EvaMAE: How Helpful Are DEM Data in Enhancing Geo-Foundation Models for Earth Imagery?

Saugat Adhikari[*], Da Yan[*], Naman Nimbale[*], Weijin Liu[#], Xiaodong Yu[#], Akhlaque Ahmad[*],
Lyuheng Yuan[*], Jiao Han[*], Zhe Jiang[+]
[*]Indiana University Bloomington, Bloomington, IN, United States
[#]Stevens Institute of Technology, Hoboken, NJ, United States
[+]University of Florida, Gainesville, FL, United States
{adhiksa,yanda,nnimbale,akahmad,lyyuan,jiaohan}@iu.edu
{wliu62,xyu38}@stevens.edu
zhe.jiang@ufl.edu

Abstract

Numerous geo-foundation models have been pre-trained recently on plentiful unlabeled Earth imagery datasets by self-supervised learning, and they have been demonstrated to enhance performance in downstream supervised geospatial tasks such as flood extent mapping. However, these approaches generally ignore the terrain data that are readily available in the format of digital elevation model (DEM) from sources such as USGS's 3D Elevation Program (3DEP). On the other hand, a few works have shown that elevation guidance can improve the performance of flood extent mapping on conventional models trained from scratch. This is intuitive since in natural disaster events such as flooding, landslide and avalanche, the floodwater, loose earth or snow moves downhill.

In this work, we explore the use of DEM data in geo-foundation models by introducing EvaMAE, a Masked Autoencoder (MAE) architecture that integrates elevation data for pre-training. Different strategies for incorporating DEM data are studied including convolution- and cross-attention-based approaches. We also explore the use of ControlNet to integrate DEM data during fine-tuning. Extensive experiments on downstream tasks such as flood and landslide segmentations demonstrate that (i) incorporating DEM data is helpful in both the pre-training and the fine-tuning stages, that (ii) the best-performing model for pre-training uses cross-attention to combine DEM and RGB features in both the MAE encoder and decoder, and that (iii) the best-performing model for fine-tuning uses ControlNet to incorporate DEM data. We also release a new large-scale annotated flood mapping dataset called EvaFlood used in our model training. All our code, pre-trained models, and the EvaFlood dataset are available at https://github.com/saugatadhikari/EvaMAE.

CCS Concepts

• **Computing methodologies** → **Reconstruction**; **Image segmentation**; • **Information systems** → **Geographic information systems**; • **Applied computing** → **Earth and atmospheric sciences**.

SIGSPATIAL '25, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2086-4/2025/11
https://doi.org/10.1145/3748636.3762739

Keywords

GeoFoundation Model, Earth imagery, DEM

ACM Reference Format:
Saugat Adhikari[*], Da Yan[*], Naman Nimbale[*], Weijin Liu[#], Xiaodong Yu[#], Akhlaque Ahmad[*], Lyuheng Yuan[*], Jiao Han[*], Zhe Jiang[+]. 2025. EvaMAE: How Helpful Are DEM Data in Enhancing Geo-Foundation Models for Earth Imagery?. In *The 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3748636.3762739

1 Introduction

High-resolution Earth imagery is now widely accessible from agencies such as NOAA (National Oceanic and Atmospheric Administration), USGS (United States Geological Survey), and NASA (National Aeronautics and Space Administration), as well as commercial platforms like Google Earth and Maxar. These aerial and satellite images support critical applications ranging from environmental monitoring to disaster response. Optical satellites (e.g., Sentinel-2, Landsat 9, Maxar WorldView-3) use passive sensors to capture reflected sunlight across multispectral bands (e.g., visible, infrared), enabling high-resolution analysis but failing under cloud cover or at night. In contrast, SAR[1] satellites (e.g., Sentinel-1, ALOS-2) employ active radar to emit microwaves, penetrating clouds and vegetation for all-weather/day-night imaging — though with lower resolution and less intuitive grayscale backscatter outputs. NOAA National Geodetic Survey (NGS) also provides Emergency Response Imagery (ERI) [2] for rapid post-disaster damage assessment, which is collected using specialized aircraft to capture high-resolution photos of disaster-affected areas, with RGB cameras and precise geotagging.

Besides the above spectral data, the USGS 3D Elevation Program (3DEP) also collects high-precision terrain data in the format of digital elevation model (DEM) collected via airborne LiDAR. In other words, we can accompany each pixel in a geotagged Earth image with its elevation to reconstruct a 3D view for free. As we have demonstrated in our prior works FloodTrace [15] and ALFA [5] on flood annotation, this 3D view is more intuitive to humans, enabling more accurate and productive data annotations (e.g., flood maps over Earth imagery). As modern neural computer vision models mimic human vision capabilities, we expect the integration of DEM would improve their model performance on Earth imagery, especially for tasks such as detecting terrain-sensitive disaster events

[1] SAR means synthetic aperture radar

such as flooding, landslide and avalanche, where the floodwater, loose earth or snow move downhill under gravitational forces.

This has been confirmed by our prior work, EvaNet [37], which extends the U-Net model for flood segmentation by integrating DEM data to the design of both the network architecture and the loss function, and has achieved much superior performance than the U-Net baseline. DEM data have also been used by our prior graphical models for accurate flood segmentation by encoding the fact that floodwater moves downhill into directed conditional dependency edges between adjacent pixel pairs [3, 20–22, 26, 27, 35, 36, 44]. However, these models are trained from scratch, but Earth imagery with ground-truth flood map annotations is scarce, limiting the segmentation accuracy and model generalizability.

Foundation models have emerged as a potential solution, which are pre-trained on large unlabeled datasets through self-supervision, and then fine-tuned for various downstream tasks with small labeled datasets. Numerous geo-foundation models have been pre-trained recently on the abundant unlabeled Earth imagery datasets, and they have been demonstrated to enhance performance on downstream supervised geospatial tasks. For example, a popular annotated dataset for flood mapping is Sen1Floods11 [10] which was curated by a startup called 'Cloud to Street' (now Floodbase) but it is very small. The geo-foundation model, Prithvi [24], when fine-tuned on Sen1Floods11, achieves an IoU score of 82.99% on the water class, while the original water IoU score in [10] is only 24.21%. However, the existing geo-foundation models generally ignore the terrain (i.e., DEM) data, which are 'free' to collect and utilize.

In this paper, we explore how to effectively integrate this 'free' DEM data source into the pre-training and fine-tuning stages of a geo-foundation model. Following the existing geo-foundation models, we adopt the Masked Autoencoder (MAE) [19] architecture for pre-training where the input patches of an Earth image are partially masked for reconstruction, using a ViT[14]-based encoder and a ViT-based decoder. We consider only RGB for spectral channels, since the multi-spectral and multi-sensor extension techniques to be reviewed in Section 2 are orthogonal to our DEM integration methods, so can be easily incorporated. We call our elevation-integrated MAE models as EvaMAE, and three approaches of integrating DEM data are explored: (1) treating elevation map as an additional input channel beyond RGB (EvaMAE-Channel), (2) passing elevation map through a convolutional layer and then adding the elevation features with the RGB features patch by patch at both the encoder and the decoder (EvaMAE-Conv), and (3) using cross-attention to fuse the elevation features of all patches with the RGB features of input patches at both the encoder and the decoder (EvaMAE-CrossAttn).

Inspired by the design of ControlNet [45] in pre-trained text-to-image diffusion models, in our fine-tuning stage, we also explore a similar design to condition the downstream tasks on the elevation map which we find to improve the downstream model performance.

To verify the effectiveness of our EvaMAE models, we consider two semantic segmentation tasks: (1) flood extent mapping and (2) landslide segmentation. For **flood extent mapping**, while we can directly pre-train models with Earth imagery datasets used by existing geo-foundation models, these datasets do not cover much imagery from flooding events. Therefore, we have curated a dataset of high-resolution aerial imagery from NOAA's Emergency Response Imagery (ERI) from various hurricane and flooding events covering a total area of 17,055.83 km^2. Our EvaMAE models are first pre-trained on this unlabeled dataset before being fine-tuned.

We also annotated a subset of Earth imagery from NOAA ERI with flood maps for the purpose of fine-tuning in our downstream flood segmentation task, which covers an area of 3630.78 km^2. This accurate flood map annotation is made possible with our 3D annotation tool FloodTrace [15], which improves annotation productivity by using elevation-guided BFS for automated label derivation[2] [4]. However, the annotating process is still time-consuming, so we can only afford to annotate a subset of images from NOAA ERI.

We have released both the pre-training and annotated datasets accompanied with their elevation maps, collectively called EvaFlood, at https://github.com/saugatadhikari/EvaMAE.

For **landslide segmentation**, while we do not find good aerial imagery datasets on landslide, Landslide4Sense [1] is probably the largest annotated satellite imagery dataset on landslide and is adopted for fine-tuning. Each image has size 128 × 128 and 14 bands, and we only take the RGB and DEM bands for fine-tuning. The landslide detection rate is expected to be low[3] in this difficult task since our models are pre-trained on our EvaFlood dataset and have not seen landslide events before, and images in Landslide4Sense have a much lower resolution and keeping only 4 out of the 14 bands loses information. Nevertheless, we show that integrating DEM increases the landslide detection rate a lot compared with using RGB only, indicating that our pre-trained models can generalize to unseen scenarios where DEM data are again helpful.

The main contributions of this paper are summarized as follows:

- This is the first work that comprehensively investigate how to effectively integrate the 'free' DEM data source into the training of a geo-foundation model. Positive results are observed which advocate the utilization of DEM data in the training of future geo-foundation models.
- Three approaches are proposed to integrate DEM data into the pre-training stage of a geo-foundation model, using convolution- and cross-attention-based designs. The cross-attention-based design, EvaMAE-CrossAttn, is found to be the most effective and clearly improves performance when compared with a baseline without using DEM.
- We explore the use of ControlNet to condition the fine-tuning stage on the DEM data, and obtain positive results which advocate the use of DEM data in geo-foundation model fine-tuning, especially for terrain-sensitive disaster events.
- We curated a high-quality dataset, EvaFlood, for DEM-enhanced training of geo-foundation model under flooding scenes, including components for both pre-training and fine-tuning.
- Extensive experiments have been conducted which verify that the 'right' way of integrating DEM which we have discovered (i.e., EvaMAE-CrossAttn for pre-training, and ControlNet for fine-tuning) can both obtain better recovery of

[2] When an annotator marks an individual pixel $\mathbf{p}$ as flooded, the label propagates to nearby pixels with lower elevations by 'pit-filling' BFS stopping when reaching pixels with elevation higher than that of $\mathbf{p}$. When an annotator marks an individual pixel as dry, the label propagates to nearby pixels by 'hill-climbing' BFS stopping when reaching pixels with elevation starting to drop.

[3] We expect the performance to be much better if annotated aerial imagery datasets on landslide are available for fine-tuning, or if we train the geo-foundation models to take all bands of satellite imagery.

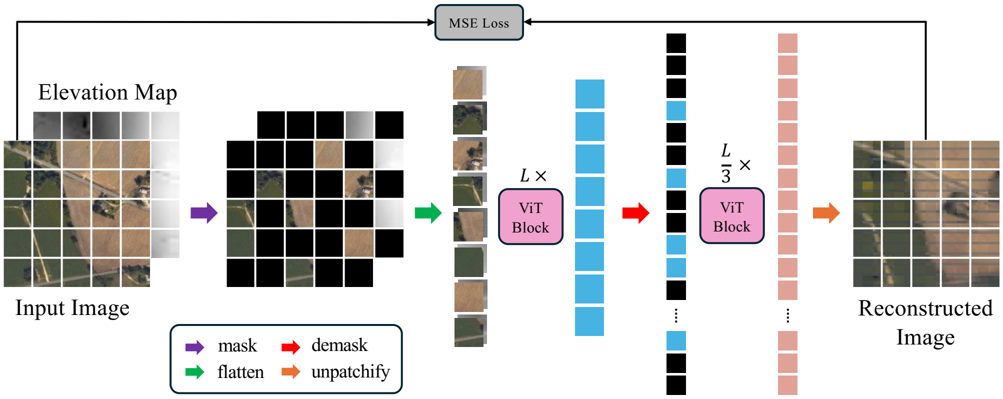

Figure 1: Illustration of MAE (Ignoring the Elevation Map) and EvaMAE-Channel

umasked patches, and achieve better results on downstream flood and landslide segmentation tasks.

The rest of this paper is organized as follows. Section 2 first reviews the preliminaries and related work on geo-foundation models. Section 3 then presents our model design, and Section 4 reports our experiments. Finally, we conclude this paper in Section 5.

2 Preliminaries and Related Work

In this section, we first briefly review ViT [14] and MAE [19] that are basic components of modern geo-foundation models. We then review the many recently proposed geo-foundation models.

2.1 ViT and MAE

The Vision Transformer (ViT) [14] adapts the Transformer model, originally developed for NLP, to computer vision tasks. ViT splits an input image into fixed-size (typically 16×16) non-overlapping patches, linearly embeds them into patch tokens, and processes them with a standard Transformer encoder. By leveraging self-attention mechanisms, ViT captures global dependencies across the entire image, overcoming the locality limitations of ConvNets. Swin Transformer [30] further enhances the efficiency by introducing locality and hierarchy to ViTs: self-attention is conducted only within each local window, and windows are shifted after each layer to enable cross-window communication; also, patch merging is used to reduce resolution, enabling multi-scale feature extraction.

The Masked Autoencoder (MAE) [19] is a self-supervised learning framework designed to efficiently pre-train ViTs by reconstructing masked portions of input images. Figure 1 illustrates the MAE architecture (let us ignore the elevation map for now), where an input image is first partitioned into patches of size 16 × 16. MAE randomly masks a high proportion (e.g., 75%) of the image patches, and flattens each unmasked patch into a 1024-dimensional embedding by a learnable linear projection. These embeddings for visible patches are then inputted into the ViT encoder to emit their 1024-dimensional feature vectors (i.e., latent representations).

While the ViT encoder takes only 25% unmasked patch tokens as the input, the ViT decoder takes all image patches as its input, so it has to be lightweight. The input embeddings to the decoder are 512-dimensional, where visible patches pass their feature vectors emitted by the encoder through a learnable linear projection to obtain 512-dimensional embeddings, while masked patches use a learnable embedding for the special [MASK] token. The patch embeddings emitted by the decoder are then converted back to 16 × 16 patches using a learnable linear projection.

Patch embeddings at the inputs of both the encoder and the decoder are added with 2D positional embeddings (particularly important for the encoder as the unmasked patches are sparse), where the positional embedding of a patch is obtained by concatenating the conventional 1D sinusoidal positional embeddings individually computed for height and width. MAE is trained to recover the masked patches by minimizing the mean squared error (MSE).

2.2 Geo-Foundation Models

Numerous geo-foundation models have been proposed in recent year, and we review them next by discussing their techniques on adding contrastive loss, multi-spectral, multi-scale and multi-sensor support, and methods for better patch masking, orientation-awareness, and efficiency enhancement. We remark that those techniques are orthogonal to the topic of DEM integration that we focus on, and can be easily integrated as needed.

There are two strategies for self-supervision during the pre-training of geo-foundation models: contrastive learning (CL) and masked image modeling (MIM) [43].

Self-Supervision by CL. Contrastive learning (CL) pre-trains an image encoder such as MoCo-v2 [11] to emit embeddings that pull augmentations of the same image closer and to push apart representations of two different images, using the well-known InfoNCE [40] loss. SeCo [32] also encourages the representation to be invariant to seasonal changes by using images at the same location but separated by approximately 3 months as positive samples, while CaCo [31] further enforces sensitivity to permanent, long-term changes (e.g., urban development) by pushing apart images at the same location if a long-term change is estimated between them (GMM clustering over feature differences is used to find two clusters 'change' and 'no change'). Both SeCo and CaCo sample images around cities to reduce redundancy (i.e., oceans and forests with low variability). Ayush et al. [8] further improves the performance of self-supervised learning by utilizing the geo-coordinates of geo-tagged remote sensing datasets. The coordinates are clustered into k areas, and an area prediction head is added to make the representation geography-aware. MATTER [6] uses contrastive learning to learn material and texture representations that stay consistent over time, using specialized network designs. The learned representations are effective in downstream tasks such as change detection and semantic segmentation. Note that the reconstruction loss of MAE as in our EvaMAE models can be easily combined with a contrastive loss as in [7, 16, 39], and is orthogonal to our focus on how DEM data can integrate with the MAE network architecture.

Self-Supervision by MIM. Most recent geo-foundation models, instead, follow the MAE architecture that uses masked image modeling (MIM) for self-supervision to reconstruct the masked patches. Prithvi [24] captures the temporal nature of satellite imagery (i.e., a region is scanned multiple times by a satellite over time, generating a tensor with a time dimension) by designing a 3D positional encoding for its patches which is computed by concatenating the 1D sinusoidal positional embeddings individually computed for height, width, and time. The MAE encoder input contains all the patches along the time dimension. Since we focus on non-recurrent disaster events in downstream tasks, we ignore the temporal dimension which is orthogonal to our focus on DEM data integration.

Multi-Spectral Support. Since spectral bands have different wavelengths and spatial resolution, SatMAE [12] proposes to further group the spectral bands into different channel groups to generate finer tokens, and to expand the positional encoding to also include the spectral encoding. To better exploit local spectral continuity and generalize to variable band counts, S2MAE [28] and SpectralGPT [23] partitions a 3D cube-shaped spectral image into non-overlapping 3D tensor tokens along both the spatial and spectral dimensions. They also support more flexible masking schemes than SatMAE's group mask design, so that different channels in a group can be masked differently. In this work, we only consider aerial imagery with RGB channels as a clean setting to study the methods and effect of DEM data integration. In fact, our aerial imagery has such high resolution that flooded areas are easily identifiable even to the naked eye. Moreover, [9] finds that using additional bands may even reduce performance on segmentation tasks.

Multi-Scale Support. Ground Sample Distance (GSD) measures the spatial resolution, denoting the physical distance between two adjacent pixels. Realizing that objects of interest can vary across wide spatial resolutions, Scale-MAE [34] introduces a GSD-aware positional embedding to also encode the scale information, and it decodes the masked image through a bandpass filter to reconstruct both low and residual high frequency images. Noman et al. [33] argue that Scale-MAE can only work with RGB channels, but multi-spectral images (e.g., from Sentinel-2) can have different GSD resolutions for different channels. SatMAE++ [33] takes input image at different scale levels, and feeds the image at the lowest scale level to SatMAE. The reconstructed output from SatMAE is then utilized by upsampling blocks to reconstruct higher-scale levels. Cross-Scale MAE [39] further enhance MIM with contrastive learning to enforce cross-scale consistency. For each image, it generates an additional image of lower GSD for the same site (e.g., by cropping and rescaling), and pass both images through a siamese MAE network. Contrastive loss is applied to the encoder output to pull the two images closer, and a prediction loss is added to the decoder output to let the embeddings from the lower GSD predict the embeddings of the original image. In our work, images in our EvaFlood dataset has the same high resolution so multi-scale support is not needed for flood segmentation, but the landslide images have a much lower resolution so the multi-scale techniques reviewed here may be used to further improve the performance of landslide segmentation, which we leave as a future work.

Multi-Sensor Support. msGFM [18] incorporates four sensor modalities in pre-training: RGB images, Sentinel-2, SAR and DSM. Each sensor has its own patch embedding layer adapted to its number of channels, but the learned embeddings of all modalities are integrated through the same encoder to learn joint representations. Each sensor has its own decoder to predict its masked patches from the encoded representations of itself or of another paired modality for the same geo-location. CROMA [16] and OmniSat [7] further applies cross-attention between different sensor modalities to learn joint representations, and they combine reconstruction loss with contrastive loss to pull representations of the same image patch of different sensor modalities closer while pushing apart different image patches. SkySense [17] is purely based on contrastive learning without MIM. It uses a factorized encoder to extract spatial features from each sensor modality and then fuse them to capture a multi-modal spatiotemporal representation. Contrastive learning is then used to pull together features of multi-grained samples (e.g., pixel- and image-level) at the same geo-location.

Note that msGFM is the only geo-foundation model considering elevation map (i.e., DSM), it simply treats DSM the same as other modalities (e.g., RGB) for joint encoding by self-attentions and for cross-modal prediction. We argue that the elevation map should be treated differently since it is available for free as a condition to improve image reconstruction and downstream predictions, so there is no need waste model capacity to reconstruct elevation maps. Moreover, using more than one transformer layer (with self-attention) to encode the elevation map is detrimental to the model performance (see Section 3), while it is favorable for spectral modalities.

Improvements to the MAE Framework. While objects in natural images are generally oriented upward due to gravity, those in Earth imagery can appear in various orientations from a bird's-eye view. To make the representations rotation-invariant, MA3E [29] crops an area of patches and randomly rotates the area content. Both rotated patches and other patches are separately masked with a ratio of 75%, and the unmasked patches are passed to an MAE to reconstruct the original image. Angle embeddings are added to the patches of the rotated crop to prompt the model, and optimal transport is used to assign similar original image patches to each patch in the rotated crop. RVSA [42] adapts the plain ViTs with isotropic structures using a learnable rotation mechanism, by proposing a rotated varied-size window attention to replace the original full attention in transformers, which also reduces the computational cost. Since many patch tokens in Earth Imagery are repetitive (e.g., oceans and forests), to reduce the cost of self-attention (quadratic to the number of tokens), LeMeViT [25] uses a small number of learnable meta tokens to compress the image information. Self-attention among image tokens are replaced by dual cross-attention between image tokens and meta tokens, which significantly reduces the computational cost. Finally, RingMo [38] proposes to better capture small and dense objects by a new patch incomplete mask (PIMask) strategy that randomly reserve some pixels in masked patches, but all patches need to be inputted to the encoder. Since

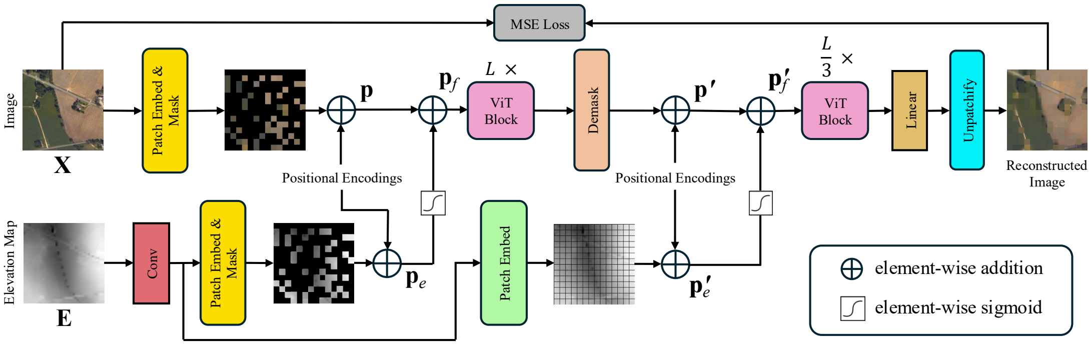

Figure 2: The Network Architecture of EvaMAE-Conv

our EvaMAE models are also MAE-based, these improvements can be easily integrated to our models.

3 Methodology

We next describe our explored architectures for elevation-integrated MAE pre-training, including: (1) DEM as an additional input channel (EvaMAE-Channel), (2) convolution-based DEM integration (EvaMAE-Conv), and (3) cross-attention-based DEM integration (EvaMAE-CrossAttn). Besides, we also explore their fine-tuning versions that use ControlNet [45] to condition on DEM, denoted by ControlNet-Channel, -Conv, and -CrossAttn, respectively.

3.1 EvaMAE-Channel

Refer to Figure 1 on Page 3 again, but now also consider the elevation map. In EvaMAE-Channel, we simply treat the elevation map as an additional input channel. Recall that in the 'flatten' operation, a linear projection $f_e(.)$ is applied to convert each patch into an embedding, to be inputted as a token to the ViT encoder. When only RGB patches are considered, each patch has size $16 \times 16 \times 3$ which is flattened into a 768-dimensional vector and then passed through a 768×1024 dense layer[4] $f_e(.)$ to obtain its 1024-dimensional token embedding. Now since EvaMAE-Channel uses the additional DEM channel, each patch has size $16 \times 16 \times 4$, so it is flattened into a 1024-dimensional vector. Therefore, the dense layer $f_e(.)$ for patch-to-token embedding mapping is now 1024×1024.

Note that we follow the typical MAE architecture as used by SatMAE [12], where in Figure 1, the input image has size 224×224, leading to $14 \times 14 = 196$ patches. The ViT encoder has $L = 24$ transformer layers and an embedding dimension of 1024. Since the ViT decoder takes more tokens as input, to keep computational cost tractable, it has 8 transformer layers and an embedding dimension of 512. All transformer layers use multi-headed self-attention with 16 parallel attention heads.

3.2 EvaMAE-Conv

Figure 2 shows the network architecture of EvaMAE-Conv, where we treat the elevation map **E** as a separate branch to be patchified, with its token embeddings then added to those of the input image **X**. To capture the local elevation cues (e.g., downhill directions), we first pass **E** through a 3×3 convolution (with replicate padding and padding size = 1) before patchification.

Let the token embedding of an image patch (after adding its positional embedding) be $\mathbf{p}$, and let the token embedding of its corresponding DEM patch be $\mathbf{p}_e$, then we obtain the fused patch embedding $\mathbf{p}_f$ to be inputted to the MAE encoder by computing

$$\mathbf{p}_f = \mathbf{p} + \sigma(\mathbf{p}_e),$$

where σ is the sigmoid function that adds non-linearity and regulates the added values from the DEM branch to be within $(0, 1)$. We also tested a version without taking σ (i.e., $\mathbf{p}_f = \mathbf{p} + \mathbf{p}_e$), but it leads to a slightly lower segmentation performance. We denote this variant as EvaMAE-Conv^{-}. Inspired by EvaNet [37], we also explored the use of GLU (gated linear unit) [13] for feature fusion by computing $\mathbf{p}_f = \mathbf{p} \otimes \sigma(\mathbf{p}_e)$ where $\otimes$ is the element-wise multiplication, but the performance drops significantly so we excluded this model. We believe this is because MAE does not work well with GLU-fused embeddings as the input tokens, in contrast to convolutional encoder-decoder networks like EvaNet.

Note that the DEM patch embeddings are fused to the inputs of both the ViT encoder and the decoder, with the difference that the encoder only takes the unmasked patches while the decoder takes all patches (masked image patches are replaced with the [MASK] token, but fused with the embeddings of their paired DEM patches).

3.3 EvaMAE-CrossAttn

Figure 3 shows the network architecture of EvaMAE-CrossAttn, where the Earth image **X** and the elevation map **E** are passed through separate branches for patchifying, embedding, and then ViT encoding. Unlike the image branch, the elevation branch copes with all DEM patches without masking, and only one ViT[5] layer is used in the encoder (instead of L layers for the image branch).

Note from Figure 3 that we use only one ViT (i.e., Transformer) layer for DEM since our experiments reveal that using more than one layer actually reduces the performance, most likely because the information in DEM (e.g., downhill directions) is much simpler than that in images, so using more Transformer layers is an overkill and backfires based on the principle of Occam's razor.

After the ViT encoding, we then use standard cross-attention operation to fuse the information of DEM patches into the image patches. As the red module in Figure 3 shows, the cross-attention

[4] Implementation-wise, $f_e(.)$ is realized as a 16×16 convolutional layer with stride = 16 operating on the entire input image. The convolutional kernel parameters exactly match weights of the dense layer when input patches are viewed as flattened vectors.

[5] We use the terms 'ViT layer' and 'transformer layer' interchangeably, since we are using transformer layers with patch tokens

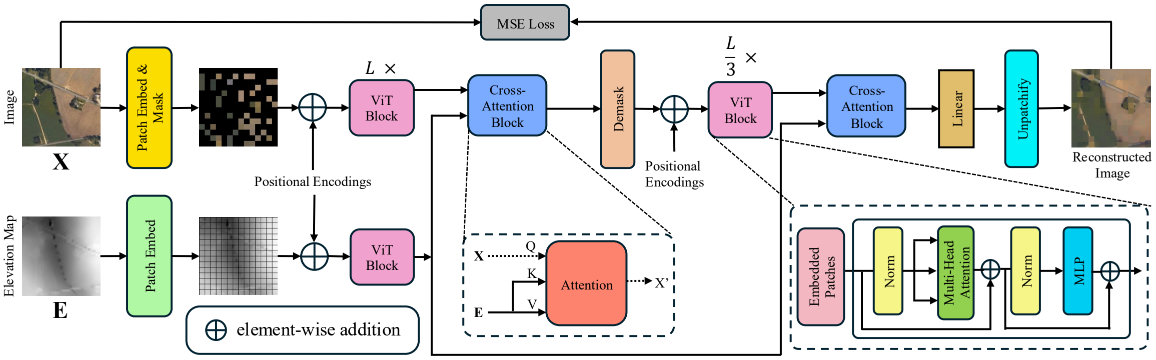

Figure 3: The Network Architecture of EvaMAE-CrossAttn (Using Cross-Attention Blocks)

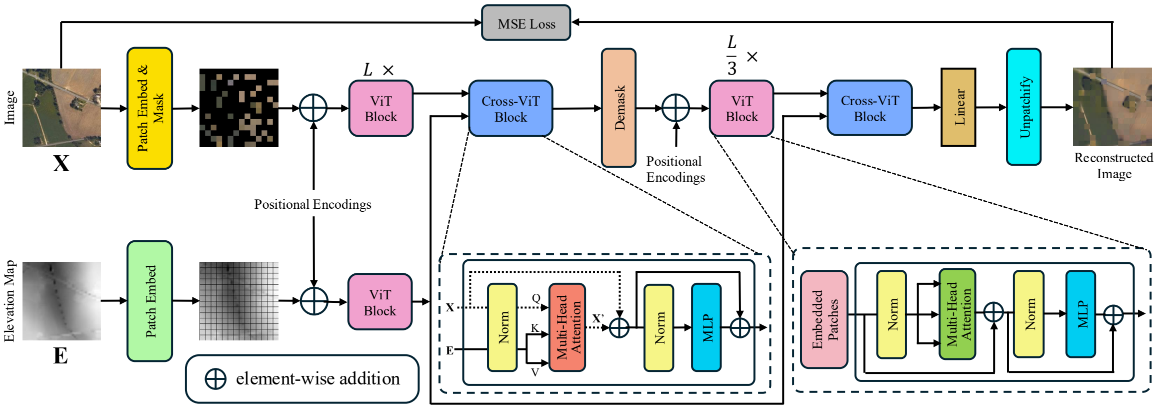

Figure 4: The Network Architecture of EvaMAE-CrossViT (Using Cross-ViT Blocks)

emits new image patch embeddings $\mathbf{X}'$ by computing the embedding of each image patch $\mathbf{p}$ as the weighted sum of the 'value' embeddings of DEM patches, where the weight of each DEM patch $\mathbf{p}_e$ is computed as the inner product between the 'query' embedding of $\mathbf{p}$ and the 'key' embedding of $\mathbf{p}_e$. Here, 'query' embeddings are computed by a dense layer over image path embeddings, and 'key' and 'value' embeddings are computed by two separate dense layers over DEM patch embeddings, as in standard Transformer [41].

In the actual implementation, we use multi-headed cross-attention with 8 attention heads for the cross-attention blocks.

The image patch embeddings after cross-attention are then demasked (i.e., by using [MASK] for masked patches) and passed through the ViT decoder, and the decoded image patch embeddings are then fused with the encoded DEM patches again via cross-attention before being used to reconstruct the image patches.

Note that we use full DEM data without masking when taking cross-attention with image patches for feature fusion at the outputs of both ViT encoder and decoder, since DEM data are available as a context to condition upon, rather than the target of reconstruction/decoding as with $\mathbf{X}$. The cost introduced by the DEM source is light though, since it is only passed through one transformer layer for encoding, and two cross-attention layers for feature fusion.

We also explore an alternative approach to taking cross-attention for feature fusion, by using full Transformer layer rather than only the cross-attention operation, as illustrated by the two blue 'Cross-ViT' blocks in Figure 4. This added model capacity is found to improve the performance of our flood segmentation task. We call this variant of EvaMAE-CrossAttn as EvaMAE-CrossViT.

3.4 Fine-Tuning Methods

Once our EvaMAE models are pre-trained, we only use the encoder during the fine-tuning stage for downstream segmentation tasks, which takes an Earth image $\mathbf{X}$ and its corresponding elevation map $\mathbf{E}$ as its inputs, and outputs 196 patch embeddings (since we use images of size 224×224, which leads to $14 \times 14 = 196$ patches of size 16×16) each with 1024 dimensions, denoted by $\mathbf{X}' \in \mathbb{R}^{196 \times 1024}$.

As the right part of Figure 5 shows, we use a convolution-based decoder to decode $\mathbf{X}'$ back to a tensor $\mathbf{Y} \in \mathbb{R}^{224 \times 224 \times N}$, where N is the number of classes and $\mathbf{Y}[i][j] \in \mathbb{R}^N$ is the logit score vector for pixel $\mathbf{p} = (i, j)$. In our flood (resp. landslide) segmentation task, $N = 2$ since a pixel is either in or not in the flooding (resp. landslide) area. As Figure 5 shows, $\mathbf{X}'$ is first reshaped to a tensor of size $14 \times 14 \times 1024$, and then upsampled by four 2×2 transposed convolutions (with a stride of 2), each doubling the width, height and reducing the number of channels by half. Finally, a 1×1 convolution is applied to reduce the number of channels from 64 to the desired number of classes N, and pixel-wise cross-entropy loss is applied to the resulting tensor to fine-tune our models with the supervision of the ground-truth flood/landslide annotations.

In the above basic fine-tuning method, the parameters of both the convolutional decoder and the EvaMAE encoder are updated during fine-tuning. An alternative way is to use ControlNet [45] originally proposed for text-to-image diffusion models. In this approach, we

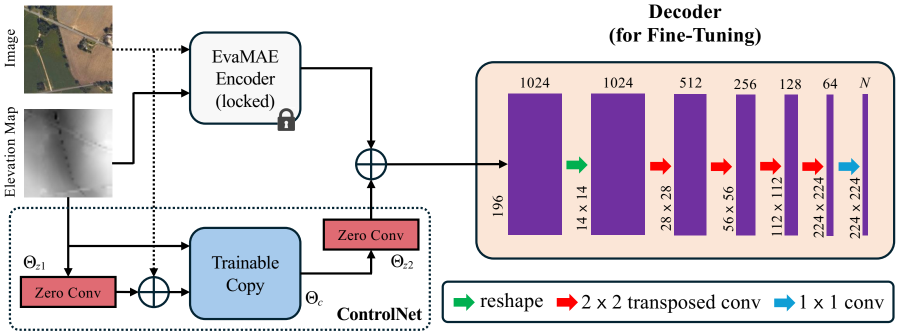

Figure 5: The Decoder for Fine-Tuning & Additional ControlNet Module

incorporate a new ControlNet module as shown by the dotted rectangle at the lower-left corner of Figure 5.

Specifically, we lock (freeze) the parameter of the original EvaMAE encoder and simultaneously clone it to a trainable copy with parameters Θ_c. The trainable copy takes the elevation map as an external condition that is convolved and then added to the image input. Here, the locked parameters preserve the well pre-trained model with plentiful Earth images, while the trainable copy reuses this pre-trained model to establish a deep, robust, and strong backbone for handling diverse input conditions.

The trainable copy is connected to the locked model with zero convolution layers, where a zero convolution layer is a 1×1 convolution with both weight and bias initialized to zeros. To build up a ControlNet, we use two instances of zero convolutions with parameters Θ_{z1} and Θ_{z2}, respectively. The output of the ControlNet is then added to that of the locked EvaMAE encoder.

In the first training step, since Θ_{z2} is initialized to zero, it adds nothing to the output of EvaMAE encoder, so harmful noise cannot influence the hidden states of the neural network layers in the trainable copy when the training starts. Moreover, since Θ_{z1} is initialized to zero, the trainable copy also receives only the input image; it is thus fully functional and retains the capabilities of the pre-trained EvaMAE encoder allowing it to serve as a strong backbone for further learning.

In this ControlNet-based method, fine-tuning updates the parameters of the convolutional decoder as well as Θ_c, Θ_{z1} and Θ_{z2}.

4 Experiments

We now report our comprehensive experimental study. We will first describe the datasets and experimental setup that we use, followed by the experimental results on pre-training and fine-tuning.

4.1 Datasets and Experimental Setup

Datasets. We obtain high-resolution aerial imagery from NOAA ERI during different storm events for both pre-training and flood segmentation fine-tuning [2]. For all the aerial images collected, we also obtain the corresponding DEM data using Google Earth Engine API which provides public access to 10-meter resolution elevation maps through the USGS 3D Elevation Program (3DEP).

Originally, all the aerial images are 0.3 m × 0.3 m in resolution and the DEM data are 10 m × 10 m. We resampled both of them into a resolution of 2 m × 2 m for alignment, which is fine enough for the purpose of flood/landslide mapping.

Table 1: Data Statistics of Pre-Training Datasets

Event	#{Images}	Coverage (km^2)		
		Training	Validation	Total
Midwest U.S. Flooding 2015	23,622	3793.90	947.12	4741.03
Hurricane Matthew 2016	13,581	2177.04	548.72	2725.76
Hurricane Harvey 2017	16,386	2625.61	663.13	3288.74
Hurricane Michael 2018	31,391	5048.11	1252.19	6300.30
Total	84,980	13,644.66	3411.17	17,055.83

Table 2: Fine-Tuning Data Statistics for Flood Segmentation

Event	Coverage (km^2)			Total Annotations (%)		
	Training	Test	Total	%Flood	%Dry	%Annotated
Hurricane Matthew 2016	245.02	241.89	486.91	35.38	40.83	76.18
Louisiana Flooding 2016	126.65	47.76	174.41	34.19	47.69	81.88
Hurricane Harvey 2017	2695.44	373.43	3068.87	43.45	28.37	71.82
Total	3067.11	663.08	3730.19	41.97	30.89	72.86

Table 1 shows the flooding-related ERI events used for pre-training and their corresponding area of imagery coverage. Each image has size 224 × 224 and there are 84,980 images in total. For each event, we selected 80% of the data for training and 20% for validation (i.e., convergence check). In terms of spatial coverage, the pre-training data covers a total of 17,055.83 square kilometers (km^2) among which 13,644.66 km^2 (80%) are used for training.

Table 2 shows the statistics of our datasets on flooding-related ERI events used for fine-tuning in our flood segmentation task. For each ERI event, we have a few large images with different sizes that are collected from non-overlapping locations: Hurricane Matthew has 9 images, Louisiana Flooding has 11 images, and Hurricane Harvey has 101 images, with a total of 121 images. We randomly selected 90% of these images for training and 10% for test, so we have 109 images in the training set and 12 images in the test set. Specifically, Hurricane Matthew has 4 images in the test set, Louisiana Flooding has 1, and Hurricane Harvey has 7. Note that imagery from Louisiana Flooding is not used during pre-training, while the two largest datasets used for pre-training are not used in fine-tuning, which assists in evaluating the model generalizability.

In terms of spatial coverage, the fine-tuning data covers a total of 3730.19 km^2 among which 3067.11 km^2 (82.14%) are used for training and 663.08 km^2 (17.86%) are for test. The ground-truth flood maps of these images are annotated using the 3D visualization tool, FloodTrace [15], which is a time-consuming process taking about 4 months. Some pixels are not annotated as either 'flooded' or 'dry' since they are ambiguous covered by tree canopy and their

Table 3: Data Statistics for Landslide Segmentation

Data	Image Size	#{Images}	Coverage (km^2)	%Landslide	%Background
Raw	128 x 128	4599	7535.00	2.25	97.75
Processed	224 x 224	2767	4533.45	3.74	96.26

labels cannot be automatically inferred by elevation-guided BFS. These pixels are excluded in loss computation and prediction IoU evaluation. We have released our dataset, EvaFlood, along with our code at https://github.com/saugatadhikari/EvaMAE.

For landslide segmentation, we use the multi-sensor satellite imagery dataset, Landslide4Sense [1], which covers landslide-affected areas around the world from 2015 through 2021. Each image is a composite of 14 bands for which we only use the RGB and DEM channels, and the spatial resolutions of both RGB and DEM are 10 m × 10 m. The dataset is fully annotated where each pixel is either 'landslide' or 'background'.

As shown in Table 3, in the original data (denoted by 'Raw'), background (non-landslide) pixels dominate landslide pixels and only 2.25% pixels are landslide ones. We preprocess the data to drop images that only contain background pixels, which boost the fraction of landslide pixels to 3.74% as shown in Table 3. Since our models are pre-trained on input images of size 224 × 224, we upsample each image and their label map from 128×128 to 224×224. Out of the 2767 images, 2231 (80%) are used for training and 536 (20%) are used for test. In terms of spatial coverage, the preprocessed dataset covers a total of 4,533.45 km^2 among which 3,655.27 km^2 are used for training and 878.18 km^2 are used for test.

Experimental Setup. We conduct all model pre-training experiments on a distributed cluster of 10 nodes, each with 4 NVIDIA A100 GPUs, so a total of 40 GPUs. Model fine-tuning is conducted on a single node with all its 4 A100 GPUs. The machines are from the Polaris supercomputer at the Argonne National Laboratory.

Following [12], RGB images are normalized using the mean and standard deviation over the entire dataset along each channels using Z-score normalization; and following [37], DEM data are normalized over the entire dataset using min-max normalization.

We have described the model architecture and most hyperparameters of EvaMAE models in Section 3, and we now provide additional information. In Figure 5, the first zero convolution layer is a 1 × 1 convolution with 1 input channel (for DEM) and 3 output channels (to allow addition with RGB). The second zero convolution layer has 1024 channels for both the input and the output (to allow addition with the EvaMAE encoder's output embeddings).

For **pre-training** our EvaMAE models, we use a learning rate of 1.5×10^{-5}, a batch size of 8, and a masking ratio of 75%. The pre-trained parameters of SatMAE (the model version for RGB imagery) are used for initialization which were pre-trained for 800 epochs, and we continue pre-training for 50 more epochs on our pre-training datasets that have been summarized in Table 1. Following SatMAE [12], our pre-training uses the AdamW optimizer, a cosine decay learning rate scheduler, and standard augmentations (RandomResizedCrop and RandomHorizontalFlip). For **fine-tuning** in both our flood and landslide segmentation tasks, we use a learning rate of 10^{-5} and a batch size of 32. We train for 30 epochs, and 20% images are randomly held out for validation.

The above numbers of epochs are selected to ensure that the training in both pre-training and fine-tuning stages converge.

Table 4: Pre-Training Time on 40 A100 GPUs

Model	Time (min)	Model	Time (min)
SatMAE	20.90	EvaMAE-Conv^{-}	26.45
SatMAE++	25.53	EvaMAE-Conv	27.31
Prithvi	8.70	EvaMAE-CrossAttn	40.25
EvaMAE-Channel	26.16	EvaMAE-CrossViT	42.57

Besides our EvaMAE model variants, we also incorporate three representative geo-foundation models for comparison. **(1) SatMAE:** There are multiple versions: multi-temporal, multi-spectral and RGB only. We use the 'RGB only' version and continue training for 50 more epochs on our pre-training datasets. The pre-trained model is obtained from https://github.com/sustainlab-group/SatMAE. **(2) SatMAE++:** There are multiple versions: multi-spectral and RGB only. We use the 'RGB only' version and continue training for 50 more epochs on our pre-training datasets. The pre-trained model is obtained from https://github.com/techmn/satmae_pp. **(3) Prithvi:** The model was originally pre-trained using 6 channels: RGB channels and Sentinel-2 bands 8A, 11, and 12. Since our EvaFlood dataset only has RGB channels, we take out the pre-trained weights for RGB channels only for use. Prithvi was initially pre-trained for 1000 epochs and we continue training for 50 more epochs on our pre-training datasets. The pre-trained model is obtained from https://github.com/isaaccorley/prithvi-pytorch.

4.2 Experimental Results on Pre-Training

Table 4 reports the total time of pre-training on our EvaFlood datasets for the various geo-foundation models we compare with. We can see that except for Prithvi which takes only 8.7 minutes, the other models take around 20–40 minutes. This is thanks to the distributed training with 40 A100 GPUs on Polaris. Compared with SatMAE, the additional overhead caused by DEM integration is light, with cross-attention-based methods being more expensive.

Figure 6 shows a visual comparison of the patch reconstruction performance of the state-of-the-art baseline model SatMAE and our two EvaMAE variants EvaMAE-Channel and EvaMAE-CrossViT. Note that as we shall see soon, EvaMAE-CrossViT is the best-performing model in our downstream flood segmentation task. In Figure 6, the first row shows 7 example images from our EvaFlood dataset, the second row shows their masked versions, and the remaining rows shows the reconstructed images for the three geo-foundation models. We can see that the reconstruction quality of EvaMAE-CrossViT is the best and significantly better than the other two models that contain some artifacts on patch borders. EvaMAE-Channel is also much better than SatMAE in terms of the similarity of pixel colors to the original images. This shows that the integration of DEM data improves the performance of patch reconstruction during the pre-training stage, and that cross-attention-based feature fusion is particularly effective.

4.3 Experimental Results on Fine-Tuning

Besides the geo-foundation models, we also train two convolutional baseline models from scratch, U-Net (with DEM as an additional channel) and EvaNet, following the default configuration in [37]. These models are trained for 30 epochs on our fine-tuning datasets that have flood/landslide annotations.

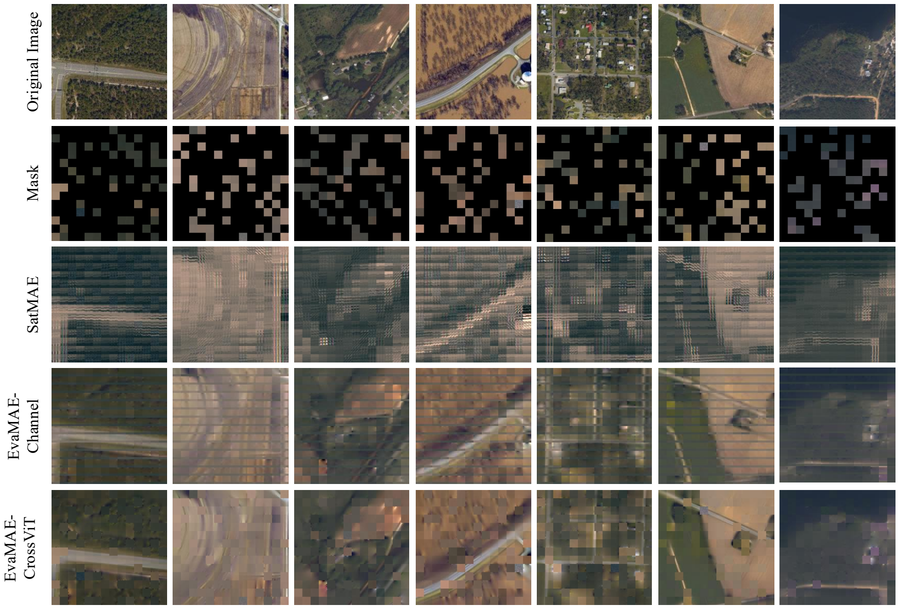

Figure 6: Visual Comparison of Patch Reconstruction Performance

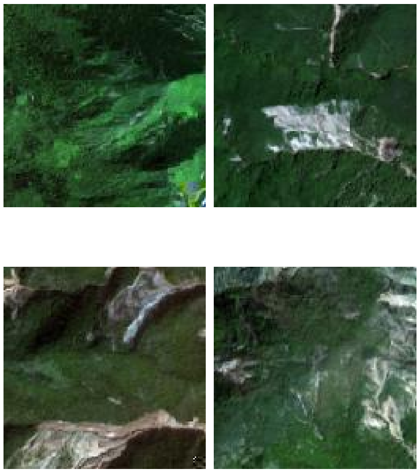

Figure 8: Landslide4Sense Images

Table 5: Flood Segmentation Result Comparison (Unit: %)

Model	Flood IoU	Dry IoU	mIoU	Time (min)
U-Net	75.92 ± 0.27	63.38 ± 0.27	68.45 ± 0.22	22.42 ± 0.43
EvaNet	69.29 ± 1.95	70.40 ± 0.89	69.54 ± 1.41	24.29 ± 0.21
SatMAE	86.24 ± 1.68	80.69 ± 1.18	83.22 ± 1.36	32.04 ± 0.56
SatMAE++	78.34 ± 0.49	68.69 ± 0.31	72.89 ± 0.34	30.86 ± 0.37
Prithvi	61.43 ± 4.11	54.28 ± 1.90	57.39 ± 1.88	**21.43 ± 0.61**
EvaMAE-Channel	82.58 ± 2.24	74.84 ± 1.66	78.77 ± 1.89	33.25 ± 0.92
EvaMAE-Conv^{-}	85.95 ± 0.22	81.07 ± 0.84	83.16 ± 0.46	26.26 ± 0.64
EvaMAE-Conv	86.31 ± 1.61	81.18 ± 1.48	83.45 ± 1.55	27.86 ± 0.38
EvaMAE-CrossAttn	85.54 ± 0.43	80.66 ± 0.57	82.90 ± 0.55	33.53 ± 0.62
EvaMAE-CrossViT	89.61 ± 1.13	85.52 ± 0.72	87.42 ± 0.84	35.01 ± 0.97
ControlNet-Channel	82.26 ± 0.56	73.44 ± 1.00	77.83 ± 0.76	34.28 ± 0.26
ControlNet-Conv^{-}	87.47 ± 1.03	83.25 ± 0.99	85.11 ± 0.95	29.34 ± 0.25
ControlNet-Conv	86.84 ± 0.79	81.89 ± 0.57	84.08 ± 0.49	28.90 ± 0.38
ControlNet-CrossAttn	80.44 ± 0.43	70.97 ± 0.24	75.57 ± 0.32	30.24 ± 0.31
ControlNet-CrossViT	**89.97 ± 0.21**	**86.15 ± 0.29**	**87.90 ± 0.19**	31.38 ± 0.12

Note that (1) we also fine-tune the geo-foundational models for 30 epochs after pre-training, and that (2) our U-Net and EvaNet models utilize DEM data though cannot benefit from pre-training, while baseline geo-foundation models SatMAE, SatMAE++ and Prithvi do not utilize DEM data but benefits from pre-training.

Table 5 shows the performance results (mean ± std calculated from 5 runs) of various models for flood segmentation, where we report metrics including Flood IoU, Dry IoU, mean IoU (mIoU), and the model training time (for 30 epochs). Here, Flood (resp. Dry) IoU measures the overlap between the predicted flood (resp. dry area) extent and the actual flood (resp. dry area) extent, and mIoU is their mean. We can see that the performance of U-Net and EvaNet are far from SatMAE and EvaMAE models, since they cannot utilize knowledge from pre-training. The performance of Prithvi is poor even though it is the fastest model, since it is originally pre-trained on satellite images rather than RGB ones. While SatMAE++ is not very competitive since it does not utilize DEM data and our data is not multi-scale, SatMAE turns out to be quite competitive even without using DEM data. Our EvaMAE models all produce very high IoU values thanks to the use of DEM data, with EvaMAE-Conv slightly beating SatMAE, and EvaMAE-CrossViT beating SatMAE by a large margin. When using ControlNet approach for fine-tuning, the performance of EvaMAE-Conv^{-}, EvaMAE-Conv and EvaMAE-CrossViT are further improved with ControlNet-CrossViT giving the best performance. A final observation is that we need to use EvaMAE-CrossViT instead of EvaMAE-CrossAttn to unleash the power of DEM integration in this downstream task.

To visually observe the model prediction quality, Figure 8 shows flood extent map predicted by EvaMAE-CrossViT on ERI imagery, with two regions (a) and (b) highlighted for comparing the predictions from different models. We can see that in Region (a) which is mostly inundated, U-Net, SatMAE and EvaMAE-Channel still wrongly predicts some dry 'holes' while EvaMAE-Conv and EvaMAE-CrossViT are able to provide correct predictions. In Region (b) where only a small piece of highland on the lower-left corner is dry, EvaMAE-CrossViT is able to predict the remaining area to be all flooded while the other models predict some wrong dry 'holes'.

Table 6: Landslide Segmentation Results (Unit: %)

Model	Landslide IoU	Background IoU	mIoU	Time (min)
U-Net	10.73 ± 0.31	90.66 ± 0.15	50.69 ± 0.13	**2.28 ± 0.43**
SatMAE	15.22 ± 0.43	97.75 ± 0.39	56.49 ± 0.19	3.45 ± 0.72
SatMAE++	11.55 ± 1.47	97.89 ± 0.26	54.73 ± 0.73	5.01 ± 0.93
Prithvi	0.42 ± 0.24	98.08 ± 0.32	49.25 ± 0.06	3.26 ± 0.21
EvaMAE-Channel	16.29 ± 0.80	98.12 ± 0.10	57.21 ± 0.43	3.24 ± 0.36
EvaMAE-Conv^{-}	0.08 ± 0.10	98.51 ± 0.08	49.30 ± 0.01	4.24 ± 0.32
EvaMAE-Conv	6.02 ± 1.57	**98.51 ± 0.02**	52.27 ± 0.78	4.36 ± 0.57
EvaMAE-CrossAttn	15.72 ± 1.51	98.40 ± 0.05	57.06 ± 0.74	4.01 ± 0.42
EvaMAE-CrossViT	15.25 ± 2.88	98.41 ± 0.06	56.84 ± 1.41	4.34 ± 0.61
ControlNet-Channel	18.23 ± 0.82	97.25 ± 0.70	57.79 ± 0.27	6.42 ± 0.78
ControlNet-Conv^{-}	2.01 ± 0.32	98.27 ± 0.31	50.14 ± 0.22	6.37 ± 0.46
ControlNet-Conv	15.41 ± 1.72	98.38 ± 0.04	56.90 ± 0.85	6.73 ± 0.23
ControlNet-CrossAttn	**21.02 ± 0.54**	97.82 ± 0.15	**59.42 ± 0.28**	6.19 ± 0.27
ControlNet-CrossViT	19.75 ± 1.15	98.03 ± 0.12	58.89 ± 0.55	6.40 ± 0.36

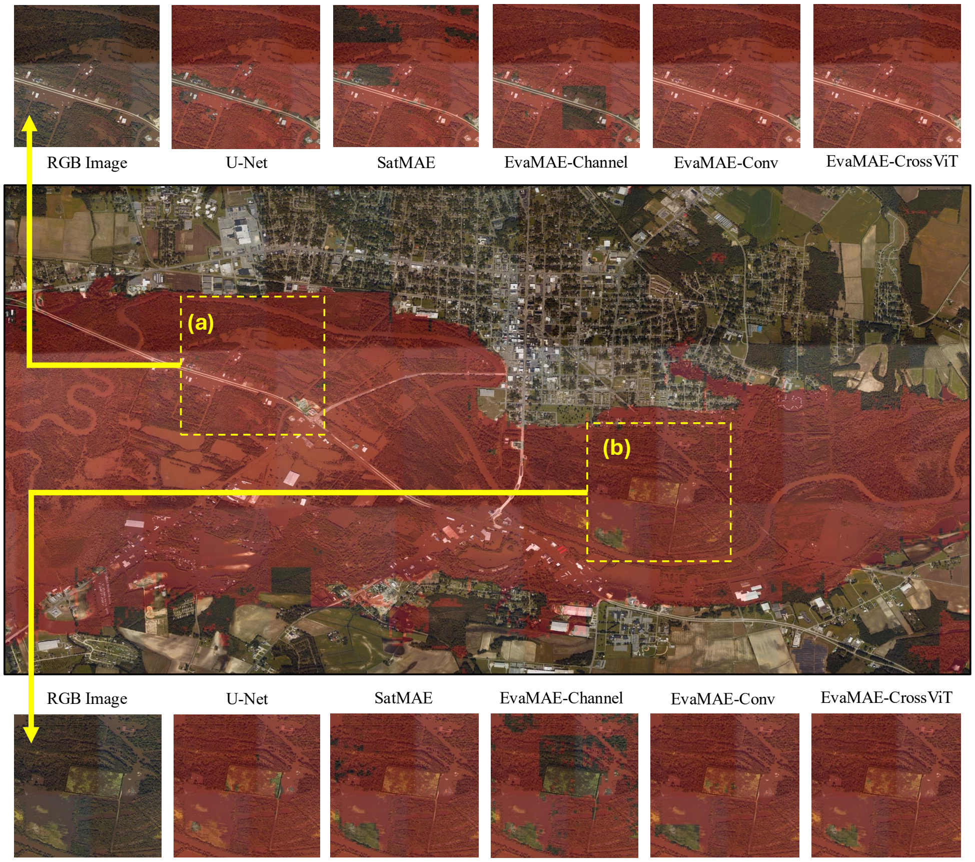

Figure 7: Visual Comparison of the Flood Segmentation Performance

Table 7: Ablation Study on # of ViT Blocks for DEM (Unit: %)

# ViT Blocks	Flood IoU	Dry IoU	mIoU
1	89.61 ± 1.13	85.52 ± 0.72	87.42 ± 0.84
2	88.61 ± 1.29	84.56 ± 0.72	86.23 ± 1.03
3	87.92 ± 1.51	84.20 ± 0.77	85.75 ± 1.17

Table 6 shows the performance results of various models for landslide segmentation, where EvaNet is excluded since it is designed for flood segmentation only. We can see that the Landslide IoU is generally low, since Landslide4Sense is low-resolution satellite imagery rather than high-resolution aerial imagery (see Figure 7), and the positive samples are limited (i.e., background dominates). In this difficult task, we observe similar results that SatMAE is still a competitive model while EvaMAE models are generally better. The differences are that (1) EvaMAE-Conv^{-} and EvaMAE-Conv do not perform well, and (2) EvaMAE-Channel is the best when ControlNet is not used, and (3) EvaMAE-CrossAttn is better than EvaMAE-CrossViT, (4) ControlNet improves the performance of all EvaMAE models with ControlNet-CrossAttn giving the best performance beating all others by a large margin.

Finally, recall that we only use one transformer layer to encode DEM data before using the encoded data for feature fusion. Table 7 shows the results of flood segmentation prediction when we vary the number of transformer layers to encode DEM data as 1, 2 and 3. We can see that using only one transformer layer to encode DEM data indeed gives the best performance.

5 Conclusion

We explored various methods to integrate DEM data into the pre-training and fine-tuning stages of geo-foundation models for Earth imagery, collectively called EvaMAE models. We found that incorporating DEM data is helpful and the most effective when using cross-attention operations, and DEM conditioning with ControlNet is helpful during fine-tuning. Our contributions also include the comprehensive experimental study and comparisons, and a new flood-related dataset, EvaFlood, for pre-training and fine-tuning.

Acknowledgments

The authors at Indiana University Bloomington (IUB) were supported by DOE ECRP Award DE-SC0025228, NSF OAC-2414474 and OAC-2414185, Radiance Technologies Innovation Bowl grand prize awarded as the competition winner in 2023, and 2024–2025 Luddy Faculty Fellow Award from IUB. Xiaodong Yu was funded by NSF 2348465 and Zhe Jiang was funded by NSF OAC-2152085, OAC-2402946, OAC-2410884, and IIS-2207072. This research used resources of the Argonne Leadership Computing Facility, a U.S. Department of Energy (DOE) Office of Science user facility at Argonne National Laboratory and is based on research supported by the U.S. DOE Office of Science-Advanced Scientific Computing Research Program, under Contract No. DE-AC02-06CH11357.

References

[1] [n. d.]. Landslide4Sense. https://eod-grss-ieee.com/dataset-detail/MkJ2T3pJM0p3eGh1QkZwRVZFa0FsZz09.

[2] [n. d.]. NOAA Emergency Response Imagery. https://storms.ngs.noaa.gov.

[3] Saugat Adhikari, Da Yan, Zhe Jiang, Jiao Han, Zelin Xu, Yupu Zhang, Arpan Man Sainju, and Yang Zhou. 2025. Scaling Terrain-Aware Spatial Machine Learning for Flood Mapping on Large Scale Earth Imagery Data. *ACM Trans. Spatial Algorithms Syst.* 11, 2 (2025), 9:1–9:29.

[4] Saugat Adhikari, Da Yan, Mirza Tanzim Sami, Jalal Khalil, Lyuheng Yuan, Bhadhan Roy Joy, Zhe Jiang, and Arpan Man Sainju. 2022. An elevation-guided annotation tool for flood extent mapping on earth imagery (demo paper). In *SIGSPATIAL*, Matthias Renz and Mohamed Sarwat (Eds.). ACM, 28:1–28:4.

[5] Saugat Adhikari, Da Yan, Tianyang Wang, Landon Dyken, Sidharth Kumar, Lyuheng Yuan, Akhlaque Ahmad, Jiao Han, Yang Zhou, and Steve Petruzza. 2025. Faster Annotation for Elevation-Guided Flood Extent Mapping by Consistency-Enhanced Active Learning. In *IJCAI*.

[6] Peri Akiva, Matthew Purri, and Matthew J. Leotta. 2022. Self-Supervised Material and Texture Representation Learning for Remote Sensing Tasks. In *CVPR*. IEEE, 8193–8205.

[7] Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loïc Landrieu. 2024. OmniSat: Self-supervised Modality Fusion for Earth Observation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXVIII (Lecture Notes in Computer Science, Vol. 15086)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 409–427. https://doi.org/10.1007/978-3-031-73390-1_24

[8] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David B. Lobell, and Stefano Ermon. 2021. Geography-Aware Self-Supervised Learning. In *ICCV*. IEEE, 10161–10170.

[9] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. 2023. SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding. In *ICCV*. IEEE, 16726–16736.

[10] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. 2020. Sen1Floods11: a georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In *CVPR*. Computer Vision Foundation / IEEE, 835–845.

[11] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020. Improved Baselines with Momentum Contrastive Learning. *CoRR* abs/2003.04297 (2020).

[12] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. 2022. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. In *NeurIPS*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.).

[13] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language Modeling with Gated Convolutional Networks. In *ICML (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 933–941.

[14] 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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. OpenReview.net.

[15] Landon Dyken, Saugat Adhikari, Pravin Poudel, Steve Petruzza, Da Yan, Will Usher, and Sidharth Kumar. 2025. Enabling Fast and Accurate Crowdsourced Annotation for Elevation-Aware Flood Extent Mapping. In *PacificVis*.

[16] Anthony Fuller, Koreen Millard, and James R. Green. 2023. CROMA: Remote Sensing Representations with Contrastive Radar-Optical Masked Autoencoders. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/11822e84689e631615199db3b75cd0e4-Abstract-Conference.html

[17] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. 2024. SkySense: A Multi-Modal Remote Sensing Foundation Model Towards Universal Interpretation for Earth Observation Imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 27662–27673. https://doi.org/10.1109/CVPR52733.2024.02613

[18] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. 2024. Bridging Remote Sensors with Multisensor Geospatial Foundation Models. In *CVPR*. IEEE, 27852–27862.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*. IEEE, 15979–15988.

[20] Wenchong He, Zhe Jiang, Marcus Kriby, Yiqun Xie, Xiaowei Jia, Da Yan, and Yang Zhou. 2022. Quantifying and Reducing Registration Uncertainty of Spatial Vector Labels on Earth Imagery. In *KDD*, Aidong Zhang and Huzefa Rangwala (Eds.). ACM, 554–564.

[21] Wenchong He, Arpan Man Sainju, Zhe Jiang, and Da Yan. 2021. Deep Neural Network for 3D Surface Segmentation based on Contour Tree Hierarchy. In *SDM*, Carlotta Demeniconi and Ian Davidson (Eds.). SIAM, 253–261.

[22] Wenchong He, Arpan Man Sainju, Zhe Jiang, Da Yan, and Yang Zhou. 2022. Earth Imagery Segmentation on Terrain Surface with Limited Training Labels: A Semi-supervised Approach based on Physics-Guided Graph Co-Training. *ACM Trans. Intell. Syst. Technol.* 13, 2 (2022), 26:1–26:22.

[23] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jón Atli Benediktsson, and Jocelyn Chanussot. 2024. SpectralGPT: Spectral Remote Sensing Foundation Model. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 8 (2024), 5227–5244.

[24] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dário A. B. Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu K. Ganti, Kommy Weldemariam, and Rahul Ramachandran. 2023. Foundation Models for Generalist Geospatial Artificial Intelligence. *CoRR* abs/2310.18660 (2023).

[25] Wentao Jiang, Jing Zhang, Di Wang, Qiming Zhang, Zengmao Wang, and Bo Du. 2024. LeMeViT: Efficient Vision Transformer with Learnable Meta Tokens for Remote Sensing Image Interpretation. In *IJCAI*. ijcai.org, 929–937.

[26] Zhe Jiang, Wenchong He, Marcus Stephen Kirby, Sultan Asiri, and Da Yan. 2021. Weakly Supervised Spatial Deep Learning based on Imperfect Vector Labels with Registration Errors. In *KDD*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 767–775.

[27] Zhe Jiang, Yupu Zhang, Saugat Adhikari, Da Yan, Arpan Man Sainju, Xiaowei Jia, and Yiqun Xie. 2023. A Hidden Markov Forest Model for Terrain-Aware Flood Inundation Mapping from Earth Imagery. In *SDM*, Shashi Shekhar, Zhi-Hua Zhou, Yao-Yi Chiang, and Gregor Stiglic (Eds.). SIAM, 316–324.

[28] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. 2024. S2MAE: A Spatial-Spectral Pretraining Foundation Model for Spectral Remote Sensing Data. In *CVPR*. IEEE, 27696–27705.

[29] Zhihao Li, Biao Hou, Siteng Ma, Zitong Wu, Xianpeng Guo, Bo Ren, and Licheng Jiao. 2024. Masked Angle-Aware Autoencoder for Remote Sensing Images. In *ECCV (Lecture Notes in Computer Science, Vol. 15066)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 260–278.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*. IEEE, 9992–10002.

[31] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. 2023. Change-Aware Sampling and Contrastive Learning for Satellite Images. In *CVPR*. IEEE, 5261–5270.

[32] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i-Nieto, David Vázquez, and Pau Rodríguez. 2021. Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data. In *ICCV*. IEEE, 9394–9403.

[33] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman H. Khan, and Fahad Shahbaz Khan. 2024. Rethinking Transformers Pre-training for Multi-Spectral Satellite Imagery. In *CVPR*. IEEE, 27811–27819.

[34] Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. 2023. Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning. In *ICCV*. IEEE, 4065–4076.

[35] Arpan Man Sainju, Wenchong He, Zhe Jiang, and Da Yan. 2020. Spatial Classification with Limited Observations Based on Physics-Aware Structural Constraint. In *AAAI*. AAAI Press, 898–905.

[36] Arpan Man Sainju, Wenchong He, Zhe Jiang, Da Yan, and Haiquan Chen. 2021. Flood Inundation Mapping with Limited Observations Based on Physics-Aware Topography Constraint. *Frontiers Big Data* 4 (2021), 707951.

[37] Mirza Tanzim Sami, Da Yan, Saugat Adhikari, Lyuheng Yuan, Jiao Han, Zhe Jiang, Jalal Khalil, and Yang Zhou. 2024. EvaNet: Elevation-Guided Flood Extent Mapping on Earth Imagery. In *IJCAI*. ijcai.org, 1200–1208.

[38] 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. 2023. RingMo: A Remote Sensing Foundation Model With Masked Image Modeling. *IEEE Trans. Geosci. Remote. Sens.* 61 (2023), 1–22. https://doi.org/10.1109/TGRS.2022.3194732

[39] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. 2023. Cross-Scale MAE: A Tale of Multiscale Exploitation in Remote Sensing. In *NeurIPS*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.).

[40] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018).

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All

You Need. In *NeurIPS*. 6000–6010.

[42] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. 2022. Advancing Plain Vision Transformer Towards Remote Sensing Foundation Model. *CoRR* abs/2208.03987 (2022).

[43] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. SimMIM: a Simple Framework for Masked Image Modeling. In *CVPR*. IEEE, 9643–9653.

[44] Zelin Xu, Tingsong Xiao, Wenchong He, Yu Wang, Zhe Jiang, Shigang Chen, Yiqun Xie, Xiaowei Jia, Da Yan, and Yang Zhou. 2024. Spatial-Logic-Aware Weakly Supervised Learning for Flood Mapping on Earth Imagery. In *AAAI*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 22457–22465.

[45] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*. IEEE, 3813–3824.