

# Harnessing Massive Satellite Imagery with Efficient Masked Image Modeling

Fengxiang Wang<sup>1</sup>, Hongzhen Wang<sup>2,3\*</sup>, Di Wang<sup>4,5</sup>, Zonghao Guo<sup>3</sup>, Zhenyu Zhong<sup>6</sup>,  
Long Lan<sup>1\*</sup>, Wenjing Yang<sup>1\*</sup>, Jing Zhang<sup>4\*</sup>

<sup>1</sup> College of Computer Science and Technology, National University of Defense Technology

<sup>2</sup> Xiaomi Corp. <sup>3</sup> Tsinghua University <sup>4</sup> School of Computer Science, Wuhan University

<sup>5</sup> Zhongguancun Academy <sup>6</sup> Nankai University

## Abstract

Masked Image Modeling (MIM) has become an essential method for building foundational visual models in remote sensing (RS). However, the limitations in size and diversity of existing RS datasets restrict the ability of MIM methods to learn generalizable representations. Additionally, conventional MIM techniques, which require reconstructing all tokens, introduce unnecessary computational overhead. To address these issues, we present a new pre-training pipeline for RS models, featuring the creation of a large-scale RS dataset and an efficient MIM approach. We curated a high-quality dataset named **OpticalRS-13M** by collecting publicly available RS datasets and processing them through exclusion, slicing, and deduplication. *OpticalRS-13M* comprises 13 million optical images covering various RS tasks, such as object detection and pixel segmentation. To enhance efficiency, we propose **SelectiveMAE**, a pre-training method that dynamically encodes and reconstructs semantically rich patch tokens, thereby reducing the inefficiencies of traditional MIM models caused by redundant background pixels in RS images. Extensive experiments show that *OpticalRS-13M* significantly improves classification, detection, and segmentation performance, while *SelectiveMAE* increases training efficiency over  $2\times$  times. This highlights the effectiveness and scalability of our pipeline in developing RS foundational models. The dataset, source code, and trained models will be released at [SelectiveMAE](https://github.com/MiliLab/SelectiveMAE) (<https://github.com/MiliLab/SelectiveMAE>).

## 1. Introduction

Over the past decade, advances in remote sensing (RS) technology and data acquisition have enhanced applications in ecosystem monitoring [3], natural disaster management [4], among others [5, 6]. These rely on essential capabilities such as scene classification [7, 8], object detection [9], change

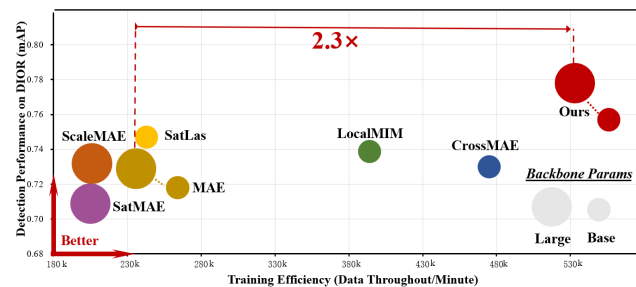


Figure 1. Advantages of SelectiveMAE on finetuning performance and pre-training efficiency. The accuracy is evaluated on the DIOR detection dataset [1] with different versions of ViT [2]. The training efficiency is measured by *Data Throughput/Minute*, i.e., the processed images per minute, on  $8 \times$  NVIDIA A100 GPUs.

detection [10], and semantic segmentation [11]. However, each task typically demands substantial computational resources to train specialized models and learn task-specific feature representations.

Recent advances in self-supervised learning, particularly Masked Image Modeling (MIM) techniques [13, 14], have significantly improved the pre-training of visual foundation models [15–20]. This progress has led to the emergence of remote sensing foundation models (RSFMs), which provide general feature representations and achieve state-of-the-art performance across various remote sensing tasks [21]. However, RSFMs still face two key challenges: (i) Compared to ImageNet-21k [22], existing RS datasets [23–26] are significantly smaller (approximately 1 million vs. 14 million samples), limiting the effective MIM training of large models; and (ii) these datasets primarily capture global scene semantics [23–25] but lack the diversity and fine-grained details essential for downstream tasks, hindering the generalization of learned representations.

To address these challenges, taking an example of visible light RS imagery, we propose a new pipeline to collect, create, and efficiently process a large RS dataset. Firstly, we reviewed publicly available remote sensing datasets from the past decade and selected them based on the DiRS (Diver-

\*Corresponding authors

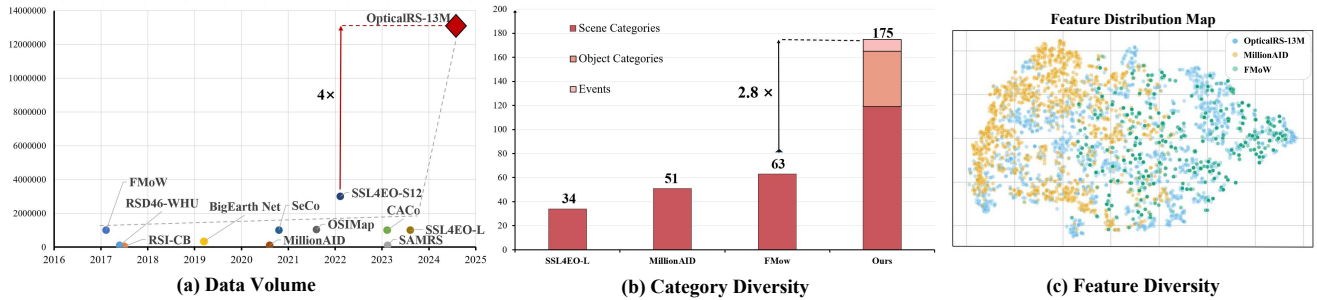


Figure 2. Advantages of OpticalRS-13M in data volume and diversity. (a) The volume of OpticalRS-13M surpasses at least  $4\times$  compared to existing RS datasets. (b) OpticalRS-13M includes significantly more sample categories than others, spanning both object and event types. y-axis denotes the data volume in (a-b). (c) OpticalRS-13M provides more diverse feature patterns than other large-scale RS datasets. Here, 1,000 samples are separately selected from different datasets, and their features from a pre-trained ViT-B are visualized using t-SNE [12].

sity, Richness, and Scalability) principle [23]. Nevertheless, issues such as inconsistent data sources, excessively large image sizes, and redundant pixels still exist. Therefore, we also apply exclusion, slicing, and deduplication processes to further improve data quality. Correspondingly, we obtain a large-scale RS dataset called **OpticalRS-13M** comprising 13 million images, which is designed to fully leverage the representation learning capabilities of MIM methods in RS applications. OpticalRS-13M exceeds previous RS datasets [21, 24, 27–42], being at least four times larger (Fig. 2 left). Moreover, OpticalRS-13M encompasses a wide range of diverse RS scenarios encountered in downstream tasks such as object-level detection and pixel-level segmentation (Fig. 2 right).

Despite substantial efforts in training RSFM using MIM methods [28, 33, 37, 43], the computational burden and slow convergence when employing MIM training on large-scale RS datasets cannot be ignored. Specifically, pre-training on 1 million RS samples requires 107 hours for the ViT-B [2] backbone on 8 Nvidia A100 GPUs [21]. This issue becomes even more pronounced when training on larger datasets, *e.g.*, OpticalRS-13M. In natural scene analysis, this issue has led to numerous studies [44–49] aimed at improving MIM training efficiency. One approach is to accelerate the token reconstruction process by using decoders with fewer parameters [44, 45]. Another approach is to reduce the number of visible patch tokens input into the vision encoder [46–48], speeding up feature extraction.

Conventional MIM approaches, following the encoding-then-decoding procedure, overlook the unique characteristics of RS images, which typically feature sparse foreground pixels and dense backgrounds [21, 31]. This raises two key questions about how to efficiently conduct MIM training in the RS field: 1) Is it necessary to reconstruct all the redundant background patches during the MIM decoding process? 2) Is there a feasible way to encode fewer image patches (*e.g.*,  $\leq 25\%$ ) to accelerate the convergence of MIM training? To address the first question, a measure-based selection process is needed to identify the appropriate patches for reconstruc-

tion. For the second question, the intuition is that the patch tokens used in the encoding-then-decoding procedure should effectively capture feature dependencies in RS images.

Regarding the above issues, in the second part of the pipeline, we introduce an MIM method called SelectiveMAE for efficiently processing RS images, which dynamically encodes and reconstructs patch tokens based on their semantic richness. Specifically, SelectiveMAE utilizes the Histogram of Oriented Gradients (HOG) algorithm to quantify the semantic richness of patches. Then, it selects a subset of patch tokens (*e.g.*,  $\leq 50\%$ ) with higher HOG values for feature encoding (*e.g.*,  $\leq 15\%$ ) and pixel reconstruction (*e.g.*,  $\leq 35\%$ ). However, using an extremely low ratio of visible patches during MIM training can lead to gradient explosion. To mitigate this, we designed a Progressive Semantic Token Selection (PSTS) module, which dynamically selects semantically relevant patch tokens during the entire training phase. In the beginning, SelectiveMAE encodes semantically rich tokens and reconstructs semantically similar ones to warm up the training process. As training advances, SelectiveMAE shifts to reconstructing high-semantic tokens from encoded lower-semantic ones to capture complementary semantic dependencies. This analogical-to-complementary strategy allows SelectiveMAE to efficiently and progressively learn robust representations of RS images while accelerating MIM convergence (Fig. 1). Our experiments reveal that 40% of RS image patches are sufficient to train a comparable model, offering new insights into MIM training on RS images.

In summary, our main contributions are as follows:

1. We introduce a new pipeline to collect, create, and efficiently process a large optical RS dataset for developing RS foundation models. Using this pipeline, we create the OpticalRS-13M dataset, which is a large-scale RS dataset comprising 13 million optical images with diversified coverage scenarios.
2. For this pipeline, we introduce SelectiveMAE, an efficient MIM method tailored for RS image pre-training. It significantly accelerates convergence and enhances representation learning compared to the original MIM

approach.

- Experiment results demonstrate the effectiveness and scalability of the proposed pipeline. OpticalRS-13M significantly enhances the performance of RS foundation models in downstream tasks, while SelectiveMAE achieves over  $2\times$  speedup in pre-training compared to MAE [13].

## 2. Related Work

**Remote Sensing Datasets.** In recent years, many RS datasets have been created for tasks such as scene classification [26, 50], object detection [51–53], and segmentation [54–56]. The availability of free, unlabeled satellite images has led to the development of large-scale RS datasets. Some works combine various sensor data to create extensive datasets. The SEN12MS dataset [57] consists of 180,662 triplets, each containing synthetic aperture radar (SAR) and multispectral Sentinel-2 image patches, and MODIS land cover maps. While SSL4EO-S12 [58], SSL4EO-L [59] and SatlasPretrain [60] contain millions of images, the major data type in these datasets are multispectral and SAR, with only a small fraction being RGB. There are also some large-scale RS visual-language datasets such as Skyscript [61] and RS5M [62], but they primarily focus on multimodal tasks. Currently, there are also some datasets that focus solely on visible light RS images. MillionAID [23], SeCo [25] and CACo [24] provide nearly a million images of the same location over different times. However, these datasets primarily target scene classification and often overlook fine-grained target information, limiting their utility for various downstream tasks. To address this gap, we introduce the OpticalRS-13M dataset, which is larger and more diverse. Using this dataset for pre-training, it significantly enhances performance across multiple downstream tasks.

**Masked Image Modeling.** Inspired by the success of Masked Language Modeling (MLM) in NLP [63, 64], MIM has been developed for visual pre-training [13, 14, 65–75]. MIM learns image representations by reconstructing masked tokens, focusing on various regression targets [76–80], masking strategies [81, 82], and reconstruction methods [83–86]. A major challenge for MIM is its high computational demand and lengthy pre-training times. To mitigate this, approaches such as asymmetric encoder-decoder strategies [44, 45], reducing input patches [46, 48], or the Difficulty-Flatten Loss [47] have been proposed. Additionally, CrossMAE [49] employs cross-attention between masked and visible tokens to enhance efficiency without sacrificing performance. However, these methods do not account for the unique characteristics of RS images, such as sparse foreground information and complex backgrounds. Considering these issues, we introduce an efficient MIM method named SelectiveMAE that significantly speeds up pre-training, enhancing the practicality of developing RS foundation models based on large-scale datasets.

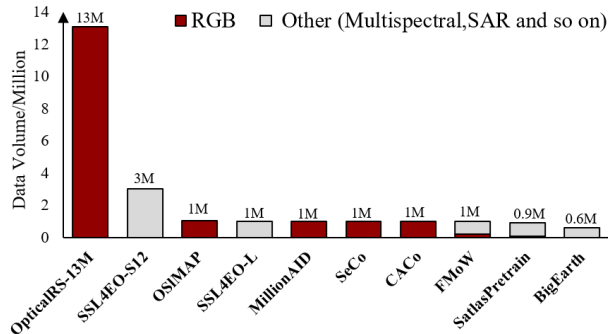


Figure 3. The comparison between OpticalRS-13M and existing RS large-scale datasets in terms of data volume and image type.

**Remote Sensing Foundation Models.** Despite the abundance of RS data, much of it remains unlabeled and thus inaccessible for supervised learning [87]. Self-supervised learning methods have recently been employed to extract representations from unlabeled RS data. Due to the inefficiency of designing pretext tasks and gathering required data for contrastive self-supervised methods [24, 25, 27, 88, 89], recent advancements have primarily centered around generative self-supervised methods, especially in MIM. Specifically, many studies aim to improve generative self-supervised algorithms by leveraging general image knowledge [28, 29, 31], scaling up parameter sizes [32, 43], integrating spatio-temporal information [33, 34, 36], encompassing geometric attributes [21, 90], handling multi-sensor data [91–95], and employing multi-scale concepts [37, 39, 40]. However, these methods have not effectively addressed the substantial computational burden associated with self-supervised pre-training in RS. In the paper, we propose a new pipeline that can collect, create, and efficiently process large amounts of optical RS data for developing RSFMs, enhancing the practicality of MIM pre-training on large-scale datasets.

## 3. Method

The proposed pipeline contains two critical components: dataset generation and efficient pre-training, which will be introduced in detail in the following text.

### 3.1. Dataset Curation

Recent progress in self-supervised pre-training for RSFMs is limited by the relatively small scale and diversity of existing RS datasets compared to natural scene datasets. To overcome this issue, we curate a large-scale RS dataset with diverse coverage scenarios, named OpticalRS-13M. Detailed data collection and preprocessing are presented in the supplementary material.

**Data Volume.** OpticalRS-13M comprises 13 million high-resolution visible light RS images, establishing a significant scale advantage over existing datasets, as illustrated in Figure 3. OpticalRS-13M contains 13,203,698

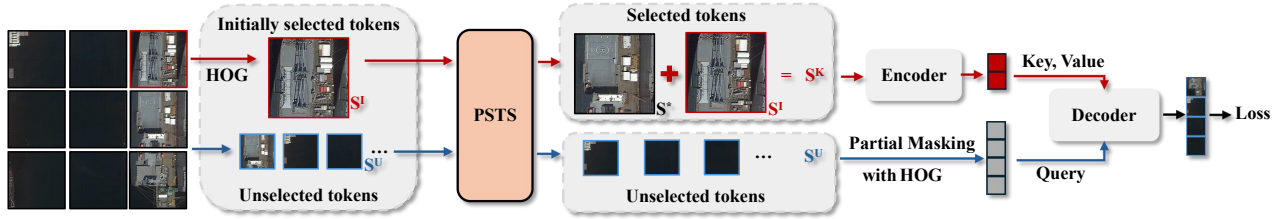


Figure 4. **Overview of SelectiveMAE.** It inputs fewer visible patches and reconstructs only partial patches to accelerate training efficiency.

images with a total of 2,630,362,174,503 pixels, averaging 199,214 pixels per image. We also offer a high-quality, lightweight version: OpticalRS-4M. OpticalRS-4M contains 3,920,829 images with a total of 688,459,799,359 pixels, averaging 175,590 pixels per image. While large-scale datasets such as SSL4EO-S12 [58] and SatLasPre-train [60] incorporate diverse data formats, *e.g.*, including multispectral imagery, OpticalRS-13M is uniquely focused on visible light data. The dataset is carefully curated from multi-sensor acquisitions, leveraging imagery from high-resolution satellite platforms such as WorldView [96], QuickBird [97], GeoEye [98], and others. This specialization positions OpticalRS-13M as the most extensive visible light RS image repository to date, offering unparalleled utility for training vision foundation models for RS applications.

**Dataset Diversity.** To demonstrate the broad scope of scenes and target categories covered by our dataset, we conducted a statistical analysis accompanied by a detailed label explanation, as shown in Fig. 2 (b) and (c). The analysis reveals that OpticalRS-13M comprises 12 main categories, each containing numerous subclasses. Unlike existing self-supervised optical RS datasets, OpticalRS-13M not only offers more comprehensive scene and target information but also introduces “Events” as a distinct high-level category. This category includes labels such as “Fire”, “Flood”, “Landslide”, “Post-Earthquake”, and so on. This expanded information enhances the versatility of OpticalRS-13M for a wider range of downstream tasks.

### 3.2. Efficient Pre-training

After obtaining a large-scale RS dataset, the next step is to conduct self-supervised pre-training. However, as the capacity of the dataset grows, vast computational and time costs are required for existing MIM approaches used in the RS community, *e.g.*, MAE [13]. To address this issue, our pipeline introduces an efficient MIM method, SelectiveMAE. In this section, we first review the preliminaries of MAE, then we present the details of SelectiveMAE.

#### 3.2.1. Preliminaries of MAE

**1) Masking.** Similar to supervised training of a standard ViT, MAE divides the image into regular, non-overlapping patches. It then samples a subset of these patches and masks the remaining ones. Typically, the masking ratio is 75%,

meaning only 25% of the patches are input to the encoder. This random sampling follows a uniform distribution according to the masking ratio. **2) MAE Encoder.** The encoder is a standard ViT applied only to the visible, unmasked patches. It linearly projects the patches, adds positional embeddings, and processes them through a series of transformer blocks. By operating on a smaller subset of patches, the encoder enables training of large models with reduced computational and memory requirements. **3) MAE Decoder.** The encoded tokens and masked tokens are fed into the decoder, which comprises transformer blocks with self-attention layers. The masked tokens are shared, learnable tensors enhanced with positional embeddings. The decoder, utilized only during pre-training, generates the output predictions for those masked tokens. **4) Reconstruction Target.** MAE predicts the pixel values for each masked patch, with each element in the decoder output representing a patch’s pixel value vector. The loss function computes the mean squared error between the reconstructed targets and original patches.

However, applying MAE for self-supervised pre-training results in considerable training costs in terms of time complexity, especially when working with large RS datasets. To address this, we observe that RS optical images often contain redundant information, which could be omitted during training to improve pre-training efficiency. Specifically, we address two issues: **1) Is it necessary to reconstruct all the masked patches given the redundancy in RS images?** **2) Can the visible patches input to the MAE encoder be further compressed to enhance acceleration?**

#### 3.2.2. Partial Reconstruction

For question 1, previous research [49] has shown that for general images, when MAE reconstructs 75% of the patches to calculate the loss, a specially designed decoder doesn’t need to fully reconstruct all remaining patches. In fact, reconstructing just 50% or even 25% of the patches can achieve similar performance and speed up training. However, for RS images, if we randomly sample patches and remove most for reconstruction, the reconstructed patches might not be semantically rich ones. Using only a random subset for reconstruction even degrades performance.

To address this issue, we propose selecting semantically rich patches for reconstruction instead of random selection. Specifically, given an input image  $x \in \mathbb{R}^{H \times W \times C}$ , it is

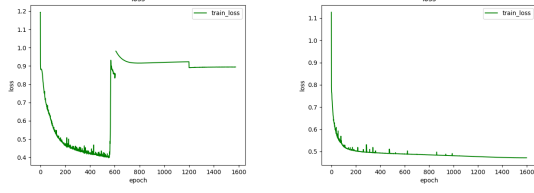


Figure 5. Effectiveness of PSTS. Left: Using only 40% of patches for encoding and reconstruction often leads to gradient explosions. Right: The training loss after adopting PSTS.

reshaped into  $N = (H \times W)/p^2$  non-overlapping patches  $x^p \in \mathbb{R}^{N \times (p^2 C)}$ , where  $p$  is the patch size,  $(H, W)$  is the size of the input image, and  $C$  is the number of channels. These patches  $\{x_i^p\}_{i=1}^N$  are then linearly mapped to patch embeddings. To retain positional information, positional embeddings are added to the patches. We select a portion of the patches to input to the encoder based on the masking ratio  $m \in [0, 1]$  ( $m = 85\%$  by default), as detailed in Sec. 3.2.1. The remaining patches serve as reconstruction targets for the decoder.

Unlike MAE’s masking ratio, we introduce a new reconstruction ratio  $r$ , the proportion of pixels to be reconstructed, denoted as  $r \in [0, m]$  ( $r = 25\%$  by default). We compute the HOG features  $HOG(\cdot)$  of the remaining patches and select those with the high HOG feature values according to the reconstruction ratio  $r$ , rather than using all patches. The process can be formulated as:

$$token_R = \{x_i^p | i \in \text{top}_{\lfloor r \times N \rfloor}(HOG(\{x_i^p\}_{i=1}^{m \times N}))\}, \quad (1)$$

where  $token_R$  denotes the selected mask tokens for reconstruction and  $\text{top}_n(\cdot)$  denotes the index set of the selected top  $n$  tokens. The decoder uses a lightweight design based on cross-attention following CrossMAE [49]. Experimental results in supplementary material show that this partial reconstruction strategy significantly increases the training throughput without affecting the learned representations.

### 3.2.3. Progressive Semantic Token Selection

For the second question, we initially tried a naive approach by increasing the masking ratio to 85%, meaning only 15% of the patches in each RS image were input to the encoder while keeping a 25% reconstruction ratio as proposed in Sec. 3.2.2. However, during training, this setup often led to issues like gradient explosions or loss divergence, as shown in Fig. 5 left. The figure shows that using an extremely low portion of patches for encoding and reconstruction caused unstable training due to gradient explosions.

How can we achieve acceleration at a high masking ratio (e.g., 85%) while ensuring MAE completes the pre-training task for RS images? Inspired by curriculum learning [99–101], which follows the principle of learning from easy to hard, we introduce the Progressive Semantic Token Selection (PSTS) module for patch selection, as depicted in Fig. 4.

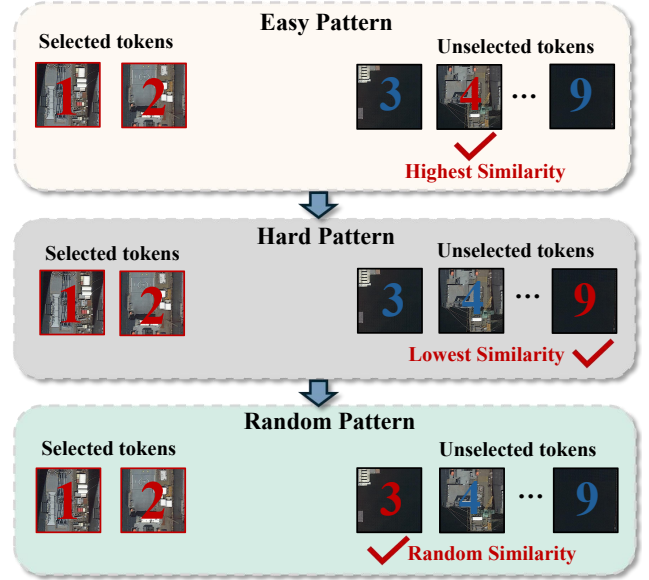


Figure 6. Illustration of the patch selection process in PSTS.

In this module, we begin by selecting a limited number of patches and then select additional patches based on the similarity measure in the training epoch, dynamically transitioning from easily learned, semantically similar patches to more challenging, complementary ones.

For initialization, we employ a HOG selection strategy to choose the initial patch set from  $S^N = \{x_i^p\}_{i=1}^N$ , with a proportion  $s \in [0, (1 - m)/2]$ .  $m$  is the mask ratio. We define the initial set of selected token as:

$$S^I = \{S^N(i) | i \in \text{top}_{\lfloor s \times N \rfloor}(HOG(S^N))\}. \quad (2)$$

We select  $\lfloor s \times N \rfloor$  tokens with the maximum HOG feature values (i.e.,  $\text{top}_{\lfloor s \times N \rfloor}$ ) from the original token set to form the initial token set. This simple yet effective strategy ensures that semantically rich tokens are selected.

After token initialization, we incrementally increase the number of tokens to guide training from easier to more challenging examples, while maintaining the final masking ratio for selected tokens, as outlined in Algorithm 1. Given the high HOG feature values of the initial token sets, nearby tokens selected by PSTS also exhibit high HOG values. In partial reconstruction method in Section 3.2.2, we filter unselected tokens to retain those with high HOG values as reconstruction targets. Thus, selecting nearby tokens brings encoding tokens and reconstruction targets closer, streamlining training. In contrast, if a token’s feature distribution differs significantly from others, we consider these more challenging patches to learn.

Specifically, we select tokens from  $S^U$  based on  $S^I$ . First, we use  $\mathbb{S}^I \in \mathbb{R}^{|S^I| \times d}$  and  $\mathbb{S}^U \in \mathbb{R}^{|S^U| \times d}$  to denote the matrix representation of the initial token set  $S^I$  and the unselected token set  $S^U$ , where  $|\cdot|$  represents the number of tokens and  $d$  the feature dimension after the embedding

---

**Algorithm 1** Progressive Semantic Token Selection

---

**Require:** Number of training epochs  $T$  and total training stages  $N_g$  (i.e.,  $T/N_g$  epochs for each stage), masking rate  $m$ , input dataset  $\mathcal{X}$

**Ensure:** Obtain the selected tokens set  $S^K$  and update  $S^U$  in each epoch

- 1: **for**  $t \leftarrow 1$  **to**  $T$  **do**
  - 2:   Sample data sample from  $\mathcal{X}$ , feed-forwarded through the embedding to obtain the output token set  $S_N$
  - 3:    $s \leftarrow \frac{1}{2}(1 - m)$
  - 4:   Obtain  $S^I$  via Eq. (2) and initialize  $S^U$
  - 5:   Obtain the current training stage  $\zeta = \lceil N_g * t/T \rceil$
  - 6:   Calculate  $distance(S^U \rightarrow S^I)_i$  for  $i \in \{1, \dots, |S^U|\}$  via Eq. (3) and Eq. (4)
  - 7:   Obtain  $S^*$  and update  $S^K, S^U$  via Eq. (5) and Eq. (6)
  - 8: **end for**
- 

layer. We use Cosine Distance to measure the dissimilarity between the tokens in these two sets:

$$\mathcal{D}(S^U, S^I) = 1 - \cos \langle S^U, S^I \rangle = 1 - \frac{S^U (S^I)^T}{\|S^U\| \cdot \|S^I\|} \quad (3)$$

where  $\mathbf{1}$  is an all-one matrix.  $\mathcal{D}(S^U, S^I) \in \mathbb{R}^{|S^U| \times |S^I|}$  represents the pairwise distances between tokens in  $S^U$  and  $S^I$ . Next, we define the distance between the tokens in  $S^U$  to the initial token set  $S^K$  based on the selection criteria in each training stage as follows:

$$distance(S^U \rightarrow S^I)_i = \begin{cases} -\min_j (\mathcal{D}(S^U, S^I)_{i,j}), & \zeta = 1 \\ \max_j (\mathcal{D}(S^U, S^I)_{i,j}), & \zeta = 2 \\ \text{random}_j (\mathcal{D}(S^U, S^I)_{i,j}), & \text{otherwise} \end{cases}, \quad (4)$$

where  $i \in \{1, \dots, |S^U|\}$ ,  $j \in \{1, \dots, |S^I|\}$ , and  $\zeta$  represents the training stage depending on the number of epochs. Finally, we sample  $\lfloor N \times (1 - m - s) \rfloor$  tokens from  $S^U$  and add them with  $S^I$  to form  $S^K$ , which can be formulated as follows:

$$S^* = \{S^U(i) | i \in \text{top}_{\lfloor N \times (1 - m - s) \rfloor} (distance(S^U \rightarrow S^I)_i)\}, \quad (5)$$

$$S^K = S^I \cup S^*, \quad S^U = S^U \setminus S^*, \quad (6)$$

where  $S^*$  represents the selected token from  $S^U$ . The operations in Eq. (5) and Eq. (6) are performed in each training epoch. This process is summarized in Algorithm 1. Fig. 6 further provides an example to illustrate which patches should be selected at different stages, helping to facilitate readers' understanding.

## 4. Experiment

In this section, we performed comprehensive comparative experiments to evaluate the performance of the proposed

method on various downstream tasks, including RS scene classification, object detection, and semantic segmentation. Then, we validated the effectiveness of the OpticalRS-13M dataset and the SelectiveMAE method through ablation experiments. We also conducted further experiments to test the scalability of the pipeline. All experiments are performed by fine-tuning. **The pre-training and fine-tuning settings of our methods can be found in the supplementary material.**

### 4.1. Main Results

We compare SelectiveMAE to state-of-the-art RSFMs. In addition to benchmarking against MIM-based approaches such as SatMAE [33] and ScaleMAE [37], we also compare with contrastive learning-based methods, including GASSL [27] and SeCo [25], as well as advanced supervised learning models like SatLas [60]. Notably, these comparative methods were specifically designed for increasing performance rather than accelerating pre-training. **The Scene Classification and Detection comparison results are taken directly from SkySense [92], while those for Semantic Segmentation are exactly sourced from MTP [38].** Although their speed performance was not reported, we expect them to lag behind the baseline MAE. The performance of ViT-B and ViT-L on various downstream tasks after pre-training is presented in Table 1. The results indicate that after pre-training on OpticalRS-13M, SelectiveMAE outperforms other RSFMs. Notably, it not only outperforms baseline (MAE) but also achieves this at 2.1 times the speed (556/264). This makes SelectiveMAE particularly well-suited for large-scale datasets, offering significant time savings during training on OpticalRS-13M.

**Scene Classification.** We first assess the pre-trained model's performance on the scene classification task to evaluate its overall representational capability without additional decoders. We use two scene classification datasets: AID [102] and RESISC-45 [103], following training details and train-test splits as in [21, 37]. Table 1 shows that SelectiveMAE performs competitively against other pre-training methods on both datasets. Additionally, when scaled to ViT-L, our model outperforms OREOLE [43] and other competitors, indicating the efficacy of SelectiveMAE pre-training on the OpticalRS-13M dataset, which enables strong feature representation learning while maintaining efficient scalability with increasing model size.

**Horizontal & Oriented Object Detection.** We utilized the well-established DIOR dataset for horizontal object detection [1] and its improved variant DIOR-R for oriented object detection [104]. Following the methodologies of previous works [21, 41], we maintained consistent experimental setups, employing Faster-RCNN [107] and Oriented-RCNN [108] as detectors for each dataset. The results are summarized in Table 1. Our approach, utilizing a ViT-B backbone, demonstrates competitive or superior performance compared to other methods with a Swin-B backbone, such

Table 1. Performance comparison results of different models. “TR” represents the ratio of training data to the entire dataset. The first and second scores are marked by **bold** and **blue**, respectively. † means pre-training on 4 million images sampled from OpticalRS-13, with the epoch is set to 800. The overall accuracy is adopted as the metric for the scene classification task.

Model	Backbone	Params (M)	Data Throughput /Minute	Scene Classification		Object Detection		Semantic Segmentation	
				AID [102]	RESISC-45 [103]	DIOR [1]	DIOR-R [104]	LoveDA [55]	SpaceNetv1 [56]
				TR=20%/50%	TR=10%/20%	mAP <sub>50</sub>	mAP <sub>50</sub>	mIoU	mF1
SeCo [25]	ResNet-50 [105]	26	-	93.47/95.99	89.64/92.91	-	-	43.63	77.09
GASSL [27]	ResNet-50 [105]	26	-	93.55/95.92	90.86/93.06	67.40	65.65	48.76	78.51
TOV [31]	ResNet-50 [105]	26	-	95.16/97.09	90.97/93.79	70.16	66.33	49.70	-
CACo [24]	ResNet-50 [105]	26	-	90.88/95.05	88.28/91.94	66.91	64.10	48.89	77.94
SatMAE [33]	ViT-L [2]	307	205k	95.02/96.94	91.72/94.10	70.89	65.66	-	78.07
ScaleMAE [37]	ViT-L [2]	307	206k	96.44/97.58	92.63/95.04	73.81	66.47	-	-
SSL4EO [58]	ViT-S [2]	22	-	91.06/94.74	87.60/91.27	64.82	61.23	-	-
RingMo [41]	Swin-B [106]	88	-	96.90/98.34	94.25/95.67	75.90	-	-	-
SatLas [60]	Swin-B [106]	88	243k	94.96/97.38	92.16/94.70	74.10	67.59	-	-
GFM [28]	Swin-B [106]	88	-	95.47/97.09	92.73/94.64	72.84	67.67	-	-
RVSA [21]	ViT-B+RVSA [21]	86	-	97.03/98.50	93.93/95.69	75.80	68.06	51.95	-
OREOLE [43]	ViT-G [43]	914	-	96.71/ -	- / -	77.40	<b>71.31</b>	<b>54.00</b>	-
MAE [13]†	ViT-B [2]	86	264k	96.58/98.02	92.44/94.43	75.40	67.35	52.80	79.41
SelectiveMAE †	ViT-B [2]	86	556k	96.90/98.12	93.35/94.58	75.70	67.78	53.05	<b>79.50</b>
SelectiveMAE†	ViT-L [2]	307	533k	<b>97.25/98.50</b>	<b>94.57/95.77</b>	<b>77.80</b>	70.31	<b>54.31</b>	79.46
SelectiveMAE	ViT-B [2]	86	556k	97.10/98.28	93.70/95.48	75.80	67.69	52.68	79.44
SelectiveMAE	ViT-L [2]	307	533k	<b>97.49/98.52</b>	<b>94.73/96.36</b>	<b>78.70</b>	<b>71.75</b>	53.92	<b>79.48</b>

as RingMo [41]. When using the larger ViT-L backbone, SelectiveMAE shows enhanced performance across both detection datasets, even outperforming OREOLE [43], which has close to 1B parameters, underscoring the excellent scalability of our method.

**Semantic Segmentation.** We further evaluate the performance of the pre-trained model on pixel-level perception tasks, particularly semantic segmentation, using two well-known RS datasets: LoveDA [55] and SpaceNetv1 [56]. Our implementation follows [21], utilizing UperNet [109] as the segmentation framework. Table 1 demonstrates the clear superiority of SelectiveMAE over its competitors in semantic segmentation tasks. SelectiveMAE focus on semantically rich patches, such as the boundaries between foreground objects and background stuff, leading to better representation learning for segmentation.

## 4.2. Ablation Study

To validate the efficacy of SelectiveMAE, we conducted ablation experiments about HOG Selecting, Similarity Measure and Selection Pattern. To further verify the pipeline’s effectiveness, we conducted separate evaluations of both the OpticalRS-13M dataset generated by the pipeline and the accelerated performance of SelectiveMAE. Additionally, we compared the SelectiveMAE with the CrossMAE, a MAE-based method for natural images. More ablation experiment results are presented in the supplementary material.

**Alternatives of HOG Selecting.** To verify the efficacy of HOG in token selecting, we conduct two ablation experiments using MillionAID dataset [23] to replace HOG: (1) we used a pretrained Swin-B [106] model and applied k-nearest neighbor clustering to the extracted features. (2) We employed a lightweight feature extractor network based on

Table 2. Ablation experiment results of HOG selection through pre-training on the MillionAID dataset for 800 epochs.

Method	Params	Data Throughput / Minute	AID		RESISC-45	
			TR=20%/50%	TR=10%/20%	TR=20%/50%	TR=10%/20%
Adamae [110]	2.36M	498k	88.78/91.25	85.72/87.44		
Swin-B	88M	356k	93.21/96.48	89.94/93.72		
HOG	-	556k	93.17/96.12	89.21/92.31		

Table 3. Ablation experiment results of different similarity measures and selection patterns in PSTS through pre-training on the MillionAID dataset for 800 epochs.

Selection Pattern	Similarity Measure	AID		RESISC-45	
		TR=20%/50%	TR=10%/20%	TR=20%/50%	TR=10%/20%
near-far-random	Euclidean distance	93.02/95.96	88.97/92.07		
near-far-random	Manhattan distance	92.92/95.89	89.17/91.92		
far-near-random	Cosine distance	90.12/92.78	85.81/88.80		
near-far-random	Cosine distance	93.17/96.12	89.21/92.31		

Adamae [110]. Results in the Table 2 demonstrate HOG’s pre-training speed advantage. While using Swin-B slightly outperforms HOG, it is highly inefficient, and pre-training a foundation model like Swin requires massive data and computational resources. For RS images with abundant low-level features, HOG effectively selects tokens while ensuring substantially higher training speeds.

### Ablation of Similarity Measure and Selection Pattern.

To comprehensively assess the PSTS, we tested different Similarity Measures and the Selection Pattern. The results in the Table 3 reveal that different similarity measures had minimal impact on performance. However, changing the selection pattern led to a performance drop, confirming the effectiveness of the PSTS.

**Efficacy of OpticalRS-13M.** As illustrated in Table 4, when pre-training with a randomly sampled subset of 1 mil-

Table 4. Performance comparison of using different pre-training settings on datasets, methods and epochs, where the ViT-B [2] is adopted as the network.

Dataset	Method	Image Number	Epoch	AID [102]	RESISC-45 [103]
				TR=20%/50%	TR=10%/20%
<i>Different Datasets</i>					
MillionAID [23]	MAE	1 million	800	94.92/97.38	89.20/93.60
OpticalRS-13M	MAE	1 million	800	<b>96.26/97.98</b>	<b>91.53/93.88</b>
<i>Equivalent Data Throughput</i>					
OpticalRS-13M	MAE	2 million	400	96.64/98.10	91.80/94.31
OpticalRS-13M	MAE	3 million	267	<b>96.67/98.18</b>	<b>92.24/94.41</b>
OpticalRS-13M	MAE	4 million	200	96.10/98.03	<b>92.38/94.30</b>
OpticalRS-13M	MAE	8 million	100	96.58/ <b>98.26</b>	91.83/93.99
OpticalRS-13M	MAE	13 million	67	96.28/98.06	91.41/93.60
OpticalRS-13M	SelectiveMAE	1 million	800	96.29/97.78	91.41/93.48
OpticalRS-13M	SelectiveMAE	4 million	200	95.96/ <b>98.06</b>	<b>92.06/94.05</b>
OpticalRS-13M	SelectiveMAE	13 million	67	<b>96.31/97.95</b>	91.88/93.76
<i>More Epoch</i>					
OpticalRS-13M	MAE	13 million	67	96.28/98.06	91.41/93.60
OpticalRS-13M	MAE	13 million	100	96.34/98.13	91.79/93.85
OpticalRS-13M	MAE	13 million	200	96.51/98.22	92.19/94.26
OpticalRS-13M	MAE	13 million	800	<b>97.10/98.28</b>	<b>93.70/95.48</b>

Table 5. The efficiency comparison between SelectiveMAE and the baseline method MAE [13] across different backbones on the part of OpticalRS-13M dataset [23] with 800 epochs. The memory is measured on a single NVIDIA A100 GPU with a batch size of 256.

Model	Backbone	Data Volume	Training Time (h)	GPU Memory (MB)
MAE	ViT-B	1 million	51	17628
SelectiveMAE	ViT-B	1 million	<b>24 (2.1×, -27)</b>	<b>9570 (1.8×)</b>
MAE	ViT-L	1 million	57	30530
SelectiveMAE	ViT-L	1 million	<b>25 (2.3×, -32)</b>	<b>18790 (1.6×)</b>
MAE	ViT-B	4 million	177	17628
SelectiveMAE	ViT-B	4 million	<b>86 (2.1×, -81)</b>	<b>9570 (1.8×)</b>

lion images from OpticalRS-13M, the model achieves superior performance compared to MillionAID [23], underscoring the efficacy of OpticalRS-13M for representation learning. To further assess dataset diversity, we conducted experiments with equivalent data throughput during pre-training. The results in Table 4 reveal that optimal performance is attained by varying training configurations rather than utilizing the entire 13 million image corpus, regardless of the pre-training methodology. In our opinion, this observation highlights the inherent diversity of OpticalRS-13M, as training with fewer epochs proved insufficient for the MIM method to fully exploit the dataset’s potential, leading to model underfitting. To validate this hypothesis, we extended the training schedule, and the experimental results presented in the last part of Table 4 confirm our claim.

**Efficiency Advantage of SelectiveMAE.** To further highlight the efficiency advantage of SelectiveMAE, we evaluate the training time and memory footprint during pre-training of different backbones on the part of OpticalRS-13M dataset, as shown in Table 5 (the corresponding accuracies are presented in Table 1). It can be seen that, since only 40% of the image patches (15% for the encoder and 25% for the decoder) are involved in the calculation, our methods present more than doubled training acceleration compared to the vanilla MAE baseline, where the acceleration advantage is more significant for larger models. It is worth noting that SelectiveMAE also reduces memory footprint by using fewer

Table 6. Performance comparison between our method and CrossMAE, where 400 million images of OpticalRS-13M are randomly sampled for pre-training 800 epochs.

Method	Data Throughput / Minute	AID	RESISC-45
		TR=20%/50%	TR=10%/20%
CrossMAE(ViT-B)	475k	96.12/97.18	92.08/94.12
SelectiveMAE(ViT-B)	556k	96.78/98.12	93.35/94.58

patches. To further evaluate the scalability of our pipeline, we organized datasets containing 4 million samples and conducted experiments. The results in Table 5 show that, when the dataset was expanded to 4 million samples, the time savings increased from 27 (51-24) to 81 hours. These findings demonstrate that our pipeline scales effectively with larger datasets, offering significant pre-training acceleration.

**Comparison with CrossMAE [49].** CrossMAE, an MAE-based self-supervised pretraining method, is originally designed for natural images. Since SelectiveMAE adopts the same decoder as CrossMAE [49] during pre-training, and incorporates a HOG-based strategy tailored to RS images inspired by its partial reconstruction concept, as detailed in Section 3.2.2. In this section, we have also complemented related comparison experiments. Nevertheless, due to CrossMAE has not been pretrained in the RS field, we conducted a fair evaluation through pretraining it on the OpticalRS-13M dataset using the official code. Table 6 shows that SelectiveMAE outperforms CrossMAE in both accuracy and efficiency. By integrating RS-specific enhancements, PSTS and partial reconstruction, SelectiveMAE proves to be a highly effective self-supervised learning approach for RS.

## 5. Conclusion

In this paper, we introduce a new pre-training pipeline for RS models, featuring the creation of a large-scale RS dataset and an efficient MIM approach. We first curated OpticalRS-13M, a large-scale optical remote sensing dataset for unsupervised learning. Unlike previous RS datasets, OpticalRS-13M offers a larger and more diverse image set with fine-grained details relevant to downstream tasks. Benchmarking representative MIM methods on OpticalRS-13M highlights its advantages in these tasks. Then, we present SelectiveMAE to reduce the computational overhead of MIM training on large-scale RS datasets. This efficient MIM method dynamically encodes and reconstructs tokens based on their semantic richness. SelectiveMAE significantly accelerates training, demonstrating that using only 40% RS image patches is sufficient for training a comparable MIM model. Extensive experiments show that OpticalRS-13M significantly contributes to improving classification, detection, and segmentation performance, while SelectiveMAE achieves over a 2× speedup along with GPU memory savings, highlighting the effectiveness and scalability of our pipeline in developing RS foundational models.

**Acknowledgements:** We gratefully acknowledge Prof. Zhiyuan Liu and Prof. Maosong Sun for their helpful discussions and support during this research. This work was partially supported by the National Natural Science Foundation of China (No. 62372459, No.62376282 and No. 624B2109) and the Major Special Project of China Innovation Challenge (Ningbo) under Grant 2024T008.

## References

- [1] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 1, 6, 7
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 1, 2, 7, 8
- [3] Nathalie Pettorelli, Henrike Schulte to Bühne, Ayesha Tulloch, Grégoire Dubois, Cate Macinnis-Ng, Ana M Queirós, David A Keith, Martin Wegmann, Franziska Schrod, Marion Stellmes, et al. Satellite remote sensing of ecosystem functions: opportunities, challenges and way forward. *Remote Sensing in Ecology and Conservation*, 4(2):71–93, 2018. 1
- [4] Olalekan Mumin Bello and Yusuf Adedoyin Aina. Satellite remote sensing as a tool in disaster management and sustainable development: towards a synergistic approach. *Procedia-Social and Behavioral Sciences*, 120:365–373, 2014. 1
- [5] Liang Huang, Fengxiang Wang, Yalun Zhang, and Qingxia Xu. Fine-grained ship classification by combining cnn and swin transformer. *Remote Sensing*, 14(13):3087, 2022. 1
- [6] Fengxiang Wang, Deying Yu, Liang Huang, Yalun Zhang, Yongbing Chen, and Zhiguo Wang. Fine-grained ship image classification and detection based on a vision transformer and multi-grain feature vector fpn model. *Geo-spatial Information Science*, pages 1–22, 2024. 1
- [7] Tongdi He and Shengxin Wang. Multi-spectral remote sensing land-cover classification based on deep learning methods. *The Journal of Supercomputing*, 77(3):2829–2843, 2021. 1
- [8] Gong Cheng, Xingxing Xie, Junwei Han, Lei Guo, and Gui-Song Xia. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3735–3756, 2020. 1
- [9] 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 (CVPR)*, pages 1829–1838, June 2022. 1
- [10] Curtis E Woodcock, Thomas R Loveland, Martin Herold, and Marvin E Bauer. Transitioning from change detection to monitoring with remote sensing: A paradigm shift. *Remote Sensing of Environment*, 238:111558, 2020. 1
- [11] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169:114417, 2021. 1
- [12] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2
- [13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 2022. 1, 3, 4, 7, 8
- [14] 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. 1, 3
- [15] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1
- [16] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, 131(5):1141–1162, 2023.
- [17] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [18] Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [19] Fengxiang Wang, Wanrong Huang, Shaowu Yang, Qi Fan, and Long Lan. Learning to learn better visual prompts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5354–5363, 2024.
- [20] Jingyi Wang, Xiaobo Xia, Long Lan, Xinghao Wu, Jun Yu, Wenjing Yang, Bo Han, and Tongliang Liu. Tackling noisy labels with network parameter additive decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1
- [21] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. doi: 10.1109/TGRS.2022.3222818. 1, 2, 3, 6, 7
- [22] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lih Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 1
- [23] 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. 1, 2, 3, 7, 8
- [24] 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. 2, 3, 7

- [25] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodríguez. Seasonal Contrast: Un-supervised Pre-Training From Uncurated Remote Sensing Data. In *ICCV*, 2021. 1, 3, 6, 7
- [26] Gencer Sumbul, Arne de Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mário Caetano, Begüm Demir, and Volker Markl. BigEarthNet-MM: A Large-Scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, 2021. 1, 3
- [27] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-Aware Self-Supervised Learning. In *ICCV*, 2021. 2, 3, 6, 7
- [28] 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*, pages 16806–16816, 2023. 2, 3, 7
- [29] Ziyue Huang, Mingming Zhang, Yuan Gong, Qingjie Liu, and Yunhong Wang. Generic knowledge boosted pre-training for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3
- [30] 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, 2024.
- [31] 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*, 2023. 2, 3, 7
- [32] Keumgang Cha, Junghoon Seo, and Taekyung Lee. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*, 2023. 3
- [33] 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. 2, 3, 6, 7
- [34] Fanglong Yao, Wanxuan Lu, Heming Yang, Liangyu Xu, Chenglong Liu, Leiye Hu, Hongfeng Yu, Nayu Liu, Chubo Deng, Deke Tang, et al. Ringmo-sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 3
- [35] Wentao Jiang, Jing Zhang, Di Wang, Qiming Zhang, Zeng-mao Wang, and Bo Du. LeMeViT: Efficient vision transformer with learnable meta tokens for remote sensing image interpretation. In *International Joint Conference on Artificial Intelligence*, 2024.
- [36] Wenyuan Li, Keyan Chen, and Zhenwei Shi. Geographical supervision correction for remote sensing representation learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. 3
- [37] 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. *CoRR*, abs/2212.14532, 2023. 2, 3, 6, 7
- [38] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, and Liangpei Zhang. Mtp: Advancing remote sensing foundation model via multi-task pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pages 1–24, 2024. 6
- [39] Maofeng Tang, Andrei Liviu Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3
- [40] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwar, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *CVPR*, 2024. 3
- [41] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 6, 7
- [42] 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. doi: 10.1109/TGRS.2023.3268232. 2
- [43] Philippe Dias, Aristeidis Tsaris, Jordan Bowman, Abhishek Potnis, Jacob Arndt, H. Lexie Yang, and Dalton Lunga. Oreole-fm: Successes and challenges toward billion-parameter foundation models for high-resolution satellite imagery. In *ACM International Conference on Advances in Geographic Information Systems*, 2024. 2, 3, 6, 7
- [44] Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. Siamese masked autoencoders. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3
- [45] Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2122–2131, 2023. 2, 3
- [46] Jin Li, Yaoming Wang, Xiaopeng Zhang, Yabo Chen, Dongsheng Jiang, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Qi Tian. Progressively compressed auto-encoder for self-supervised representation learning. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3
- [47] Jun Chen, Ming Hu, Boyang Li, and Mohamed Elhoseiny. Efficient self-supervised vision pretraining with local masked reconstruction. *arXiv preprint arXiv:2206.00790*, 2022. 3
- [48] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Yunhe Wang, and Chang Xu. Fastmim: Expediting masked image modeling pre-training for vision. *arXiv preprint arXiv:2212.06593*, 2022. 2, 3

- [49] Letian Fu, Long Lian, Renhao Wang, Baifeng Shi, Xudong Wang, Adam Yala, Trevor Darrell, Alexei A Efros, and Ken Goldberg. Rethinking patch dependence for masked autoencoders. *arXiv preprint arXiv:2401.14391*, 2024. 2, 3, 4, 5, 8
- [50] Ben G Weinstein, Sergio Marconi, Stephanie A Bohlman, Alina Zare, Aditya Singh, Sarah J Graves, and Ethan P White. A remote sensing derived data set of 100 million individual tree crowns for the national ecological observatory network. *Elife*, 10:e62922, 2021. 3
- [51] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 3
- [52] Jan Gasienica-Jozkoway, Mateusz Knapik, and Bogusław Cyganek. An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance. *Integrated Computer-Aided Engineering*, 28(3):221–235, 2021.
- [53] Long Lan, Fengxiang Wang, Shuyan Li, Xiangtao Zheng, Zengmao Wang, and Xinwang Liu. Efficient prompt tuning of large vision-language model for fine-grained ship classification. *arXiv preprint arXiv:2403.08271*, 2024. 3
- [54] Gaetan Bahl, Mehdi Bahri, and Florent Lafarge. Single-shot end-to-end road graph extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1403–1412, 2022. 3
- [55] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *NeurIPS Track on Datasets and Benchmarks*, volume 1, 2021. 7
- [56] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 3, 7
- [57] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms—a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019. 3
- [58] 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*, 11(3):98–106, 2023. 3, 4, 7
- [59] Adam Stewart, Nils Lehmann, Isaac Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. Ssl4eo-l: Datasets and foundation models for landsat imagery. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [60] 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. 3, 4, 6, 7
- [61] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5805–5813, Mar. 2024. 3
- [62] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv preprint arXiv:2306.11300*, 2023. 3
- [63] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3
- [64] Huayue Cai, Xiang Zhang, Long Lan, Guohua Dong, Chuanfu Xu, Xinwang Liu, and Zhigang Luo. Learning deep discriminative embeddings via joint rescaled features and log-probability centers. *Pattern Recognition*, 114:107852, 2021. 3
- [65] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 3
- [66] Xiao Teng, Long Lan, Jing Zhao, Xueqiong Li, and Yuhua Tang. Highly efficient active learning with tracklet-aware co-cooperative annotators for person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [67] Long Lan, Xiao Teng, Jing Zhang, Xiang Zhang, and Dacheng Tao. Learning to purification for unsupervised person re-identification. *IEEE Transactions on Image Processing*, 2023.
- [68] Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *Machine Intelligence Research*, 21:617–630, 2024.
- [69] Yifan Zhao, Jia Li, and Yonghong Tian. Parsing objects at a finer granularity: A survey. *Machine Intelligence Research*, 21(3):431–451, 2024.
- [70] Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, and Zhaoxiang Zhang. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10375–10385, 2023.
- [71] Haochen Wang, Junsong Fan, Yuxi Wang, Kaiyou Song, Tiancai Wang, Xiangyu Zhang, and Zhaoxiang Zhang. Bootstrap masked visual modeling via hard patches mining. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [72] Zhirui Gao, Renjiao Yi, Yaqiao Dai, Xuening Zhu, Wei Chen, Chenyang Zhu, and Kai Xu. Curve-aware gaussian splatting for 3d parametric curve reconstruction, 2025. URL <https://arxiv.org/abs/2506.21401>.
- [73] Zhirui Gao, Renjiao Yi, Yuhang Huang, Wei Chen, Chenyang Zhu, and Kai Xu. Self-supervised learning of hybrid part-aware 3d representation of 2d gaussians and superquadrics, 2025. URL <https://arxiv.org/abs/2408.10789>.
- [74] Zhirui Gao, Renjiao Yi, Chenyang Zhu, Ke Zhuang, Wei Chen, and Kai Xu. Generic objects as pose probes for few-shot view synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

- [75] Zhirui Gao, Renjiao Yi, Zheng Qin, Yunfan Ye, Chenyang Zhu, and Kai Xu. Learning accurate template matching with differentiable coarse-to-fine correspondence refinement. *Computational Visual Media*, 10(2):309–330, 2024. 3
- [76] 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. 3
- [77] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021.
- [78] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.
- [79] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2021.
- [80] Hao Liu, Xinghua Jiang, Xin Li, Antai Guo, Yiqing Hu, Deqiang Jiang, and Bo Ren. The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1649–1656, 2023. 3
- [81] Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, pages 20026–20040. PMLR, 2022. 3
- [82] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35:14290–14302, 2022. 3
- [83] Shubham Tulsiani and Abhinav Gupta. Pixeltransformer: Sample conditioned signal generation. In *International Conference on Machine Learning*, pages 10455–10464. PMLR, 2021. 3
- [84] Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, and Christoph Feichtenhofer. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16284–16294, 2023.
- [85] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 34:13165–13176, 2021.
- [86] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yanis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pages 300–318. Springer, 2022. 3
- [87] 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, 2023. 3
- [88] Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. Dino-mc: Self-supervised contrastive learning for remote sensing imagery with multi-sized local crops. *arXiv preprint arXiv:2303.06670*, 2023. 3
- [89] Fengxiang Wang, Hongzhen Wang, Yulin Wang, Di Wang, Mingshuo Chen, Haiyan Zhao, Yangang Sun, Shuo Wang, Long Lan, Wenjing Yang, et al. Roma: Scaling up mamba-based foundation models for remote sensing. *arXiv preprint arXiv:2503.10392*, 2025. 3
- [90] Zhihao Li, Biao Hou, Siteng Ma, Zitong Wu, Xianpeng Guo, Bo Ren, and Licheng Jiao. Masked angle-aware autoencoder for remote sensing images. *arXiv preprint arXiv:2408.01946*, 2024. 3
- [91] Jeremy Irvin, Lucas Tao, Joanne Zhou, Yuntao Ma, Langston Nashold, Benjamin Liu, and Andrew Y Ng. Usat: A unified self-supervised encoder for multi-sensor satellite imagery. *arXiv preprint arXiv:2312.02199*, 2023. 3
- [92] 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 preprint arXiv:2312.10115*, 2023. 6
- [93] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. *arXiv preprint arXiv:2404.01260*, 2024.
- [94] Nikolaos Ioannis Bountos, Arthur Ouaknine, and David Rolnick. Fomo-bench: a multi-modal, multi-scale and multi-task forest monitoring benchmark for remote sensing foundation models. *arXiv preprint arXiv:2312.10114*, 2023.
- [95] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Pedram Ghamisi, Naoto Yokoya, Hao Li, Xiuping Jia, Antonio Plaza, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. DOI:10.1109/TPAMI.2024.3362475. 3
- [96] Nathan Longbotham, Fabio Pacifici, Seth Malitz, William Baugh, and Gustau Camps-Valls. Measuring the spatial and spectral performance of worldview-3. In *Hyperspectral Imaging and Sounding of the Environment*, pages HW3B–2. Optica Publishing Group, 2015. 4
- [97] Thierry Toutin and Philip Cheng. Quickbird—a milestone for high resolution mapping. *Earth Observation Magazine*, 11(4):14–18, 2002. 4
- [98] Manuel A Aguilar, María del Mar Saldaña, and Fernando J Aguilar. Assessing geometric accuracy of the orthorectification process from geoeye-1 and worldview-2 panchromatic images. *International Journal of Applied Earth Observation and Geoinformation*, 21:427–435, 2013. 4
- [99] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021. 5
- [100] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

- [101] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pages 2535–2544. PMLR, 2019. [5](#)
- [102] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. [6](#), [7](#), [8](#)
- [103] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. [6](#), [7](#), [8](#)
- [104] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. [6](#), [7](#)
- [105] 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 (CVPR)*, June 2016. [7](#)
- [106] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. [7](#)
- [107] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. [6](#)
- [108] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3520–3529, October 2021. [6](#)
- [109] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. [7](#)
- [110] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14507–14517, 2023. [7](#)