Self-Supervised Learning for Remote Sensing

Jeegn Dani
Purdue University
West Lafayette, IN
jdani@purdue.edu

Abstract

Self-supervised learning (SSL) has emerged as a transformative approach for satellite imagery analysis, enabling models to learn from vast unlabeled datasets critical for addressing global challenges. This study studies and evaluates state-of-the-art SSL methods like SeCo, SatMAE, and ScaleMAE—on diverse land cover classification tasks using datasets such as EuroSAT, Optimal-31, RESISC-45, and the self created GlobalPatches. GlobalPatches addresses the lack of geographic diversity in benchmarks by incorporating Sentinel-2 imagery and DynamicWorld labels across underrepresented regions. Results highlight the superior performance of MAE-based models in multiscale tasks but underscore the need for lightweight SSL approaches and globally inclusive benchmarks. The paper also identifies opportunities in hyperspectral band selection, multimodal fusion, and impact-driven benchmark design to align research with real-world needs.

1. Introduction

Self-supervised Learning (SSL) has emerged as a transformative paradigm in machine learning that focuses on learning useful representations from unlabeled data. Rather than relying on manually annotated datasets, which can be expensive and time-consuming to create, SSL uses the intrinsic structure of the data itself to create training signals during the pretraining phase. Recently, SSL has gained significant attention in computer vision and has achieved major breakthroughs by already outperforming supervised pretraining on many vision-based tasks [5, 6, 7]. During this pretraining phase the model is exposed to a large amount of unlabeled data, where it learns general patterns and structures of the data. These learned representations can then be fine-tuned on labled datasets for specific downstream tasks, which is called transfer learning. For instance, models pretrained on large-scale datasets like ImageNet using SSL methods can be fine-tuned for tasks such as object detection, image segmentation, flow estimation and many more [19, 20].

This reusability not only accelerates the deployment of machine learning solutions across diverse applications but also drastically reduces the compute cost, as finetuning a pretrained model saves quite a bit of compute compared to training a model from scratch. This make SSL a highly efficient approach in resource constrained setting with an abundance of unlabeled data.

Hundreds of remote sensing satellites continuously monitor the Earth's surface, creating petabyte-scale datasets. Machine learning for these satellite datasets (SatML) can be used to address pressing planetary-scale challenges including climate change [45], poverty [22], food insecurity [25], biodiversity loss [35], and many other goals. Satellite imagery often covers large geographical areas over extended periods, resulting in vast amounts of unlabeled data. The upcoming NASA-ISRO Synthetic Aperture Radar (NISAR) mission will generate up to 85 TB/day [2] and in comparison, the Common Crawl database use to train GPT-3 was 45 TB [4]. So manually labelling these datasets is impractical and costly. This is where SSL offers a natural solution. By utilizing SSL we can learn meaningful representations from the unlabeled satellite data which can be used for a variety of downstream tasks, such as land cover classification, flood and disaster detection, soil moisture estimation, wildlife habitat mapping and many more.

On the surface it might seem straightforward to apply deep learning algorithms developed for natural images directly to satellite imagery. However, satellite data comes with unique challenges and characteristics that make this "lift-and-shift" approach insufficient and hence satellite data has to be considered as distinct modality [32]. Satellite data is stored similar to natural images and videos with height, width, and channel dimensions. With temporal information included by stacking spatially aligned images along a fourth dimension. What makes it unique is the logarithmic spatial scale, requiring models to recognize patterns from fine details like individual trees or animals to large-scale phenomena like forest coverage and wildfires. Similarly, temporal scales in satellite data contain patterns that can manifest over hours (earthquakes, tsunamis), weeks (urban development, flooding), seasons (crop cultivation, snow cover), years (glacial retreat, deforestation), and decades (sea level rise, erosion) [32]. In addition to this, satellite data often includes multispectral channels covering the visible, near-infrared, and shortwave infrared spectrum, combined with with active sensors such as synthetic aperture radar (SAR), LiDAR, and radar. For example Sentinel-2, a mission under the European Space Agency Copernicus Program, provides 13-channel optical images. These additional complexities warrants specialized methods and models to handle satellite data. And in this project I investigated SSL methods tailored for SatML and tested them on geographically diverse landcover mapping tasks.

Historically, progress in machine learning has been closely tied to the availability of benchmark datasets and associated challenges. These benchmarks have not only driven innovation but have also provided a standardized way to measure and compare progress. A prime example is ImageNet, The annual ImageNet Large Scale Visual Recognition Challenge gave us breakthrough architectures like AlexNet, ResNet and Vision Transformers. However, while working over this project I realized that in the satellite imagery domain, there are no such universally adopted benchmark datasets that currently exist. There are several large-scale annotated image datasets like Functional Map of the World (1,047,691 images)[9], BigEarthNet (590,326 images) [36], MLRSNet (109,161 images)[27] but they are still more densely sampled in the global North (Figure 1). This makes a compelling argument for the urgent need for a geographically inclusive benchmark dataset that include underrepresented regions and also accounts for varying environmental conditions.

It is also important that benchmarks in satellite imagery are designed not just to improve performance metrics but to have a real-world impact on a variety of downstream tasks. These tasks include land cover classification, disaster detection, biodiversity monitoring, and climate modeling, all of which are directly linked to addressing pressing societal and environmental challenges. By ensuring that benchmark challenges are aligned with practical applications, we can drive innovation that goes beyond theoretical improvements and contributes meaningfully to solving real-world problems.

2. Related Works

SSL models learn meaningful representations from the large amounts of unlabeled data. This is achieved by designing pretext tasks (or proxy loss) that help the model to learn what we care about the data and solve the tasks. A variety of pretext tasks have been proposed for natural images, such as predicting relative position of patches [12], solving jigsaw puzzles [26], predicting rotations [16] or colorization [47]. More recently Contrastive Learning and Masked Auto Encoders have dominated SSL for vision based tasks [6, 14, 17, 19, 20].

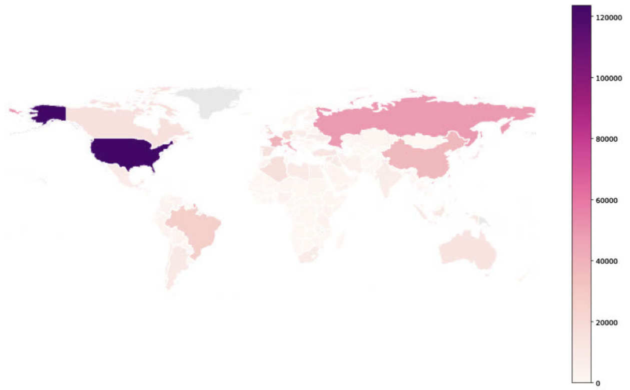

Figure 1. Geographic distribution of functional Map of the World (fMoW) Sentinel images by country [9]

2.1. Contrastive Learning

Intuitively, contrastive learning methods learn representations by pulling similar image pairs (positives) closer and pushing apart the dissimilar ones (negatives) in the feature space. Since there are no labels for the data, contrastive learning assumes that each data point is unique and defines its own class. To generate training data, we create positive pairs by applying random augmentations to the same image, such as cropping, flipping, or color jittering. These augmentations simulate different views of the same data, and the model is tasked with learning that they represent the same thing. Meanwhile, negative pairs are created by sampling other images in the dataset, which are assumed to represent different classes. As illustrated in Figure 2, after creating positive and negative pairs we pass them through an encoder network $f(\cdot)$ to get the latent representation h_i, h_j in the embedding space. This is followed by a small neural network projection head $g(\cdot)$ that maps h_i, h_j to z_i, z_j which lie in the space where a contrastive loss is applied. A common choice for this contrastive objective is the InfoNCE loss [39]:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]}\exp(z_i \cdot z_k/\tau)} \tag{1}$$

where τ is a temperature hyper-parameter scaling the the distribution of distance. Controlling how strictly the model must treat the positive pairs as similar and negatives as dissimilar.

2.2. Seasonal Contrast

Seasonal Contrast (SeCo) [24] is a SSL method based on Contrastive Learning tailored for Satellite imagery. It

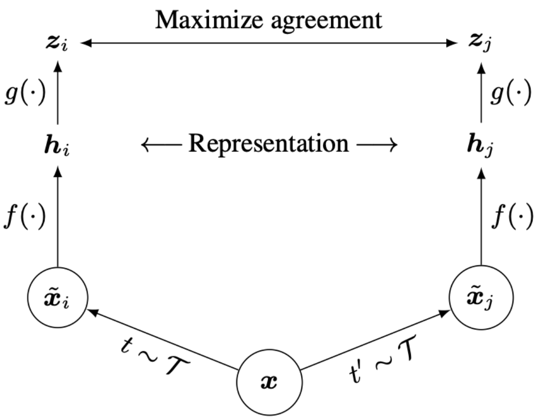

Figure 2. Contrastive Learning framework. Where t and t' are sampled from the same family of augmentations and applied to the data example. The base encoder $f(\cdot)$ and a projection head $g(\cdot)$ are trained trained to maximize agreement using the InfoNCE loss (Eqn 1). Post training, we throw away the projection head $g(\cdot)$ and use the encoder $f(\cdot)$ to get meaningful representations for the downstream tasks [6]

leverages the seasonal variability in satellite images of the same geographic location to create positive pairs rather than solely relying on artificial augmentations. As encouraging the representation to be invariant to seasonal changes is a strong inductive bias for downstream tasks like land-cover classification and building detection. However, it might not be suitable for downstream tasks like change detection or deforestation tracking where seasonal variations are important. So, in order to get the shared representation contain both time-varying and invariant features, this method uses the idea of having multiple embedding subspaces [44]. So instead of having a common embedding space that is invariant to seasonal changes, the common representation is projected into several embedding sub-spaces which are variant or invariant to time. And the contrastive loss is a weighted sum of the InfoNCE loss across all the embedded sub-spaces.

As illustrated in Fig 3, for each geographic location, we obtain 3 images at different times $x^{t_0}, x^{t_1}, x^{t_2}$. We let $q = x^{t_0}$, $k_0 = \mathcal{T}(x^{t_1})$, $k_1 = x^{t_2}$ and $k_2 = \mathcal{T}(x^{t_0})$. Where $\mathcal{T}$ is a set of commonly used artificial augmentations, such as random cropping, color jittering, and random flipping. This results in creating 3 views where the first view k_0 contains both seasonal and artificial transformations, k_1 only has seasonal transformation and k_2 containing only artificial transformations. Thereafter these images are encoded into their latent representations $v^q, v^{k_0}, v^{k_1}, v^{k_2}$ in a common embedding space $\mathcal{V}$. Next, these intermediate representaions are finally projected into 3 different subspaces $\mathcal{Z}_0, \mathcal{Z}_1, \mathcal{Z}_2$ by the projection heads h_0, h_1, h_2. Subspace $\mathcal{Z}_0$ is designed to pull together all z_0^i of the same instance. So this means that $(z_0^q, z_0^{k_0})$ form the positive pair and the negative pairs are all the embeddings of different locations. This makes $\mathcal{Z}_0$ invariant to all augmentations. For $\mathcal{Z}_1$, $(z_1^q, z_1^{k_1})$ are the positive pairs and the negative pairs consists of all other location embeddings plus $(z_1^{k_0}, z_1^{k_2})$ making this embedding sub-space invariant to seasonal augmentations but variant to to artificial augmentations. Finally $\mathcal{Z}_2$ has $(z_2^q, z_1^{k_2})$ as the positive pairs and the negative pairs consists of all other location embeddings plus $(z_2^{k_0}, z_1^{k_1})$ making this embedding sub-space invariant to artificial augmentations but variant to to seasonal augmentations. The final learning objective is computed as the sum of contrastive losses across all embedding subspaces, encouraging the encoder f to preserve both time-invariant and time-varying features in the general representation space $\mathcal{V}$, which can be used for transfer learning for all kinds of downstream tasks.

The authors of this paper used a ReseNet [18] as the feature extractor $f(\cdot)$ and a 2-layer MLP with a ReLU activation as their projection head $g(\cdot)$ for each embedding sub-space. And additionally, they only retained the RGB channels from the 13 different channels available for the Sentinel-2 images. SeCo outperforms ImageNet supervised pre-training, and traditional contrastive learning self-supervised training using MoCo-v2 [7] in Land-Cover Classification tasks on satellite imagery datasets like BigEarthNet [36] and EuroSAT [21] in both linear probing and fine-tuning.

2.3. Masked Auto Encoder

The contrastive Learning methods we saw above, strongly rely on the quality of data augmentation in order to achieve good positive pairs. However, recently another conceptually different SSL method - Masked Auto Encoder (MAE) has gained popularity [10, 14, 30, 43]. MAE proposes the pretext task of reconstruction of masked patches of the input, inspired by the highly successful pre-training methods in NLP - BERT [11] and GPT [4, 28, 29]. MAE's use Vision Tranformers (ViT) [13] as the backbone network for encoding and reconstructing the masked images, this is different from the CNN based backbones used in contrastive learning SSL methods. The MAE architecture breaks down an image $I \in \mathbb{R}^{C \times H \times W}$, where H, W are the image height and width, and C is the number of channels, into a sequence of fixed-sized, non overlapping patches, $S \in \mathbb{R}^{L \times P^2 C}$ where P is the height and width of the patch, and $L = (H/P) \cdot (W/P)$ is the number of patches. Each patch is flattened into a 1D vector and passed through a linear projection $f_p : \mathbb{R}^{P^2 C} \mapsto \mathbb{R}^D$ that creates a sequence $S' \in \mathbb{R}^{L \times D}$ of embedded patch ”tokens”. These patch embeddings encode the spatial and pixel-level information of the patch into a feature vector that the encoder (ViT) architecture can process. Now, in order to learn general representations about the image a fraction p_m of the L tokens

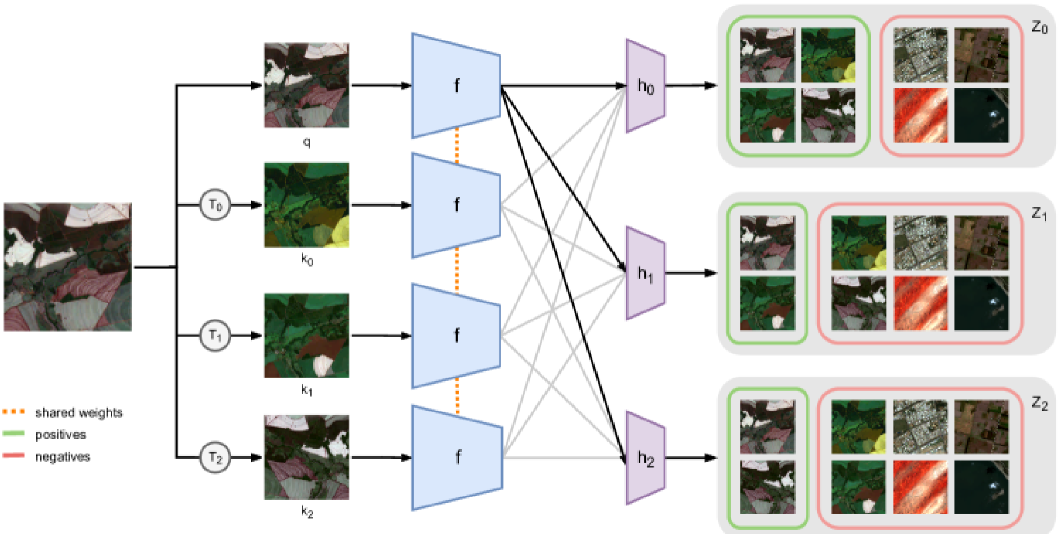

Figure 3. **Seasonal Contrast method**. An image query (q) is augmented with temporal (k_0, k_1) and synthetic (k_0, k_2) transformations $\mathcal{T}$. Image embeddings produced by the encoder f are projected into three different sub-spaces by heads h_0, h_1, h_2. Green boxes represent positive pairs while red boxes represent negative pairs (i.e. including images from other locations). Sub-space $\mathcal{Z}_0$ is invariant to all transformations, thus all augmented images belong to the same class as the query. $\mathcal{Z}_1$ is invariant to seasonal augmentations, while $\mathcal{Z}_2$ is invariant to synthetic augmentations. [24]

are masked and only the remaining $(1 - p_m)L$ visible tokens are fed to the encoder. Since the ViT architecture is permutation-invariant, we concat the patch tokens with positional embeddings (Eqn 2)[40], which indicate the position of each patch within the original image, ensuring that the model is aware of spatial relationships.

$$\begin{aligned} \texttt{Encode}(k, 2i) &= \sin\frac{k}{\Omega^{\frac{2i}{d}}} \\ \texttt{Encode}(k, 2i+1) &= \cos\frac{k}{\Omega^{\frac{2i}{d}}} \end{aligned} \tag{2}$$

Here k is defined as the index of the patch along the x or y axes, and i is the index of feature dimension in the encoding, d is the number of possible patch positions, and Ω is a large constant (normally set to 10000).

After the encoding phase, we have an assymetric decoder that is also a series of transformer blocks that operates on all L tokens and their positional encodings, with $(1 - p_m)L$ encoded outputs from the encoder placed in their original sequence position and the remaining $p_m L$ masked patches are represented by a learnable masked token. From this the decoder gives a reconstructed image $\hat{I} \in \mathbb{R}^{C \times H \times W}$. The reconstruction loss is computed as the mean squared error (MSE) between the predicted and original pixel values of the masked patches. By focusing the loss only on masked regions, MAE forces the encoder to capture contextual information from visible patches, ensuring the learned representations are robust and semantically meaningful.

2.4. SatMAE

SatMAE [10] is a SSL technique that builds up on MAE framework to address the unique challenges of satellite data. Namely temporal sequences at irregular sampling intervals and multispectral bands.

Handling temporal dimension. When dealing with satellite data, we have input tensors $I_T \in \mathbb{R}^{T \times C \times H \times W}$, where T denotes the number of images of the given geographic location in a temporal sequence. This is very common to video data. The key difference however, is the that the frames in the video data are uniformly spaced but temporal satellite imagery rarely has images taken at regular intervals and the sampling frequency drastically varies over years and across different regions. To address the temporal dimension, SatMAE divides I_T to patches $S_T \in \mathbb{R}^{L_T \times P^2 C}$, where $L_T = (H/P) \cdot (W/P) \cdot T$. Then they use the same patch embedding as MAE to get the embedded sequence of tokens $S'_T \in \mathbb{R}^{L_T \times D}$. To retain the temporal information, SatMAE introduces temporal encoding alongside positional encodings for each patch. Keeping in mind the seasonal and diurnal patterns that required to be captured for the downstream tasks, the temporal encoding inlcudes the month, year and hour of the image as specified in Eqn 3

$$t_{k,i} = [\texttt{Encode}(k_{\text{yr}}, i), \texttt{Encode}(k_{\text{mth}}, i), \texttt{Encode}(k_{\text{hr}}, i)] \tag{3}$$

The final embedding is generated by concatenating these temporal and positional (Eqn 2) encodings and fed to the Encoder (ViT) as illustrated in Figure 4. SatMAE also ap-

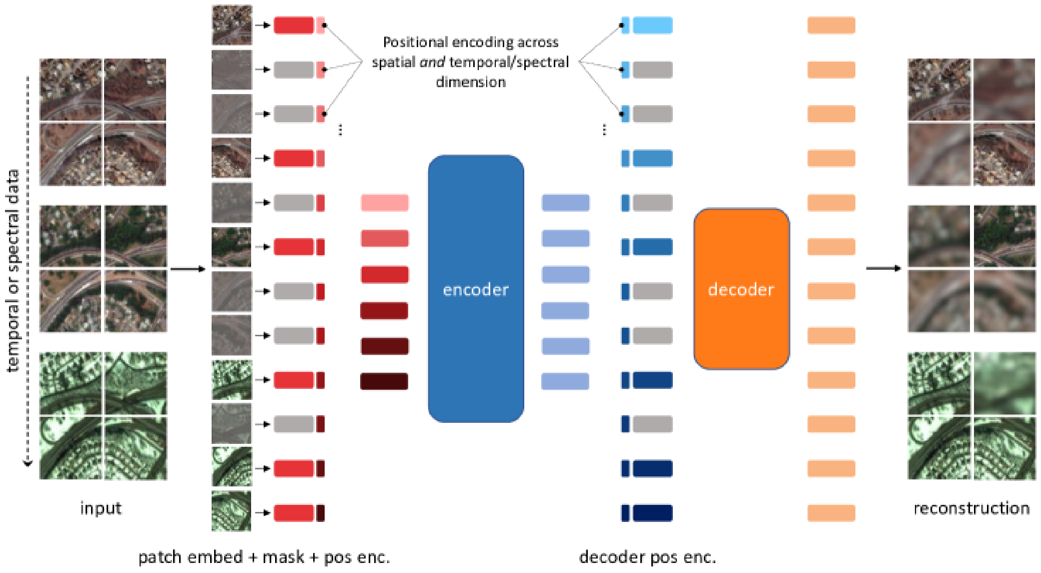
Positional encoding across spatial *and* temporal/spectral dimension
temporal or spectral data
input
patch embed + mask + pos enc.
encoder
decoder pos enc.
decoder
reconstruction

Figure 4. The MAE architecture breaks down an image into a set of patches which are embedded as vectors using a linear projection and concatenated with positional, temporal and spectral encoding according to Eqns 2, 3, 4. Then a fraction p_m of the patches are masked and the unmasked vectors are passed through an encoder. The latent representations of the unmasked patches is then joined by a learned mask token which are together fed into the decoder which reconstructs the original image per-pixel [10]

plies the an Independent masking strategy across the temporal dimension, so the patches are masked independently per each frame keeping the ratio p_m of maksed regions per image. In VideoMAE [37], it was found that independent masking leads to simplifying the reconstruction task as the mode can ”cheat” and exploit spatial redundancy across the temporal dimension. However, this kind of ”cheating” is less feasible in temporal satellite imagery, given the strong impact of seasonal variation and changing human activity over periods of time and much larger and inconsistent time deltas between temporally consecutive images.

Handling Multi-spectral Channels. Satellite imagery contains multiple spectral channels like near-infrared, shortwave infrared and ultraviolet. Rather than treating all spectral bands uniformly, SatMAE groups channels based on their spectral characteristics. For example, bands with similar wavelenghts or resolutions are grouped together. So now our input tensor I is sliced according to G spectral groups $g_1, g_2, \ldots, g_G$ such that thay all add up to C which is the total number of channels. Hence, we get G images $I_1, I_2, \ldots, I_G$, where $I_j \in \mathbb{R}^{g_j \times H \times W}$ and we use a separate patch embedding $f_{p_j} : \mathbb{R}^{P^2 g_j} \mapsto \mathbb{R}^D$ for each spectral group. All the spectral group embedded token are concatenated with position encodings given by Eqn 2 and spectral encoding given by Eqn 4 as illustrated in Figure 4:

$$g_{k_g,i} = \texttt{Encode}(k_g, i) \tag{4}$$

SatMAE outperforms outperforms supervised pretraining, SeCo and MAE on land-cover classification tasks on datasets like EuroSAT, Functional Map of the World (RGB and multispectral) and NAIP for both finetuning and linear probing tasks. It also matched SeCo's accuracy in the BigEarthNet multilabel classification task.

2.5. ScaleMAE

ScaleMAE [30] is a state-of-the-art self-supervised learning (SSL) technique tailored specifically for satellite imagery. It modifies the positional embeddings during the encoding stage and adapts the decoding task to better align with the unique characteristics of remotely sensed data. The downstream tasks for large pretrained satellite imagery models often involve imagery captured at a variety of spatial scales. Benchmark datasets such as EuroSAT [21], AiRound [23], and CV-BrCT [23] are commonly used for land cover classification tasks, which involve large-scale object detection of features such as forests, built areas, water bodies, and hilly terrains. Conversely, datasets like MLRSNet [27], Optimal-31 [41], and RESISC-45 [8] focus on classifying smaller-scale objects such as schools, airports, thermal power stations, and stadiums. Capturing and leveraging scale-specific information in satellite imagery is crucial for accurate modeling, and ScaleMAE explicitly learns relationships between data across different scales during the pretraining process. In order to do this, ScaelMAE introduces several key modifications to the standard MAE architecture, illustrated in Figure 5

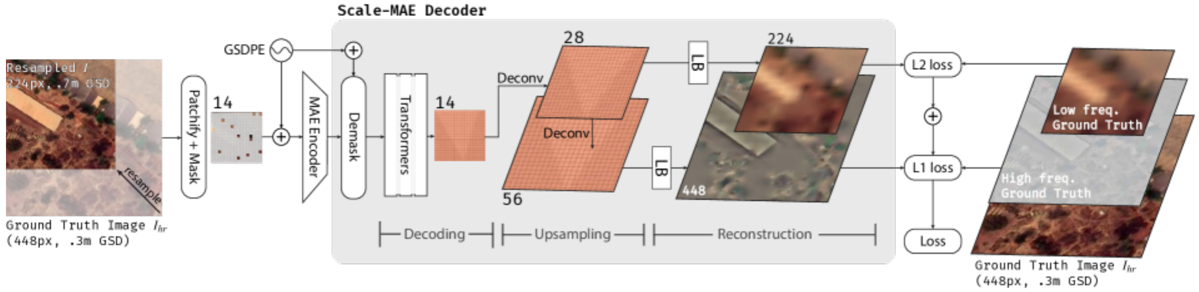

Figure 5. **ScaleMAE**: The input image is masked and decomposed to patches along with the Ground Sample Distance Positional Embedding (GSPDE), which scales the positional embeddings to the area of ground covered by the image. This is then fed to the ViT encoder like in the classical MAE architecture. The triphased ScaleMAE takes the encder output and decodes it using a reduced Transformer block. Then comes the upsampling using deconvolution blocks to create two feature maps for the low and high frequency reconstructions. This is then followed by the Laplacian Block (LB) which gives the reconstructed high and low frequency images

Super-Resolution Reconstruction. Rather than reconstructing the input image directly, ScaleMAE reconstructs both high frequency and low frequency components of the input image I inspired by the Laplacian pyramid-based super-resolution methods [46]. Super-resolution has proven effective in improving accuracy within remote sensing images due to the extremely small size of objects within the image [34]. ScaleMAE, by reconstructing both higher and lower frequency images, learns fine-grained details such as roads and rooftops from the high frequency reconstructions and borad contextual features like forests and urban development through low frequency reconstructions. In order to get the high frequency component I_{hf} from the ground truth input image I, I is donwsampled to an intermediate resolution, upsampled to back to the orginal resolution and then subtracted from I. For the low frequency component I_{lf}, I is simply interpolated to a much lower resolution and then upsampling it to the same resolution as I.

GSD Positional Encoding. Ground Sample Distance (GSD) refers to the physical distance between the two adjacent pixels in an image ad is a key metric for understanding the spatial resolution of satellite imagery. GSD determines the absolute scale of the image and so far the modesl we discussed are generally unaware of this absolute scale when learning over a set of aerial data. ScaleMAE extends the positional embeddings from Eqn 2 by scaling the by scaling it with $\frac{g}{G}$. Where g is the GSD of the input image and G is a reference GSD, which was nominally set to 1.

Progressive Laplacian Decoder. ScaleMAE follows [42] to learn a higher resolution, high frequency image and a lower resolution low frequency image. As visualized in Figure 5, where low frequency and high frequency images are 224px and 448px respectively. MAE and SatMAE, uses a traditional transformers decoder block, whereas ScaleMAE replaces it with a progressive Laplacian decoder architecture, consisting of three key stages: decoding, upsampling, and reconstruction. The decoding phase is typical to the MAE, where masked patches are reintroduced into their original positions with a learned mask token along with their positional embeddings. These are then passed through a reduced number of Transformer layers (3, instead of 8 in standard MAE) giving us latent feature maps rather than a reconstructed single image. This feature maps are progressively upsampled using deconvolution blocks followed by a LayerNorma, GLEU and another decovolution block, each iteration of this block outputs a feature map at 2x the previous resolution. The feature map after the first upsampling is used for reconstructing the low resolution low frequency image and the feature map for the higher frequency image is upsampled twice as shown in Figure 5. The final stage involves passing the high and low frequency latent feature maps through a Laplacian Block, denoted as LB in Figure 5. Each Laplacian Block comprises of three sub-blocks: a *Laplacian Feature Mapping Block*, that projects latent features from each layer of the Laplacian pyramid into the RGB space. *Laplacian Upsampling Block*, mapping latent features from one layer of the Laplacian Pyramid to a higher level. And finally the *Laplacian Pyramid Reconstruction Block*, used to reconstruct different frequency images in the RGB space. Which are then compared t the ground truth images using the L1 loss for high frequency image and L2 loss for low frequency image following [1]

These architectural innovation enables to ScaleMAE to outperform all the previously discussed SSL methods for all popular benchmark remote sensing datasets, with a significant performance increase in small-scale scene classification tasks. Like identifying airplanes, parking lots and roundabouts.

3. Approach

While reviewing these SSL methods for satellite imagery, I noticed a lack of geographic diversity in the labeled benchmark datasets used for evaluating the intermediate representations learned by the pretrained SSL models.

These datasets tend to heavily focus on the global North, particularly regions in North America and Europe, leaving large portions of the globe underrepresented. This lack of diversity can result in SSL models that generalize poorly to imagery from less represented regions such as Africa, South America, and Southeast Asia, where environmental, agricultural, and urban features differ significantly. Satellite imagery features are inherently diverse across the globe, reflecting variations in climate, vegetation, infrastructure, and land use. For instance, tropical rainforests in the Amazon and Congo basins exhibit dense, multilayered canopies that differ significantly in spectral and structural properties compared to temperate forests in Europe or North America. Similarly, urban development in Southeast Asia, characterized by high-density settlements and informal housing patterns, contrasts starkly with the sprawling, organized cities in developed regions. In agricultural areas, techniques like terraced farming in Southeast Asia and smallholder farming in Africa create distinct spatial patterns that are vastly different from large-scale mechanized farming in the global North.

3.1. GlobalPatches

To address the lack of geographic and temporal diversity in existing labeled datasets, I developed GlobalPatches, a geographically diverse and temporally representative land cover dataset. GlobalPatches leverages high-resolution Sentinel-2 satellite imagery available through Google Earth Engine (GEE) and pixel-level labels from Google's DynamicWorld [3] project. DynamicWorld is a project developed at Google to provide near realtime land cover classifications at a global scale. It uses a Machine Learning model trained on high resolution Sentinel-2 satellite imagery to generate pixel-level land cover maps. The labels comprise of Water, Trees, Grass, Flooded Vegetation, Crops, Shrub and Scrub, Built-Up Area, Bare Ground and Snow and Ice. GlobalPatches includes imagery from underrepresented regions such as . Thus ensuring coverage across continents, climates, terrains, and land use patterns.

Image Collection The Sentinel-2 imagery in GlobalPatches is carefully selected to account for both seasonal and regional variations in satellite data. Geographic boundaries for regions of interest were sourced from GeoBoundaries [33] and included areas such as Sejong (South Korea), Bali (Indonesia), Khartoum (Sudan), Rwanda, Canterbury (New Zealand), La Paz (Bolivia), Lebanon, Rio de Janeiro (Brazil), Ahmedabad (India) and Tippecanoe County (USA). Thus ensuring coverage across continents, climates, terrains, and land use patterns. These boundaries were converted into geospatial features for use in GEE, allowing precise spatial queries for image and label extraction. To capture seasonal variations, the dataset includes Sentinel-2 imagery for each quarter of the year, spanning four time periods: January–March, April–June, July–September, and October–December. This ensures that the dataset represents temporal diversity, accounting for seasonal changes in land cover, vegetation, and weather conditions. To ensure high quality data, Sentinel-2 images were filtered to include only those with less than 5% cloud cover, significantly reducing noise from clouds.

Labelling Using the same query parameters as the Sentinel-2 images, DynamicWorld images with thier per-pixel class labels are obtained and these two sets of images are joined using their timestamps. The joined data now has the RGB channels along with the labels. This joined data is now processed to to create a median composite for each time period, eliminating intra-season temporal noise while preserving essential features. Thereafter we divide this median composite image into patches of 224×224 pixels and each patch gets the label of the majority pixel labels.

3.2. Experiments

Along with GlobalPatches, I use 3 other datasets to evaluate the performance of the studied SSL methods, whose details are mentioned in Table 1. These datasets were specifically chosen due to their diversity in class categories, which encompass features observed at multiple spatial scales. Where EuroSAT and GlobalPatches classes are predominantly large-scale land cover types. In contrast, RESISC-45 and Optimal-31 include more complex and localized categories.

Linear probing is a widely used technique for evaluating the quality of representations learned by pretrained models. In this method, the pretrained model is frozen, and a lightweight linear classifier, such as a single-layer fully connected neural network is trained on top of the frozen representations. The objective of linear probing is to measure how well the fixed representations encode information relevant to the downstream classification task. Since the pretrained model's weights remain unchanged during this process, the performance of the linear classifier directly reflects the quality of the representations learned by the model during pretraining. By conducting linear probing on the EuroSAT, Optimal-31, RESISC-45, and GlobalPatches datasets, I was able to assess how well the representations from SeCo, SatMAE, and ScaleMAE generalized across diverse spatial scales, land cover types, and classification complexities. To highlight the shortcomings of using a lift-and-shift approach of using ImageNet pretrained models on satellite imagery downstream tasks, I also conduct linear probing on ReseNet50 [18] model pretrained on ImageNet-21k [31] as an example of a supervised pretraining method.

4. Results

Implementation Details. All linear probing tests were conducted on an M1 MacBook Pro, with pretrained models

Dataset	Resolution (px)	Number of Images	Classes	Category Examples
EuroSAT [21]	64	27,000	10	River, Pasture, Industrial, Forest, Residential
Optimal-31 [41]	256	1,860	31	Church, Bridge, Airplane, Lake, Meadow, Freeway
RESISC-45 [8]	256	31,500	45	Stadium, Tennis court, Terrace, Wetland, Mountain
GlobalPatches	224	39,486	10	Trees, SnowIce, Water, Bare Ground, Crops

Table 1. Overview of the datasets used in my experimentation

obtained from their official GitHub repositories. A batch size of 64 was used for all experiments. Due to the large size of the SatMAE and ScaleMAE models, which utilize a Vision Transformer (ViT) encoder, linear probing for these models was limited to 10 epochs. In contrast, the SeCo and ResNet tests were run for 50 epochs using the Adam optimizer. For SatMAE and ScaleMAE, the AdamW optimizer was employed with a weight decay of 0.05 and a layer-wise learning rate decay of 0.75 to account for the hierarchical structure of the ViT encoder. The base learning rate was set to 1e-3, unless explicitly specified otherwise in the respective literature, and was reduced by a factor of 10 at 60% and 80% of the total epochs.

Table 2 presents the top-1 and top-5 accuracies for the linear probing tests across all model and dataset pairs. The results align largely with expectations, with the exception of the underperformance of the SeCo model on the RESISC-45 and Optimal-31 datasets. Notably, despite using a ResNet50 backbone, SeCo performs worse than the ImageNet-pretrained ResNet50 on these datasets. This outcome is particularly surprising, as SeCo outperforms supervised ResNet50 training on datasets such as EuroSAT and GlobalPatches, which feature large-scale categories. This suggests that SeCo struggles to generalize effectively across varying scales.

	Top-1/Top-5 Validation Accuracy (%)			
Dataset	**ResNet50**	**SeCo**	**SatMAE**	**ScaleMAE**
RESISC-45	57.80/83.94	23.71/49.80	74.54/92.82	82.35/95.5
Optimal-31	69.53/80.64	43.37/76.70	66.04/78.13	73.68/87.09
EuroSAT	74.35/94.16	78.74/97.33	83.51/94.5	88.51/94.72
GlobalPatches	74.67/93.66	76.28/94.37	85.65/99.28	90.32/99.24

Table 2. Linear Probing Tests Top-1/Top5 accuracy

In Figure 6, we can clearly see superior intermediate representations for MAE-based models (ScaleMAE and SatMAE). Where-in the validation accuracy is above 70% in the first epoch itself compared to 20% of the other two models. This trend is visible across all the datasets tested. The significant performance gap observed between these MAE-based models and ResNet-based models can be attributed, at least partially, to differences in model architecture and complexity. The Vision Transformer (ViT) backbone of MAE-based models inherently provides greater representational capacity than the ResNet backbone, which likely contributes to their superior performance in these tasks.

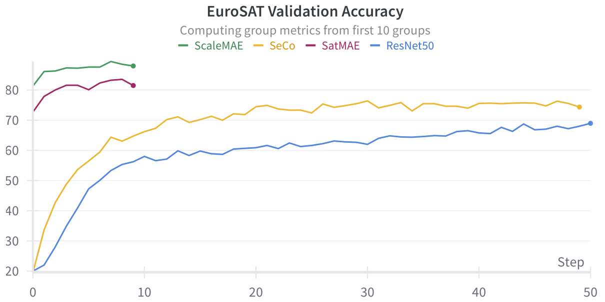

Figure 6. Validation Accuracy for all the four models on EuroSAT

ScaleMAE outperforms SatMAE by an average of 6.28 % in top-1 accuracy across all datasets. This highlights the importance of multiscale representations in satellite imagery.

5. Conclusions

Model Size. The results of our linear probing tests provide valuable insights but do not fully capture the practical implications of using SSL methods tailored for downstream satellite imagery tasks. One critical factor to consider is resource constraints during deployment. Large-scale models, while powerful, may not be ideal for applications requiring rapid processing or operating within limited computational budgets. This underscores the need for developing lightweight SSL methods for pretraining that balance performance and efficiency. An example is the Pretrained Remote Sensing Transformer (Presto) [38], which achieves competitive results comparable to MAE-based models while requiring up to four orders of magnitude fewer FLOPs to encode an image.

Multispectral Bands and Sensors. Among the methods discussed, SatMAE uniquely considers the use of multispectral bands in aerial imagery to learn richer representations. However, it relies on manually selecting and grouping bands, which may provide additional information beyond standard RGB channels. This manual approach, while effective, leaves room for improvement. A promising research direction could involve leveraging automated learning-based methods to optimize hyperspectral band selection In addition to multispectral data, satellite imagery often includes information from multiple sensors. Exploring multimodal fusion techniques to integrate these diverse data sources offers an exciting avenue for improving the quality of intermediate representations. For instance, methods like those proposed in [15] demonstrate how combining data from different modalities can enhance representation learning

Benchmark Datasets. As mentioned earlier, a critical limitation in the field of satellite machine learning (SatML) is the lack of globally diverse benchmark datasets with multiscale labels. Addressing this gap is essential not only for ensuring fairness and generalizability but also for aligning SatML research progress with real-world impact.

References

[1] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey, 2020. 6

[2] Josh Blumenfeld. Getting ready for nisar - and for managing big data using the commercial cloud. Feature Article, NASA Earthdata, 2021. Accessed: 2024-11-26. 1

[3] C. F. Brown, S. P. Brumby, B. Guzder-Williams, et al. Dynamic world, near real-time global 10m land use land cover mapping. *Scientific Data*, 9:251, 2022. 7

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1, 3

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021. 1

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 1, 2, 3

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 1, 3

[8] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 5, 8

[9] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world, 2018. 2

[10] 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. In *Advances in Neural Information Processing Systems*, 2022. 3, 4, 5

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 3

[12] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 3

[14] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners, 2022. 2, 3

[15] Anthony Fuller, Koreen Millard, and James R Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 9

[16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018. 2

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3, 7

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. 1, 2

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 1, 2

[21] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2019. 3, 5, 8

[22] Neal Jean, Marshall Burke, Ma Xie, William Davis, Xinyu Liang, Rachel Licker, Lu You, and David B. Lobell. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. 1

[23] 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. 5

[24] Oscar Mañas, Alexandre Lacoste, Xavier Giro i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data, 2021. 2, 4

[25] Catherine Nakalembe et al. Urgent and critical need for sub-saharan african countries to invest in earth observation-based agricultural early warning and monitoring systems. *Environmental Research Letters*, 15(12):121002, 2020. 1

[26] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles, 2017. 2

[27] Xiaoman Qi, PanPan Zhu, Yuebin Wang, 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, 2020. 2, 5

[28] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. OpenAI research paper. 3

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2019. OpenAI research paper. 3

[30] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *arXiv preprint arXiv:2212.14532*, 2022. 3, 5

[31] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021. 7

[32] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical – satellite data is a distinct modality in machine learning, 2024. 1, 2

[33] D. Runfola et al. geoboundaries: A global database of political administrative boundaries. *PLoS ONE*, 15(4):e0231866, 2020. 7

[34] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery, 2019. 6

[35] Andrew K. Skidmore, Nicholas C. Coops, Ehsan Neinavaz, Amr Ali, Michael E. Schaepman, Massimiliano Paganini, W. D. Kissling, Päivi Vihervaara, Ramin Darvishzadeh, Hendrik Feilhauer, et al. Priority list of biodiversity metrics to observe from space. *Nature Ecology & Evolution*, 5(7):896–906, 2021. 1

[36] Gencer Sumbul, Marcela Charfuelan, Begum Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019. 2, 3

[37] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. 5

[38] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries, 2024. 8

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 2

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 4

[41] 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, 2019. 5, 8

[42] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. A fully progressive approach to single-image super-resolution, 2018. 6

[43] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training, 2023. 3

[44] Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning, 2021. 3

[45] J. Yang, P. Gong, R. Fu, et al. Erratum: The role of satellite remote sensing in climate change studies. *Nature Climate Change*, 3(10):1001, 2013. 1

[46] 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. 6

[47] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization, 2016. 2