refined single-stage detector with feature refinement for rotating object. In *AAAI*, 2021. 7

[153] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. Learning high-precision bounding box for rotated object detection via kullback-leibler divergence. *Advances in Neural Information Processing Systems*, 34:18381–18394, 2021. 7

[154] Ting Yao, Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, and Tao Mei. Seco: Exploring sequence supervision for unsupervised representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2021. 8

[155] Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024. 9

[156] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, pages 173–190. Springer, 2020. 8

[157] Angelos Zavras, Dimitrios Michail, Xiao Xiang Zhu, Begüm Demir, and Ioannis Papoutsis. Gaia: A global, multi-modal, multi-scale vision-language dataset for remote sensing image analysis. *arXiv preprint arXiv:2502.09598*, 2025. 6

[158] Ying Zeng, Yushi Chen, Xue Yang, Qingyun Li, and Junchi Yan. Ars-detr: Aspect ratio-sensitive detection transformer for aerial oriented object detection. *TGRS*, 2024. 7

[159] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 6

[160] Hongyan Zhang, Manhui Lin, Guangyi Yang, and Liangpei Zhang. Escnet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):28–42, 2021. 8

[161] Hao Zhang, Feng Li, Huaizhe Xu, Shijia Huang, Shilong Liu, Lionel M Ni, and Lei Zhang. Mp-former: Mask-piloted transformer for image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 8

[162] Haotian Zhang, Keyan Chen, Chenyang Liu, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Cdmamba: Incorporating local clues into mamba for remote sensing image binary change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 8

[163] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on com-*

puter vision and pattern recognition, pages 9759–9768, 2020. 7

[164] Tianwen Zhang, Xiaoling Zhang, Jianwei Li, Xiaowo Xu, Baoyou Wang, Xu Zhan, Yanqin Xu, Xiao Ke, Tianjiao Zeng, Hao Su, et al. Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 2021. 6, 7

[165] Tong Zhang, Peng Gao, Hao Dong, Yin Zhuang, Guanqun Wang, Wei Zhang, and He Chen. Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain. *Remote Sensing*, 14 (22):5675, 2022. 3

[166] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 9

[167] Xin Zhang, Xue Yang, Yuxuan Li, Jian Yang, Ming-Ming Cheng, and Xiang Li. Rsar: Restricted state angle resolver and rotated sar benchmark. *CVPR*, 2025. 6, 7

[168] 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. 1, 3, 12

[169] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 8

[170] Sijie Zhao, Xueliang Zhang, Pengfeng Xiao, and Guangjun He. Exchanging dual-encoder–decoder: A new strategy for change detection with semantic guidance and spatial localization. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. 8

[171] Sijie Zhao, Hao Chen, Xueliang Zhang, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. Rs-mamba for large remote sensing image dense prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 8

[172] Zhi Zheng, Yi Wan, Yongjun Zhang, Sizhe Xiang, Daifeng Peng, and Bin Zhang. Clnet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:247–267, 2021. 8

[173] Zhuo Zheng, Stefano Ermon, Dongjun Kim, Liangpei Zhang, and Yanfei Zhong. Changen2: Multi-temporal remote sensing generative change foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 8

[174] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 7