arXiv:2212.14532v2 [cs.CV] 2 Jan 2023

Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning

Colorado J Reed[1,2]*, Ritwik Gupta[1]*, Shufan Li[1]*,
Sarah Brockman[3], Christopher Funk[3], Brian Clipp[3],
Kurt Keutzer[1], Salvatore Candido[2], Matt Uyttendaele[2], Trevor Darrell[1]

[1]Berkeley AI Research; [2]Meta AI, FAIR; [3]Kitware Inc.

correspondence to cjrd@cs.berkeley.edu

Abstract

Remote sensing imagery provides comprehensive views of the Earth, where different sensors collect complementary data at different spatial scales. Large, pretrained models are commonly finetuned with imagery that is heavily augmented to mimic different conditions and scales, with the resulting models used for various tasks with imagery from a range of spatial scales. Such models overlook scale-specific information in the data. In this paper, we present Scale-MAE, a pretraining method that explicitly learns relationships between data at different, known scales throughout the pretraining process. Scale-MAE pretrains a network by masking an input image at a known input scale, where the area of the Earth covered by the image determines the scale of the ViT positional encoding, not the image resolution. Scale-MAE encodes the masked image with a standard ViT backbone, and then decodes the masked image through a bandpass filter to reconstruct low/high frequency images at lower/higher scales. We find that tasking the network with reconstructing both low/high frequency images leads to robust multiscale representations for remote sensing imagery. Scale-MAE achieves an average of a 5.0% *non-parametric kNN classification improvement across eight remote sensing datasets compared to current state-of-the-art and obtains a* 0.9 *mIoU to* 3.8 *mIoU improvement on the SpaceNet building segmentation transfer task for a range of evaluation scales.*

1. Introduction

Remote sensing data is captured from satellites and planes through a mixture of sensors, processing pipelines, and viewing geometries. Depending on the composition and relative geometry of the sensor to the Earth, each image's Ground Sample Distance (GSD - the physical distance between two adjacent pixels in an image) can vary from 0.3m to 1km, so a 100x100 pixel image could span anywhere from an Olympic-size swimming pool (900 m^2) to almost the entire country of Jamaica (10,000 km^2). The data within each image, and the corresponding objects and points of interest, can therefore vary across wide spatial ranges. Data from these multiscale sensors provide critical and complementary information for various operational and research applications in areas such as atmospheric, hydrologic, agricultural, and environmental monitoring [43, 50].

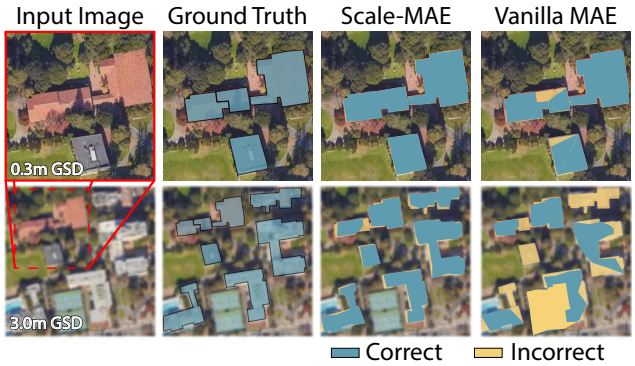

Figure 1. ***Scale-MAE* learns better representations for multiscale tasks compared to vanilla MAE.** (Column 1) The top image spans an area at 0.3m GSD and the bottom image shows the same region at a coarser GSD. (Columns 2-4) The following columns show a ground truth building segmentation, *Scale-MAE* segmentation from a finetuned UperNet, and segmentation from an analogously finetuned UperNet from a vanilla MAE, respectively. *Scale-MAE* demonstrates better performance across images at both scales. See the supplementary material for more examples.

Few modern computer vision methods have explicitly addressed multiscale remote sensing imagery [34]. Nevertheless, the remote sensing vision community has increasingly

*Denotes co-first authorship. Co-first authors will prioritize their names on their resumes/websites.

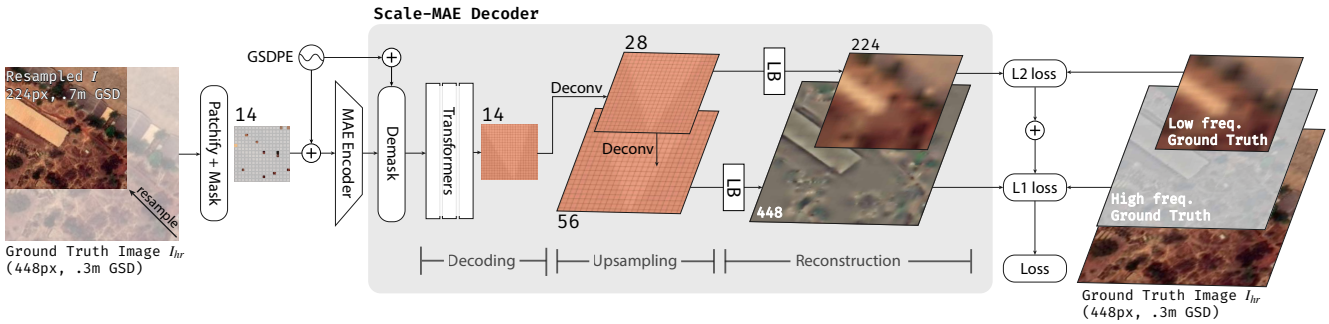

Figure 2. ***Scale-MAE* architecture.** Following the Masked Autoencoder framework, an input image is patchified and masked before being passed into an MAE encoder. A Ground Sample Distance Positional Encoding (GSDPE) is added to the encoder input, which scales the positional encodings to the area of ground covered. The *Scale-MAE* decoders has three stages: (1) Decoding, which uses a smaller number of transformer layers than MAE to decode the encoded values (2) Upsampling, which progressively deconvolves the decoded feature map to a larger size before being passed through the Laplacian Blocks (abbreviated LB, see Section 3), (3) Reconstruction, which then reconstructs low and high frequency features at different scales. These outputs are used to compute an aggregate loss with ground truth low and high frequency features, where following super resolution literature [2], an L1 loss is used for high frequency output to better reconstruct edges and an L2 loss is used for low frequency output to better reconstruct average values.

used large, pretrained models [13, 20], where such applications finetune a pretrained model for a single source of data at a specific scale [13, 20, 21, 31, 40]. In this paper we present *Scale-MAE*, a masked reconstruction model that explicitly learns relationships between data at different, known scales throughout the pretraining process. By leveraging this information, *Scale-MAE* produces a pretrained model that performs better across a wide range of GSDs and tasks.

Scale-MAE is a self-supervised pretraining framework based on the Masked Autoencoder (MAE) [25]. A standard MAE resizes/crops an image, masks the majority of the transformed image, and then tasks a Vision Transformer (ViT) based autoencoder with embedding the unmasked components. A decoding ViT then decodes the full image from these learned embeddings, where the decoder is later discarded and the encoder is used to produce representations for an unmasked input image.

Scale-MAE has two key differences that lead to its strong performance across spatial scales (Figure 1). First, a standard MAE uses relative positional encodings to inform the ViT of the position of the unmasked components. *Scale-MAE* uses a GSD-based positional encoding derived from the land area covered in the image which informs the ViT of both the position *and scale* of the input image. Second, *Scale-MAE* uses a Laplacian-pyramid decoder to encourage the network to learn multiscale representations. The embeddings are decoded to two images containing low and residual high frequency information, respectively – see Figure 2. As we discuss in Section 3, this structure allows the ViT decoder to use fewer parameters than MAE while still producing strong representations across multiple scales.

We show that *Scale-MAE* leads to better performing, more robust multiscale representations than both a standard MAE and a recently proposed, state-of-the-art remote sensing MAE called SatMAE [13] across remote sensing datasets with a variety of scale and resolution characteristics. Our contributions can be summarized as follows:

1. We present *Scale-MAE*, a pretraining method that leads to better representations for remote sensing data that are more robust across a range of spatial scales. *Scale-MAE* has two key contributions: a scale-aware positional encoding and a Laplacian pyramid decoder that decodes low and high frequency features at multiple scales.

2. *Scale-MAE* achieves an average of a 5.0% nonparametric kNN classification improvement across eight remote sensing datasets compared to current state-of-the-art.

3. *Scale-MAE* obtains a 0.9 mIoU to 3.8 mIoU improvement on the SpaceNet building segmentation transfer task for a range of evaluation scales (see Figure 1).

2. Related Work

Representation learning and the Masked Autoencoder. Representation learning aims to extract meaningful, intrinsic features from data for downstream use [5]. In practice, this often entails pretraining a deep network so that a lightweight learning routine can then finetune it for a particular downstream task, see [15–17, 23, 26, 29, 36, 47, 63]. The Masked Autoencoder (MAE) is a recent state-of-the-art self-supervised representation learning method in computer vision that pretrains a ViT encoder by masking an image, feeding the unmasked portion into a transformer-based encoder, and then tasking the decoder with reconstructing the input image [25]. *Scale-MAE* builds upon the MAE and learns representations specifically designed to be robust with multiscale remote sensing imagery.

Remote Sensing Representation Learning Neumann *et al*. [44] were one of the first to exhaustively share results on existing representation learning and semi-supervised learning techniques for remote sensing imagery. Gao *et al*. [21] demonstrated the effectiveness of MAE pretraining for remote sensing image classification. Ayush *et al*. [3] leveraged the metadata from remote sensing images via spatially aligned but temporally separated images as positive pairs for contrastive learning and predicted the latitude and longitude as pretext tasks. Gupta *et al*. [24] demonstrated the use of MAEs as a pretraining approach for passive and active remote sensing imagery. Their method introduced flexible "adapters" which could be used interchangeably with an encoder for a set of input imagery modes. Cong *et al*. [13] introduced the SatMAE, which used temporal and spectral metadata in a positional encoding to encode spatio-temporal relationships in data. The temporal data contains the year, month, and hour enabling understanding of long-term change with the year, weather information from the month, and hour information for the time of day. Further Liu *et al*. [40] and Ibañez *et al*. [31] have shown that MAE architectures can be used for band selection in hyperspectral remote sensing images, significantly reducing data redundancy while maintaining high classification accuracy. *Scale-MAE* leverages inherent absolute scale information present in remote sensing data as a way to learn robust, multiscale features that reduce data usage downstream.

Super-resolution Super-resolution has proven effective in improving accuracy within remote sensing images due to the extremely small size of objects within the image [49]. Previous works have aimed to learn continuous implicit representations for images at arbitrary resolutions to aid the super-resolution task. These representations are used to upsample the images either to specific scales [37] or to arbitrary resolutions [10, 30, 58]. Most super-resolution work aims to increase the resolution of the input image, whereas *Scale-MAE* produces both higher *and lower* resolution images. There is some work on super-resolution for satellite imagery, but much of this work is focused on synthetically creating high-resolution datasets for use with models trained specifically for high-resolution data [27, 34]. *Scale-MAE*, however, utilizes super-resolution as a means to obtain multiscale representations during pretraining.

Features at multiple scales. Because images can contain objects of many different pixel resolutions, the vision community has proposed many methods to extract multiscale features. These include spatial pyramids [6, 33, 35, 48] and dense sampling of windows [32, 59, 60]. These approaches have been combined by methods such as [19], in which dense histogram-of-gradient features are computed for each feature pyramid level. Rather than using classical computer vision techniques to extract multiscale features, convolutional neural networks have been used to build deep multiscale features. CNNs with subsampling layers inherently build feature pyramids, a property exploited explicitly by [22, 38, 39]. Recently, this multiscale idea has been extended to vision transformers by [18], who show that this architecture improves various video recognition and image classification tasks. *Scale-MAE* reconstructs a version of Laplacian pyramids as a way to force an encoder to learn multi-scale features.

3. Scale-MAE

This section describes the *Scale-MAE* pretraining framework as illustrated in Figure 2. *Scale-MAE* is a self-supervised pretraining framework based on the Masked Autoencoder (MAE) [25]. *Scale-MAE* has two key differences: first, a standard MAE uses relative positional encodings to inform the ViT of the position of the unmasked components, where an image at resolution r will have the same positional encodings regardless of the image content. *Scale-MAE*, however, uses a Ground Sample Distance (GSD) based positional encoding that scales in proportion to the area of land in an image, regardless of the resolution of the image. Second, to encourage the network to learn multiscale representations, *Scale-MAE* uses a Laplacian-pyramid decoder, where the embeddings are decoded to a lower resolution image that captures the lower frequency information and a higher resolution image that captures the high-frequency information. We formalize *Scale-MAE* in the following subsections by first specifying the necessary MAE background, describing the GSD-based positional encoding, and then explaining the Laplacian-pyramid decoder.

Setup Let $I \in \mathbb{R}^{H \times W \times C}$ represent an input image of height H, width W, and C channels. The MAE *patchifies* I into a sequence S of independent patches of height and width P pixels, where each of the N_p patches, $s \in S$ has dimension $s \in \mathbb{R}^{P^2C}$. A fraction, m, of the patches are then removed and the remaining patches are then passed through a projection function (e.g., a linear layer) to project the patches S into D dimensions, $f_E : \mathbb{R}^{P^2C} \rightarrow \mathbb{R}^D$, to obtain embedded patches $S_E = f_E(S)$. An $\mathbb{R}^2$ positional encoding vector, is then added to the embedded patches with

$$v_x(pos, 2i) = \sin \frac{pos}{10000^{\frac{2i}{D}}} \quad (1)$$

$$v_y(pos, 2i+1) = \cos \frac{pos}{10000^{\frac{2i}{D}}} \quad (2)$$

where pos is the position of the patch along the given axis and i is the feature index (visualized in Figure 3), exactly as introduced in [52]. These positional encodings are then concatenated and added to the embedded patches, which are then fed into a ViT encoder. After the encoder, the removed

m patches are then placed back into their original location in the sequence of patches where a *learned mask token* represents the masked patches that were not encoded. Another positional encoding vector is added to all patches and a sequence of transformer blocks decodes these patches to form the original input image, which is used as the learning target.

Input Unlike MAE, *Scale-MAE* performs a super resolution reconstruction, where the input image I is downsampled from a higher resolution image I_{hr} at the ground truth GSD. Instead of targeting the input image, *Scale-MAE* targets high frequency and low frequency components of I_{hr}, which is common in Laplacian pyramid super resolution models [61], where the high frequency component is at the same resolution as the ground truth image I_{hr} and the low frequency component is at the same resolution as the input image I, as shown in Figure 2. Following many works in super resolution [61], the low frequency target image is obtained by interpolating I_{hr} to a much lower resolution, r_{low} and then interpolating to the same resolution as the input image I. The high frequency target image is obtained by downsampling I_{hr} to another lower resolution $r_{\text{high-low}}$, and then upsampling to the same resolution as the ground truth image I_{hr} and subtracting this image $I_{\text{hf}} = I_{\text{hr}} - I_{\text{high-low}}$. The supplementary material provide more information on the upsampling/downsampling methodology. The key components for *Scale-MAE* are described next.

GSD Positional Encoding Remote sensing images inherently have the concept of GSD, a metric which defines the absolute scale for the image. GSD is critical to understanding, conceptually, the kinds of features that will be available in an image. An image with finer GSD (lower number) will have higher frequency details than an image with coarser GSD (high number). Models are generally unaware of absolute scale when learning over a set of data. Specifically, even if they implicitly learn that all images in a dataset share a varying resolution from input-space augmentations, then these models do not explicitly condition on the GSDs encountered in unseen data.

We extend the positional encoding from Equation (2) to include GSD by scaling the positional encoding relative to the land area covered in an image as depicted in Figure 3 and mathematically:

$$v_{gsd,x}(pos, 2i) = \sin \frac{G}{g} \frac{pos}{10000^{\frac{2i}{D}}} \tag{3}$$

$$v_{gsd,y}(pos, 2i+1) = \cos \frac{G}{g} \frac{pos}{10000^{\frac{2i}{D}}} \tag{4}$$

where g is the GSD of the image and G is a reference GSD, nominally set to 1m. Intuitively, an object imaged at a finer resolution has more pixels representing it. When imaging the same object at a coarser resolution, those pixels must map to fewer pixels. In Equation (4), we interpolate the positional encoding by a factor of $\frac{G}{g}$ to account for the ordering of the coarser set of pixels. This simple idea underpins the GSD Positional Encoding, visualized in Figure 3.

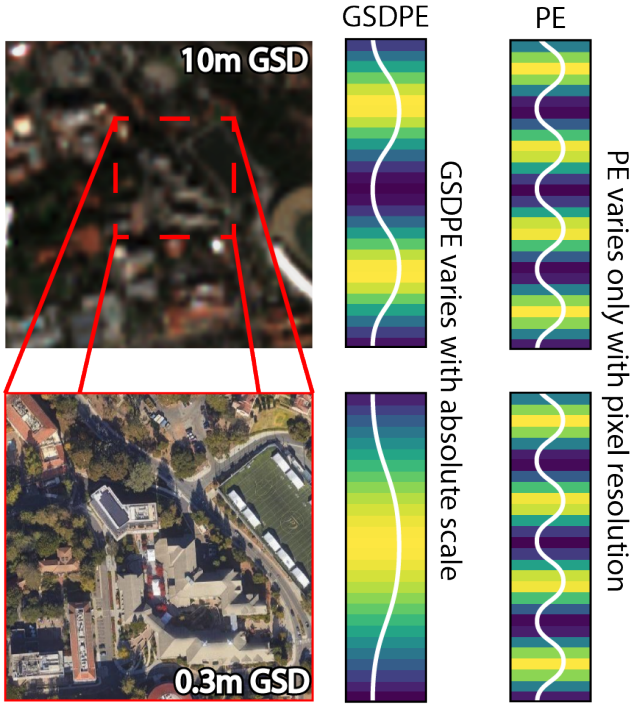

Figure 3. **The difference between GSDPE and a standard Positional Encoding (PE).** (Left) Input images at the same pixel resolution but different GSDs are shown. The image on the top is a subset of the image on the bottom. (Center) This overlap in location, albeit at a different resolution, is reflected in the GSDPE. The finer image with smaller spatial extent is represented by a corresponding subsection of the overall sine wave on the bottom. (Right) A standard positional encoding is strictly dependent on the image resolution and uses the same embedding for both. The colors behind the sine waves show the intensity and quantization of the encoding.

Scale-MAE decoder The standard MAE learns representations by tasking a network with reconstructing an image after masking out most of its pixels. While the standard MAE decoder reconstructs the input image at the same scale as its input, the objective of *Scale-MAE* is to learn multi-scale representations. We draw on works from progressive super-resolution such as [54], that learn a high resolution, high frequency image and a lower resolution low frequency image, that when combined together, yield the input image at a higher resolution.

The standard MAE decoder reconstructs to the input image scale by using a series of transformer layers. The

Scale-MAE decoder, on the other hand, decodes to multiple scales by replacing the majority of the MAE decoder's transformer layers with a progressive Laplacian decoder architecture. This architecture consists of three stages: decoding, upsampling, and reconstruction, which are shown in Figure 2 and detailed below.

Decoding follows the standard MAE decoder where following the encoder, the removed m patches are then placed back into their original location in the sequence of patches where a *learned mask token* represents the masked patches that were not encoded, a positional encoding is added, and then a series of transformer layers decode all patches. In contrast to the standard MAE decoder, the *Scale-MAE* decoder uses fewer transformer layers (e.g. 3 layers instead of 8), which more than offsets the parameter complexity as quantified in Section 5. The output of these layers is then fed into the upsampling stage.

Upsampling The latent feature maps from the decoding stage are progressively upsampled to 2x and 4x resolution using deconvolution blocks, where the first deconvolution is 2x2 with stride 2 that outputs a feature map at 2x the input resolution (28 in Figure 2), followed by a LayerNorm and GELU, and then another 2x2 deconvolution layer that outputs a feature maps at 2x the previous resolution (56 in Figure 2). See the supplementary material for a full architectural diagram.

Reconstruction After having been upsampled, the lower resolution and higher resolution feature maps are passed into *Laplacian Blocks* (LBs in Figure 2) that reconstruct high and low resolution images for the high and low frequency reconstruction, respectively. Architecturally, the Laplacian Blocks consist of a sequence of three sub-blocks: a Laplacian Feature Mapping Block, a Laplacian Upsample Block, and a Laplacian Pyramid Reconstruction Block. The Feature Mapping Block is used to project features within a particular layer of the Laplacian Pyramid back to the RGB space. The Laplacian Upsample Block represents a learnable upsample function that maps latent features from one layer of the Laplacian Pyramid to a higher level. Finally, the Laplacian Pyramid Reconstruction Block is used to reconstruct information at the different frequencies in RGB space. Following super resolution literature [2], an L1 loss is used for high frequency output to better reconstruct edges and an L2 loss is used for low frequency output to better reconstruct average values. The supplementary material has architectural diagrams for each block.

4. Experiments

We investigate the quality of representations learned from *Scale-MAE* pretraining through a set of experiments that explore their robustness to scale as well as their transfer performance to additional tasks. First, we present our main experiments in Section 4.1 and compare with SatMAE [13], a current state-of-the-art MAE for remote sensing imagery, as well as several other approaches detailed throughout. The exact implementation of *Scale-MAE* for the main experiments was determined through a set of ablation experiments presented in Section 4.2.

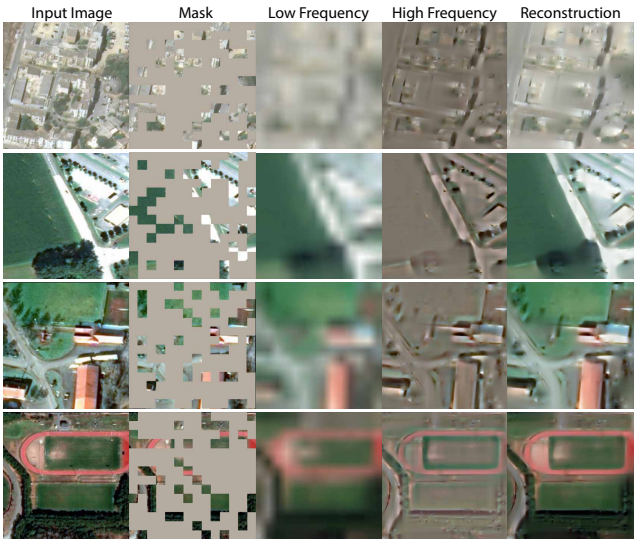

Figure 4. ***Scale-MAE* reconstruction.** Examples from Functional Map of the World are shown. From left to right, an input image at 224x224 resolution is shown. Its corresponding mask is visualized as well. Columns 3 and 4 show the low and high frequency produced by the *Scale-MAE* decoder. The last column is the reconstruction obtained from summing the low and high frequency features together.

In order to compare with SatMAE [13], for our main experiments, we pretrain *Scale-MAE* with a ViT-Large model using the Functional Map of the World (FMoW) [12] RGB training set, which consists of 363.6k images of varying image resolution and GSD. The initial higher resolution image I_{hr} is taken as a random 448px^2 crop of the input image, and the input image I is then a downsampled 224px^2 from I_{hr}. The low frequency groundtruth is obtained by downscaling I_{hr} to 14px^2 and then upscaling to 224px^2, while the high frequency groundtruth is obtained by downscaling I_{hr} to 56px^2 and then upscaling to 448px^2 and subtracting this image from I_{hr}.

Figure 4 shows examples of the masked input, low resolution/frequency, high resolution/frequency, and combined reconstruction of FMoW images during training. The low resolution/frequency images capture color gradients and landscapes, while the residual high resolution/frequency images capture object edges, roads, and building outlines.

4.1. Representation Quality

We evaluate the quality of representations from *Scale-MAE* by freezing the encoder and performing a non-

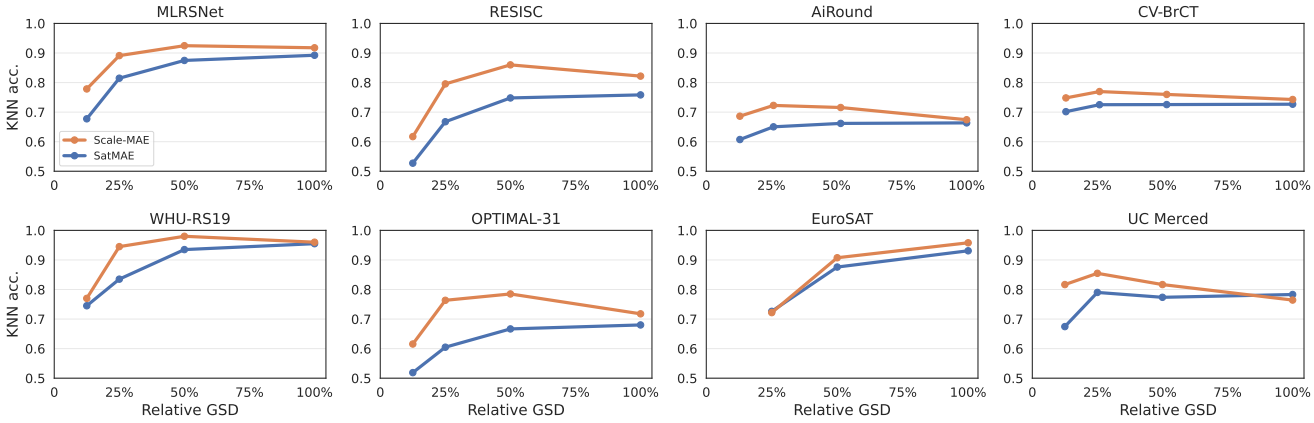

Figure 5. **Learning better representations at all scales.** *Scale-MAE* (orange) features perform better than state-of-the-art. We evaluate kNN accuracy on eight datasets with a large variance in GSD. *Scale-MAE* consistently produces better results at coarser resolutions. In addition to using evaluation datasets at different GSDs, to further test the multiscale representations, we create multiple test sets for each dataset in which we downsampled the full resolution validation set to coarser GSDs at fixed percentages: $X_{val}^{G\%}, G \in \{12.5, 25, 50, 100\}$, where Eurosat does not include the 12.5% because the images are at a resolution of 64px, our patch size is 16px, and an input image of 8px is too small.

parametric k-nearest-neighbor (kNN) classification with eight different remote sensing imagery classification datasets with different GSDs, none of which were encountered during pretraining. The kNN classifier operates by encoding all train and validation instances, where each embedded instance in the validation set computes the cosine distance with each embedded instance in the training set, where the instance is classified correctly if the majority of its k-nearest-neighbors are in the same class as the validation instance.

The reasoning behind the kNN classifier evaluation is that a strong pretrained network will output semantically grouped representation for unseen data of the same class. This evaluation for the quality of representations occurs in other notable works [7, 9, 55]. In addition to using evaluation datasets at different GSDs, to further test the multiscale representations, we create multiple test sets for each dataset. Since we cannot synthesize data at a finer GSD than the provided ground truth, we only downsample the full resolution validation set to coarser GSDs at fixed percentages: $X_{val}^{G\%}, G \in \{12.5, 25, 50, 100\}$.

Our analysis uses eight different land-use classification datasets: RESISC-45 [11], the UC Merced Land Use Dataset [62], AiRound and CV-BrCT [42], MLRSNet [46], EuroSAT [28], Optimal-31 [53], WHU-RS19 [14], SpaceNet v1 and v2 [51], and Functional Map of the World [12]. The datasets used span a wide range of GSDs, e.g., MLRSNet consists of data captured from aerial platforms with 0.1m GSD, while RESISC45 has imagery from medium-resolution satellites at >30m GSD. In some cases, the datasets present imagery at mixed GSDs which are not specified, in which case we assume an approximate constant GSD: see the supplementary material for all details. Furthermore, we provide an expanded set of experiments with linear probing and fine-tuning in the supplementary material.

Method	Backbone	Model	mIoU
Sup. (Scratch)	ResNet50	PSANet	75.6
GASSL [3]	ResNet50	PSANet	78.5
Sup. (Scratch)	ViT-Large	PSANet	74.7
SatMAE [13]	ViT-Large	PSANet	78.1
Sup. (Scratch)	ViT-Large	UperNet	71.6
Vanilla MAE	ViT-Large	UperNet	77.9
SatMAE	ViT-Large	UperNet	78.0
Scale-MAE	ViT-Large	UperNet	**78.9**

Table 1. Semantic segmentation results on SpaceNet v1. *Scale-MAE* outperforms other methods across backbone and segmentation architectures, where Sup. (Scratch) indicates a supervised model trained from scratch (a randomly initialized network).

We run kNN classification with $k = 20$. Figure 5 shows that *Scale-MAE* outperforms SatMAE across GSD scales in the different evaluation datasets and across relative GSD scales within individual datasets. At the lowest fixed GSD, UC Merced has a GSD of 0.3m, where evaluating at scales [12.5%,100%] provides an artificial GSD range of [0.3m, 2.4m], we see that *Scale-MAE* has the largest performance gap at the 2.4m GSD, with similar performance at 0.3m.

Across the other datasets and range of GSDs, *Scale-MAE* outperforms SatMAE, where the performance gap tends to grow as the GSD varies from the original GSD, indicating that *Scale-MAE* learns representations that are more robust

Dataset	Average Accuracy (%) *Scale-MAE*	SatMAE
AiRound	**62.7** (+4.9)	57.8
CV-BrCT	**69.6** (+3.4)	66.2
EuroSAT	**86.2** (+1.8)	84.4
MLRSNet	**80.8** (+5.8)	75.0
OPTIMAL-31	**64.1** (+8.4)	55.7
RESISC	**69.1** (+8.1)	61.0
UC Merced	**74.0** (+4.2)	69.8
WHU-RS19	**81.5** (+3.0)	78.5

Table 2. *Scale-MAE* performs better, across all GSDs (as in Figure 5), for all datasets we experimented with compared to SatMAE. The average improvement across all dataset for *Scale-MAE* compared to SatMAE is 5.0%.

to changes in scale for remote sensing imagery. We outperform SatMAE by an average of 5.0% across all resolutions and datasets (see Table 2). UC Merced at 100% of the true GSD is the only evaluation where SatMAE outperforms *Scale-MAE*. The supplementary material contains an extensive table demonstrating kNN classification results with varying k.

Semantic segmentation transfer We use the SpaceNet v1 building segmentation dataset [51] to evaluate semantic segmentation results on contrastive and MAE-based pretraining methods. Prior methods relied on the PSANet [64] segmentation architecture, while *Scale-MAE* uses the UperNet [56] segmentation architecture which is more common for ViT backbones. For even comparison, we test the current state-of-the-art SatMAE method with UperNet as well. Results are detailed in Table 1.

Scale-MAE outperforms SatMAE by 0.9 mIoU and a vanilla MAE (with the same pretraining settings) by 1.0 mIoU. *Scale-MAE* outperforms all other prior work, including GASSL [3], which SatMAE did not outperform on the mean Intersection over Union (mIoU) metric for semantic segmentation. Particularly, *Scale-MAE* increases the gap in performance as the resolution of input imagery becomes coarser, highlighting the absolute scale-invariance introduced by our method.

In Figure 6, we compare SpaceNet v1 evaluations across downscaled images (at 50%, 75%, and 100% of the original image size) for *Scale-MAE* and SatMAE. Similar to the classification results, *Scale-MAE* maintains higher semantic segmentation performance over SatMAE, even with images at a coarser GSD. In fact, the performance gap grows at coarser GSDs, where at the input GSD, *Scale-MAE* is 0.9 mIoU higher, at 75% GSD *Scale-MAE* is 2.9 mIoU higher, and at 50% *Scale-MAE* is 3.8 mIoU higher.

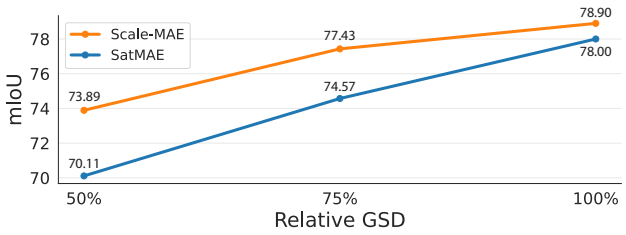

Figure 6. SpaceNet v1 evaluation across downscaled images for both *Scale-MAE* and SatMAE. *Scale-MAE* maintains higher semantic segmentation performance over SatMAE, even with images of coarser GSD.

Method	GSDPE	KNN 50%	KNN 100%
Vanilla MAE		72.8	77.8
Vanilla MAE	✓	75.4	78.5
Scale-MAE		75.3	79.6
Scale-MAE	✓	**78.1**	**80.7**

Table 3. Ablation results indicating the importance of GSDPE as determined by a KNN classification on RESISC-45 at a relative GSD of 50% and 100% of its native GSD. Using the GSDPE leads to better performance for both *Scale-MAE* and the Vanilla MAE.

Decoding Layers	KNN 50%	KNN 100%
1	76.0	78.4
2	77.9	80.4
3	**78.1**	**80.7**
4	77.5	80.0
8	77.7	78.9

Table 4. Ablation results indicating that fewer transformer layers in the decoding stage tend to work better for *Scale-MAE* as determined by a KNN classification on RESISC-45 at a relative GSD of 50% and 100% of its native GSD.

Mask Rate	KNN 50%	KNN 100%
70%	77.3	79.3
75%	**78.1**	**80.7**
80%	**78.1**	79.9

Table 5. Ablation results indicating that a 75% mask rate is optimal as determined by a KNN classification on RESISC-45 at a relative GSD of 50% and 100% of its native GSD.

4.2. Ablations

We ablate the key components of the *Scale-MAE* pretraining framework. For these experiments, we use a lightweight pretraining setting, where we pretrain for 300 epochs on FMoW (rather than 800) and use a ViT-Base encoder (rather than ViT-Large), and evaluate using a kNN evaluation on

Low Res	High Res	Combined	KNN 50%	KNN 100%
		✓	77.6	80.2
✓			72.9	74.3
	✓		77.2	80.3
✓	✓		**78.1**	**80.7**

Table 6. These ablation results indicate that reconstructing both the low resolution and high resolution components lead to robust performance. Note: when the high resolution component is reconstructed, the low-resolution residual is not used—the high resolution result is directly reconstructed. The "Combined" entry merges the low and high resolution results instead of treating them as separate losses. The evaluations are a kNN classification (k=20) on RESISC-45 at relative GSDs 50% and 100% of its native GSD.

RESISC-45 at 100% and 50% of its native GSD. The key contributions that we ablate are as follows: the GSD positional encoder in Table 3, in which we find that the GSD postional encoder benefits both *Scale-MAE* and Vanilla MAE across resolutions. In Table 4, we see that the number of transformer layers can be reduced from 8 to 3 compared to a Vanilla MAE, which results in a performance improvement. The standard masking rate of 75% still appears optimal for *Scale-MAE* according to the results in Table 5.

In Table 6 we ablate the necessity of the low and high resolution reconstructions. Specifically, we test reconstructing the low resolution image only, the high resolution image, and a combined image (rather than independent low/high reconstructions). In this case, when the high resolution component is reconstructed, we do not use the low-resolution residual, but rather, directly reconstruct the high resolution result. The "Combined" entry combines the low and high resolution results instead of treating them as separate learning objectives. The separate low/high resolution reconstructions obtain the best performance and robustness to changes in scale.

5. Discussion

In this section, we share observations about *Scale-MAE*, sketch our vision for future work, and discuss high-level questions about *Scale-MAE*.

Computational complexity. *Scale-MAE* requires a much smaller decoder than vanilla MAE—instead of a decoder depth of eight, *Scale-MAE* works well with a depth of three. In fact, with 322.9M vs 329.5M parameters using ViT-Large, *Scale-MAE* is smaller than vanilla MAE. However, GPU memory usage for equal batch sizes are higher for *Scale-MAE* since we reconstruct a higher resolution image in the *Scale-MAE* Decoder.

Multi-spectrality and modality. Electro-optical (EO) satellites, such as the ones comprising the datasets mentioned in this work, capture light at different wavelengths. Each wavelength has a different sensor, and each sensor can have a different resolution. *Scale-MAE* requires input tensors to be stacked to pass through the model. This means that we are unable to use *Scale-MAE* when the input image's bands are all of different GSDs. Additionally, synthetic aperture radar (SAR) imagery is another form of remote sensing where resolution varies across a single band. Extending *Scale-MAE* to work with different resolution bands and modalities is reserved for future work.

Can the *Scale-MAE* methodology be applied to other backbones? Methods such as ConvNeXt [41] provide competitive performance compared to Transformers. The core components of our work can be integrated, with additional work, into different architectures. The Laplacian Decoder in *Scale-MAE* can be engineered to ingest convolutional feature maps. Existing work on scale-aware CNNs can be extended to work with the Laplacian Decoder.

Evaluating on more remote sensing datasets. The field of remote sensing has had a renaissance in the last five years with the amount of available datasets. These can be generic, like Functional Map of the World, to highly specific, such as identifying illegal airstrips in Brazil [1, 8] or identifying illegal fishing vessels [45]. In fact, there are so many small, specific remote sensing datasets that entire review papers are written to enumerate them [57]. We chose to focus datasets with *properties* of remote sensing that are relevant to multi-scale representation learning.

6. Conclusion

Remote sensing imagery has accelerated the rate of scientific discovery in a broad set of disciplines. With increasingly precise methods to extract environmental indicators using computer vision methods, automated understanding of remotely sensed sources has become a mainstay in scientific literature. Remote sensing payloads are diverse and capture data at a wide range of resolutions, a feature heavily utilized by scientists. Current computer vision methods for remote sensing necessitate the training of a new model per input resolution. Not only is the training process expensive, but the overhead of curating a dataset at multiples scales makes this a daunting task.

We introduce *Scale-MAE*, a pretraining framework which introduces scale invariance into encoders that are used for a diverse set of downstream tasks. Our insights into scale-inclusive positional encodings and progressive multi-frequency feature extraction result in models that perform significantly better than state-of-the-art pretraining methods across (1) multiple scales and (2) many benchmarks.

Our goal is to take the extremely diverse and rich source

of information present in remote sensing imagery and make it simple to use with minimal training iterations required. With the introduction of *Scale-MAE*, we hope to further accelerate the rate at which scientific disciplines create impact.

Acknowledgements

Satellite imagery and derived images used in this paper in are from datasets which redistribute imagery from Google Earth, DigitalGlobe, and Copernicus Sentinel 2022 data. Professor Darrell's group was supported in part by funding from the Department of Defense as well as BAIR's industrial alliance programs.

References

[1] Manuela Andreoni, Blacki Migliozzi, Pablo Robles, and Denise Lu. The Illegal Airstrips Bringing Toxic Mining to Brazil's Indigenous Land. *The New York Times*, 2022. 8

[2] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020. 2, 5

[3] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-Aware Self-Supervised Learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10161–10170, Montreal, QC, Canada, Oct. 2021. IEEE. 3, 6, 7

[4] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 15

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 2

[6] P. Burt and E. Adelson. The Laplacian Pyramid as a Compact Image Code. *IEEE Transactions on Communications*, 31(4):532–540, Apr. 1983. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, Montreal, QC, Canada, Oct. 2021. IEEE. 6, 15

[8] Fen Chen, Ruilong Ren, Tim Van de Voorde, Wenbo Xu, Guiyun Zhou, and Yan Zhou. Fast Automatic Airport Detection in Remote Sensing Images Using Convolutional Neural Networks. *Remote Sensing*, 10(3):443, Mar. 2018. 8

[9] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 6, 15

[10] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 3

[11] 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, Oct. 2017. 6, 12

[12] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional Map of the World. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, Salt Lake City, UT, June 2018. IEEE. 5, 6, 12

[13] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery, Oct. 2022. 2, 3, 5, 6, 14, 15

[14] Dengxin Dai and Wen Yang. Satellite Image Classification via Two-Layer Sparse Coding With Biased Image Representation. *IEEE Geoscience and Remote Sensing Letters*, 8(1):173–176, Jan. 2011. 6, 12

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[16] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014. 2

[17] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010. 2

[18] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 3

[19] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008. 3

[20] Anthony Fuller, Koreen Millard, and James R. Green. SatViT: Pretraining Transformers for Earth Observation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 2

[21] Yuan Gao, Xiaojuan Sun, and Chao Liu. A general self-supervised framework for remote sensing image classification. *Remote Sensing*, 14(19):4824, 2022. 2, 3

[22] Golnaz Ghiasi and Charless C. Fowlkes. Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 519–534, Cham, 2016. Springer International Publishing. 3

[23] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org. 2

[24] Ritwik Gupta, Colorado Reed, Anja Rohrbach, and Trevor Darrell. Accelerating Ukraine Intelligence Analysis

with Computer Vision on Synthetic Aperture Radar Imagery. http://bair.berkeley.edu/blog/2022/03/21/ukraine-sar-maers/, Mar. 2022. 3

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, Dec. 2021. 2, 3, 15

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 15

[27] Yutong He, Dingjie Wang, Nicholas Lai, William Zhang, Chenlin Meng, Marshall Burke, David Lobell, and Stefano Ermon. Spatial-temporal super-resolution of satellite imagery via conditional pixel synthesis. *Advances in Neural Information Processing Systems*, 34:27903–27915, 2021. 3

[28] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing Eurosat: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 204–207, July 2018. 6, 12

[29] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020. 2

[30] X Hu, H Mu, X Zhang, Z Wang, T Tan, and J Meta-SR Sun. A magnification-arbitrary network for super-resolution. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA*, pages 15–20, 2019. 3

[31] Damian Ibañez, Ruben Fernandez-Beltran, Filiberto Pla, and Naoto Yokoya. Masked auto-encoding spectral-spatial transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 2, 3

[32] Pengpeng Ji, Shengye Yan, and Qingshan Liu. Region-Based Spatial Sampling for Image Classification. In *2013 Seventh International Conference on Image and Graphics*, pages 874–879, July 2013. 3

[33] Jan J. Koenderink. The structure of images. *Biological Cybernetics*, 50(5):363–370, Aug. 1984. 3

[34] Pawel Kowaleczko, Tomasz Tarasiewicz, Maciej Ziaja, Daniel Kostrzewa, Jakub Nalepa, Przemyslaw Rokita, and Michal Kawulok. Mus2: A benchmark for sentinel-2 multi-image super-resolution. *arXiv preprint arXiv:2210.02745*, 2022. 1, 3

[35] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178, June 2006. 3

[36] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 2

[37] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 3

[38] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 3

[39] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 21–37, Cham, 2016. Springer International Publishing. 3

[40] Yufei Liu, Xiaorun Li, Ziqiang Hua, Chaoqun Xia, and Liaoying Zhao. A band selection method with masked convolutional autoencoder for hyperspectral image. *IEEE Geoscience and Remote Sensing Letters*, 2022. 2, 3

[41] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 8

[42] Gabriel Machado, Edemir Ferreira, Keiller Nogueira, Hugo Oliveira, Matheus Brito, Pedro Henrique Targino Gama, and Jefersson Alex dos Santos. AiRound and CV-BrCT: Novel Multiview Datasets for Scene Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:488–503, 2021. 6, 12

[43] Malachy Moran, Kayla Woputz, Derrick Hee, Manuela Girotto, Paolo D'Odorico, Ritwik Gupta, Daniel Feldman, Puya Vahabi, Alberto Todeschini, and Colorado J. Reed. Snowpack Estimation in Key Mountainous Water Basins from Openly-Available, Multimodal Data Sources, Aug. 2022. 1

[44] Maxim Neumann, André Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *ArXiv*, abs/1911.06721, 2019. 3

[45] Fernando Paolo, Tsu-ting Tim Lin, Ritwik Gupta, Bryce Goodman, Nirav Patel, Daniel Kuster, David Kroodsma, and Jared Dunnmon. xView3-SAR: Detecting Dark Fishing Activity Using Synthetic Aperture Radar Imagery. In *Thirty-Sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, Oct. 2022. 8

[46] Xiaoman Qi, Panpan Zhu, Wang Yuebin, Liqiang Zhang, Junhuan Peng, Mengfan Wu, Jialong Chen, Xudong Zhao, Ning Zang, and P. Takis Mathiopoulos. MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:337–350, Nov. 2020. 6, 12

[47] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *Preprint*, 2018. 2

[48] A. Rosenfeld and M. Thurston. Edge and Curve Detection for Visual Scene Analysis. *IEEE Transactions on Computers*, C-20(5):562–569, May 1971. 3

[49] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition Workshops, pages 0–0, 2019. 3

[50] Beth Tellman, Jonathan A. Sullivan, and Colin S. Doyle. Global Flood Observation with Multiple Satellites. In *Global Drought and Flood*, chapter 5, pages 99–121. American Geophysical Union (AGU), 2021. 1

[51] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. SpaceNet: A Remote Sensing Dataset and Challenge Series, July 2019. 6, 7, 12

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3

[53] Qi Wang, Shaoteng Liu, Jocelyn Chanussot, and Xuelong Li. Scene Classification With Recurrent Attention of VHR Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):1155–1167, Feb. 2019. 6, 12

[54] Zhongrui Wang, Saumil Joshi, Sergey Savel'ev, Wenhao Song, Rivu Midya, Yunning Li, Mingyi Rao, Peng Yan, Shiva Asapu, Ye Zhuo, et al. Fully memristive neural networks for pattern classification with unsupervised learning. *Nature Electronics*, 1(2):137–145, 2018. 4

[55] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 6, 15

[56] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11209, pages 432–448. Springer International Publishing, Cham, 2018. 7

[57] Zhitong Xiong, Fahong Zhang, Yi Wang, Yilei Shi, and Xiao Xiang Zhu. EarthNets: Empowering AI in Earth Observation, Oct. 2022. 8

[58] Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. *arXiv preprint arXiv:2103.12716*, 2021. 3

[59] Shengye Yan, Xinxing Xu, Dong Xu, Stephen Lin, and Xuelong Li. Beyond Spatial Pyramids: A New Feature Extraction Framework with Dense Spatial Sampling for Image Classification. In *European Conference on Computer Vision 2012*, pages 473–487, Aug. 2012. 3

[60] Shengye Yan, Xinxing Xu, Dong Xu, Stephen Lin, and Xuelong Li. Image Classification With Densely Sampled Image Windows and Generalized Adaptive Multiple Kernel Learning. *IEEE Transactions on Cybernetics*, 45(3):381–390, Mar. 2015. 3

[61] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, 2019. 4

[62] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, pages 270–279, New York, NY, USA, Nov. 2010. Association for Computing Machinery. 6, 12

[63] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2

[64] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. PSANet: Point-wise Spatial Attention Network for Scene Parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018. 7

A. Datasets

In our experiments, we used a total of ten datasets (Table 7) for the tasks of land-use/land-cover classification and semantic segmentation. There are a large amount of remote sensing datasets in existence. Many remote sensing datasets fundamentally capture the same data with minor changes in location or distribution. We selected datasets with key, representative *properties*. These properties include (1) a diversity in the amount of kinds of classes/objects represented, (2) a large spectrum of ground sample distances from (ideally) known sensor configurations, and (3) pansharpened, othrorectified, and quality controlled imagery and labels. We capture these properties in Table 7.

A.1. Diversity in classes

For both pretraining and downstream evaluations, it is a desirable property to include as much geographic and class diversity as possible. In order to capture a wide amount of classes in remote sensing, it is necessary to include multiple localities and environments. This property serves as a proxy for the amount of unique "features" available in the dataset.

Dataset	Resolution (px)	GSD (m)	Number of Images	Number of Classes	Task Type
AiRound [42]	500	0.3 - 4800	11,753	11	C
CV-BrCT [42]	500	0.3 - 4800	24,000	9	C
EuroSAT [28]	64	10	27,000	10	C
MLRSNet [46]	256	0.1 - 10	109,161	46	C
Optimal-31 [53]	256	0.5 - 8	1,860	31	C
RESISC-45 [11]	256	0.2 - 30	31,500	45	C
UC Merced [62]	256	0.3	2,100	21	C
WHU-RS19 [14]	256	0.5	950	19	C
fMoW [12]	Various	0.3	1,047,691	62	C
SpaceNet v1 [51]	Various	0.5	6,940	2	SS

Table 7. Statistics of all datasets used in our experiments. Task types are classification (C) and semantic segmentation (SS).

A.2. Spectrum of GSDs

Scale-MAE is built to be invariant to the input absolute scale of the dataset. Many datasets are collected from a single sensor and processed in a uniform fashion. To validate that our method works with many resolutions, we included datasets which are collected from a variety of sensors but then processed in a uniform fashion. This excludes differences in processing as a factor affecting our experiments and narrowly targets resolution instead.

A.3. Quality control

It is hard to assess the quality of remote sensing datasets without manually verifying a majority of instances of the data. We mandated that images used are pansharpened (and therefore the highest resolution possible to extract from the sensor), orthorectified (and therefore well-aligned with the geodetic ellispoid), and projected to the same coordinate reference system. This eliminates large differences in sensor-to-image processing.

B. Laplacian and Upsampling Block Architectures

Figure 7 illustrates the architecture of Laplacian and Upsampling block architectures described below.

B.1. Laplacian Block

Laplacian Blocks are used to reconstruct the target at a specific resolution and frequency. A Laplacian Block consists of a chain of Feature Mapping Block, which distills information at a specific frequency, followed by one final Reconstruction Block, which generates the final output. A Feature Mapping Block consists of a 3x3 depth-wise convolution layer with GELU activation, followed by 1x1 convolution. A Reconstruction Block consists of a 4x4 transpose convolution layer followed by a

		$k = 20$		$k = 100$		$k = 5$	
Dataset	**Resolution**	***Scale-MAE***	**SatMAE**	***Scale-MAE***	**SatMAE**	***Scale-MAE***	**SatMAE**
AiRound	16	**0.384**	0.375	**0.371**	0.367	**0.363**	0.355
	32	**0.577**	0.510	**0.551**	0.491	**0.562**	0.492
	64	**0.684**	0.607	**0.638**	0.579	**0.693**	0.604
	128	**0.719**	0.650	**0.668**	0.600	**0.724**	0.660
	256	**0.714**	0.662	**0.667**	0.621	**0.717**	0.663
	496	**0.672**	0.664	**0.613**	0.613	**0.685**	0.669
CV-BrCT	16	**0.496**	0.478	0.442	**0.443**	**0.509**	0.475
	32	**0.660**	0.615	**0.588**	0.560	**0.695**	0.644
	64	**0.748**	0.701	**0.679**	0.635	**0.790**	0.727
	128	**0.770**	0.725	**0.716**	0.662	**0.806**	0.758
	256	**0.760**	0.725	**0.697**	0.666	**0.798**	0.770
	496	**0.743**	0.727	**0.671**	0.657	**0.790**	0.771
EuroSAT	16	0.721	**0.727**	0.676	**0.695**	0.723	**0.729**
	32	**0.908**	0.876	**0.875**	0.854	**0.909**	0.871
	64	**0.958**	0.931	**0.936**	0.913	**0.963**	0.934
MLRSNet	16	**0.529**	0.491	**0.488**	0.461	**0.517**	0.479
	32	**0.778**	0.677	**0.735**	0.625	**0.786**	0.684
	64	**0.891**	0.815	**0.841**	0.754	**0.909**	0.839
	128	**0.925**	0.875	**0.881**	0.814	**0.945**	0.899
	256	**0.918**	0.892	**0.861**	0.840	**0.940**	0.913
OPTIMAL-31	16	0.309	**0.322**	0.287	**0.298**	0.285	**0.319**
	32	**0.617**	0.500	**0.569**	0.508	**0.596**	0.519
	64	**0.777**	0.609	**0.705**	0.561	**0.793**	0.646
	128	**0.787**	0.670	**0.713**	0.646	**0.809**	0.694
	256	**0.710**	0.681	**0.644**	0.638	**0.726**	0.731
RESISC	16	**0.361**	0.347	**0.360**	0.327	**0.336**	0.323
	32	**0.618**	0.527	**0.589**	0.505	**0.603**	0.508
	64	**0.793**	0.667	**0.749**	0.631	**0.803**	0.667
	128	**0.858**	0.748	**0.803**	0.699	**0.875**	0.762
	256	**0.821**	0.758	**0.751**	0.708	**0.846**	0.771
UC Merced	16	0.469	**0.472**	**0.380**	0.370	**0.509**	0.488
	32	**0.807**	0.670	**0.686**	0.535	**0.825**	0.682
	64	**0.858**	0.795	**0.738**	0.729	**0.910**	0.842
	128	**0.825**	0.788	0.736	**0.738**	**0.908**	0.847
	256	0.778	**0.802**	0.691	**0.757**	0.837	**0.842**
WHU-RS19	16	0.395	**0.445**	0.345	**0.380**	0.435	**0.490**
	32	0.700	**0.710**	0.575	**0.675**	**0.730**	0.690
	64	**0.920**	0.805	**0.780**	0.730	**0.935**	0.840
	128	**0.960**	0.910	0.880	**0.890**	**0.970**	0.895
	256	**0.955**	0.940	0.900	**0.925**	**0.965**	0.945

Table 8. ***Scale-MAE* outperforms SatMAE on kNN classification across a variety of k, across a variety of resolutions.** kNN Classification results for *Scale-MAE* and SatMAE across a variety of k. Resolution is reported in pixels.

3x3 depth-wise convolution layer, a 1x1 convolution layer, and a 2x2 transpose convolution layer. In our experiments, we have two Feature Mapping Blocks per Laplacian Block.

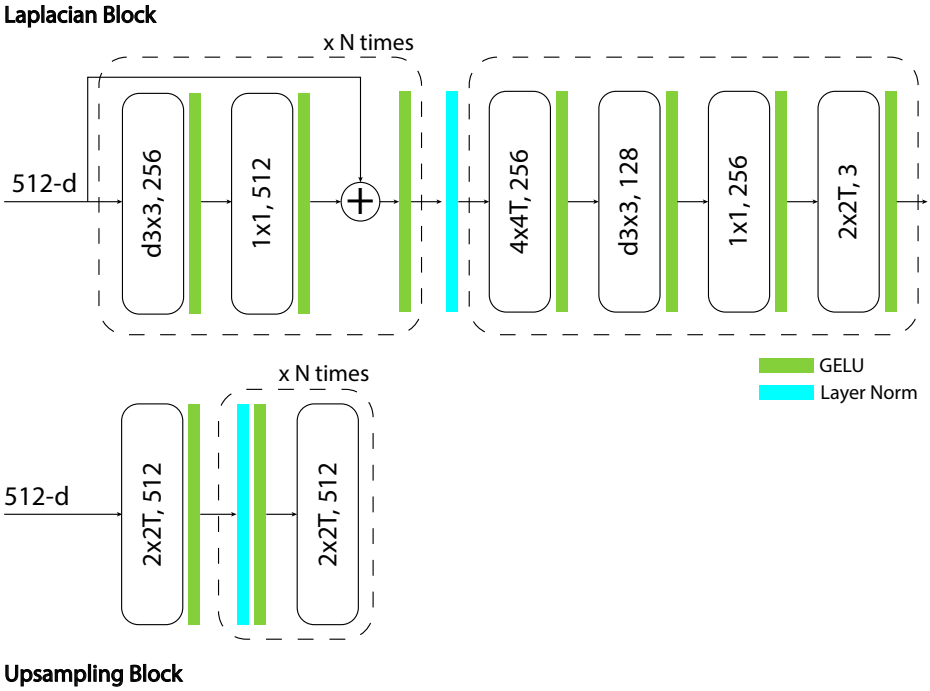

Figure 7. (top) The Laplacian Block (LB) is a fully convolutional architecture consists of a chain of Feature Mapping Block followed by one final Reconstruction Block. (bottom) The UpSampling Block (UB) consists of a series of transpose convolution layers separated by LayerNorm and GELU activation.

B.2. Upsampling Block

Upsampling Blocks are used to upsample the feature map to a higher resolution. It consists of a series of 2x2 transpose convolution layers with LayerNorm and GELU activation between them. The number of such transposed convolution layers are a function of the output and input resolution. This is a progressive process in which we repetitively upsample the feature map by a factor of 2 until we reach the desired target resolution. Figure 7 illustrates the architecture of these two blocks.

C. Evaluation Details

As discussed in the main experimental section, we investigated the quality of representations learned from *Scale-MAE* pretraining through a set of experiments that explore their robustness to scale as well as their transfer performance to additional tasks. We provide more information and details on these evaluations here. In order to compare with SatMAE [13], for our main experiments, we pretrained *Scale-MAE* with a ViT-Large model using the Functional Map of the World (FMoW) RGB training set, which consists of 363.6k images of varying image resolution and GSD. The initial higher resolution image I_{hr} is taken as a random 448px^2 crop of the input image, and the input image I is then a downsampled 224px^2 from I_{hr}. The low frequency groundtruth is obtained by downscaling I_{hr} to 14px^2 and then upscaling to 224px^2, while the high frequency groundtruth is obtained by downscaling I_{hr} to 56px^2 and then upscaling to 448px^2 and subtracting this image from I_{hr}. This is a common method for band pass filtering used in several super resolution works, where a high to low to high resolution interpolation is used to obtain only low frequency results, and then high frequency results are obtained by subtracting the low frequency image.

As further discussed in the main experimental section, we evaluate the quality of representations from *Scale-MAE* by freezing the encoder and performing a nonparametric k-nearest-neighbor (kNN) classification with eight different remote sensing imagery classification datasets with different GSDs, none of which were encountered during pretraining. Results are in Table 8. The kNN classifier operates by encoding all train and validation instances, where each embedded instance in the validation set computes the cosine distance with each embedded instance in the training set, where the instance is classified correctly if the majority of its k-nearest-neighbors are in the same class as the validation instance. The justification for a kNN classifier evaluation is that a strong pretrained network will output semantically grouped representation for unseen data of the

same class. This evaluation for the quality of representations occurs in other notable works [7, 9, 55].

D. Linear Classification Results

We perform Linear classification on the RESISC-45 and FMoW-RGB datasets. We fine-tune for 50 epochs using the same hyperparameter settings as SatMAE [13]: a base learning rate of 5×10^{-3}, a weight decay of 5×10^{-3}. We do not use temporal data for classification. For RESISC-45, we fine-tune for 100 epochs with a base learning rate of 4×10^{-3}, a weight decay of 5×10^{-3}, and a global batch size of 256 across 2 GPUs. The learning rate on the backbone is multiplied by a factor of 0.1. We use RandomResizedCrop for augmentation. We train on 224x224 images and evaluate 256x256 images because we found evaluating at a higher scale improves the performance of all models. We report both the performance of end-to-end fine-tuning and linear probing with a frozen backbone. The linear probing setup was the same as finetuning except the learning rate was 0.1. The results are shown in Table 9 and Table 10.

Model	Frozen/Finetune
Scale-MAE	**89.6/95.7**
SatMAE [13]	88.3/94.8
MAE [25]	88.9/93.3

Table 9. **Transfer classification results on RESISC-45.** Frozen indicates a linear probe and finetune is a full end-to-end finetuning of the entire model.

Model	Backbone	Top-1/Top-5
Scale-MAE	ViT-Large	**77.9/94.3**
SatMAE † [13]	ViT-Large	72.4/91.9
MAE [25]	ViT-Large	68.4/90.3
SatMAE ∗ [13]	ViT-Large	77.8/-
GASSL [4]	ResNet-50	71.55/-
MoCo-V2 [26]	ResNet-50	64.34/-

Table 10. **Linear probe and finetune classification results on FMoW-RGB.** †: We reproduce the SatMAE by taking the publicly available codebase and pretrain on FMoW dataset for 800 epochs. The results differ from their reported results, but are evaluated consistently with ours. * Reports the results from the SatMAE paper [13].

E. Visualization of SpaceNet Segmentation

Figure 8 shows an additional set of segmentation examples comparing *Scale-MAE* and vanilla MAE pre-trained on FMoW and finetuned on SpaceNet v1. The left, center, right columns are ground truth labels, *Scale-MAE* and vanilla MAE respectively. The top row shows a 0.3m GSD image and the bottom row shows a 3.0m GSD image. As shown in the figure, *Scale-MAE* performs better at both higher and lower GSDs.

F. Glossary

F.1. Ground sample distance

Ground sample distance (GSD) is the distance between the center of one pixel to the center of an adjacent pixel in a remote sensing image. GSD is a function of sensor parameters (such as its dimensions and focal length), image parameters (the target dimensions of the formed image), and the geometry of the sensor with respect to the object being imaged on the Earth. Remote sensing platforms frequently have multiple sensors to capture different wavelengths of light. Each of these sensors have varying parameters, resulting in different GSDs for an image of the same area. Additionally, the ground is not a uniform surface with changes in elevation common across the swath of the sensor. In total, a remote sensing platform has a sense of absolute scale that varies along two dimensions: (1) spectrally depending on the sensor used to capture light, and (2) spatially depending on surface elevation.

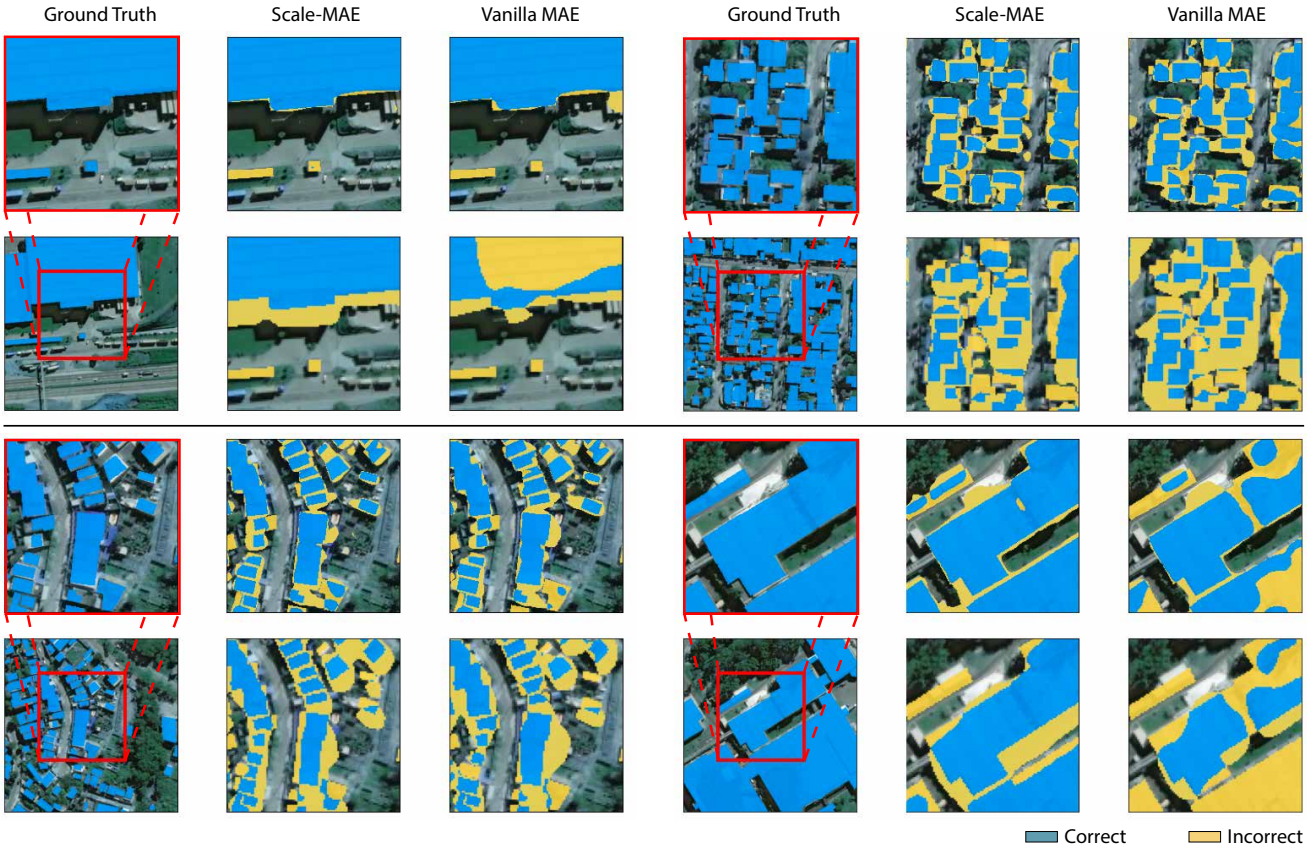

Figure 8. Visualization of Segmentation Results on SpaceNet. The left, center, right columns are ground truth labels, *Scale-MAE* and vanilla MAE, respectively. The top row shows a 0.3m GSD image and the bottom row shows a 3.0m GSD image. As shown in the figure, *Scale-MAE* performs better at both higher and lower GSDs.