

Feature Guided Masked Autoencoder for Self-supervised Learning in Remote Sensing

Yi Wang, *Student Member, IEEE*, Hugo Hernández Hernández, Conrad M Albrecht, *Member, IEEE*,
Xiao Xiang Zhu, *Fellow, IEEE*

Abstract—Self-supervised learning guided by masked image modelling, such as Masked AutoEncoder (MAE), has attracted wide attention for pretraining vision transformers in remote sensing. However, MAE tends to excessively focus on pixel details, thereby limiting the model’s capacity for semantic understanding, in particular for noisy SAR images. In this paper, we explore spectral and spatial remote sensing image features as improved MAE-reconstruction targets. We first conduct a study on reconstructing various image features, all performing comparably well or better than raw pixels. Based on such observations, we propose *Feature Guided Masked Autoencoder* (FG-MAE): reconstructing a combination of Histograms of Oriented Gradients (HOG) and Normalized Difference Indices (NDI) for multispectral images, and reconstructing HOG for SAR images. Experimental results on three downstream tasks illustrate the effectiveness of FG-MAE with a particular boost for SAR imagery. Furthermore, we demonstrate the well-inherited scalability of FG-MAE and release a first series of pretrained vision transformers for medium resolution SAR and multispectral images.

Index Terms—remote sensing, Earth observation, geospatial foundation models, self-supervised learning, masked autoencoder

I. INTRODUCTION

SELF-SUPERVISED Learning has brought breakthroughs to the remote sensing (RS) community with the ability to learn generic representations from large-scale unlabeled data [1]. The pretrained encoders (recently also called foundation models) can then be transferred to various downstream applications. While convolutional neural networks have been long studied as model backbones with contrastive learning [2], there is a growing trend of pretraining vision transformers (ViT) [3] with masked image modeling (MIM), particularly, masked autoencoder (MAE) [4] and its variants [5].

MAE works as masking some patches of an input image, encoding the unmasked patches, and reconstructing the masked patches. Such asymmetric encoder-decoder design makes it highly efficient compared to contrastive learning. However, reconstructing raw input makes MAE over-focus pixel details, sensible to artifacts and noise, and potentially diverting attention from high-level semantic representations. These challenges are exacerbated in synthetic aperture radar (SAR) scenarios, in which the existence of speckle noise,

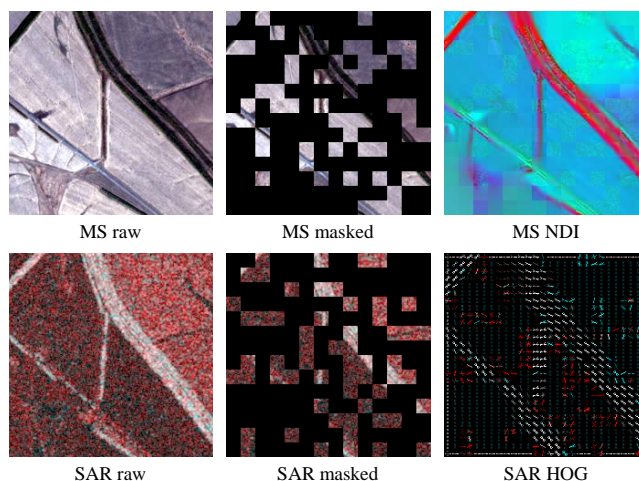


Fig. 1. Sample data of the proposed FG-MAE method—columns from left to right: Sentinel-2 multispectral (MS) and Sentinel-1 (SAR) imagery, masked model inputs, model-reconstructed features (HOG: Histogram of Gradients, NDI: Normalized Difference Index). False color of the raw SAR image is coded by $[\text{VV}, \text{VH}, (\text{VV}+\text{VH})/2]$. False color of the reconstructed MS NDI is coded by $[\text{NDVI}, \text{NDWI}, \text{NDBI}]$.

which appears as a granular disturbance and usually modeled as a multiplicative noise, limits MAE’s performance.

In this work, we propose a new simple variant of MAE for RS imagery, termed Feature Guided Masked Autoencoder (FG-MAE), by replacing raw images with image features as reconstruction targets. Looking back at traditional RS image analysis, human designed feature descriptors (e.g. edge or vegetation index) have been widely used to extract semantic information of the Earth’s surface [6, 7]. These image features incorporate expert knowledge, and can guide the model’s learning process when introduced to MAE. To demonstrate that, we conduct a study on popular features for multispectral and SAR imagery: 1) CannyEdge [8], 2) histograms of oriented gradients (HOG) [9], 3) scale-invariant feature transform (SIFT) [10], and 4) normalized difference indices (NDI) [11, 12, 13]. We show that each of these features alone works comparably well or even better than the original MAE.

We then search for the best candidates among the popular features, and propose FGMAE-MS and FGMAE-SAR. For multispectral imagery, we combine the spatial feature HOG and the spectral feature NDI, using two separate prediction heads at the end of the decoder. This combination allows the spatial and spectral features to complement each other. For SAR imagery, we simply use HOG to enhance spatial information and reduce the influence of speckle noise.

Y. Wang, H. H. Hernández and X. Zhu are with the chair of Data Science in Earth Observation, Technical University of Munich (TUM), Germany.

Y. Wang and C. M. Albrecht are with Remote Sensing Technology Institute, German Aerospace Center (DLR), Germany.

X. Zhu is with Munich Center for Machine Learning, Munich, Germany.

Codes and pretrained models are available at <https://github.com/zhu-xlab/FGMAE>. The collected EuroSAT-SAR dataset is available at <https://huggingface.co/datasets/wangyi111/EuroSAT-SAR>.

We evaluate FG-MAE on scene classification and semantic segmentation downstream tasks with BigEarthNet-MM [14], EuroSAT [15] and DFC2020 [16] datasets for both multispectral and SAR images. For EuroSAT, we match the geocoordinates of EuroSAT-MS and collect the EuroSAT-SAR dataset. Results demonstrate the effectiveness of FG-MAE on all tasks, particularly in SAR scenarios. In addition, FG-MAE remains as efficient as MAE, making it possible to scale up to big foundation models. We show that both FGMAE-MS and FGMAE-SAR scale well up to ViT-Huge with 0.7B parameters under linear evaluation protocols.

Our main contributions are listed as follows:

- We demonstrate the effectiveness of using RS image features as reconstruction targets for masked image modeling based self-supervised learning;
- We propose FG-MAE, a new variant of MAE that works well for both multispectral and SAR imagery;
- We show the benefits of FG-MAE pretrained models on three popular MS&SAR datasets;
- We verify the scalability of FG-MAE, and release a first series of pretrained ViTs for multispectral and SAR images with parameter sizes ranging from 22M to 0.7B.

II. RELATED WORK

Masked image modeling for self-supervised learning

Masked image modeling (MIM) is a recent family of generative self-supervised learning that focus on pretraining vision transformers by reconstructing the masked input, such as iGPT [17], BEiT [18] and SimMIM [19]. Of particular interest, MAE [4] drew wide attention with substantial improvements on fine-tuning downstream tasks and efficient pretraining.

Our work, FG-MAE, is a simple variant of MAE. Instead of reconstructing raw images, we propose to reconstruct features that are better suited for RS imagery. FG-MAE is also closely related to MaskFeat [20], where the authors introduced masked feature prediction for self-supervised video representation learning. We propose to use the asymmetric encoder-decoder structure of MAE for efficiency, and explore the best features for multispectral and SAR imagery.

Masked image modeling in remote sensing Most existing MIM works in RS are based on MAE [21, 22]. SatViT [21] presents the benefits of a straightforward implementation of MAE on satellite images. Wang *et al.* [23] showcased the potential of MAE on PolSAR images. RingMo [22] modified the masking strategy by reversing some pixels in the masked patches to avoid complete lost of small objects. MAEST [24] implemented MAE on hyperspectral images with spectral masking. SatMAE [5] proposed temporal and spectral masking and positional encoding in multispectral remote sensing time series. Scale-MAE [25] introduced ground sampling distance positional encoding and multiscale reconstruction to capture the geospatial scale information of RS images. Our work differs from all aforementioned approaches by improving MAE for RS imagery from the perspective of reconstructing image features as targets.

Exploiting image features in remote sensing Image feature descriptors play a big role in traditional RS image analysis.

The normalized difference indices have long been used for Earth surface monitoring since the last century [26, 12]. Similarly, spatial features like HOG are widely used as input to machine learning algorithms [27]. In this work, we revisit these well-known human-designed features and let them be learned by deep neural networks. This approach leverages the expertise of human analysts to guide the training process and facilitate the learning of better representations.

III. METHODOLOGY

Our proposed FG-MAE is a simple variant of MAE [4] that replaces the reconstruction target with RS image features. As is illustrated in Figure 2, the image is divided into non-overlapping patches, and a random subset of these patches are masked out. The remaining visible patches are sent through the ViT encoder. The full set of encoded visible patches and learnable mask tokens are fed into the lightweight ViT decoder to reconstruct target features. During training, mean squared error or L2 loss is minimized only on masked patches. In the following subsections, we will discuss different feature candidates in III-A, and present the specific target designs for multispectral and SAR imagery in III-B, respectively.

A. Target features

We consider two categories and four types of RS image features: spatially, 1) CannyEdge [8], 2) HOG [9], and 3) SIFT [10]; spectrally, 4) NDI, including vegetation index [11], water index [12] and built-up index [13].

CannyEdge CannyEdge [8] is an edge detection algorithm that identifies the edges in an image by tracing the gradient of pixel intensities. The algorithm works by convolving the image with a Gaussian filter to reduce noise, and then computing the gradient magnitude and direction of each pixel. Non-maximum suppression is applied to suppress non-max edge contributors, and edges are detected by applying a Hysteresis threshold to the gradient magnitude.

Edge descriptors can simplify complex images by highlighting object boundaries, facilitating object identification and tracking in computer vision algorithms. As one of the most popular algorithms in this family, CannyEdge has the ability to accurately detect edges while minimizing false positives. It can also adapt to changes in lighting and contrast, which can often cause issues for other edge detection algorithms. Additionally, CannyEdge is able to accurately detect edges regardless of their orientation or position within the image. This makes it a powerful tool for remote sensing applications [28].

CannyEdge is easy to compute in any deep learning framework by convolution, non-max suppression and thresholding. We use the filter toolbox of kornia [29] to extract the edges as MAE targets (one edge map from one image channel). The same process as reconstructing the raw image follows, including patchifying and normalization within each small patch.

HOG Histograms of Oriented Gradients [9] is a feature descriptor to describe the distribution of gradient orientations within a local subregion of an image. The algorithm calculates

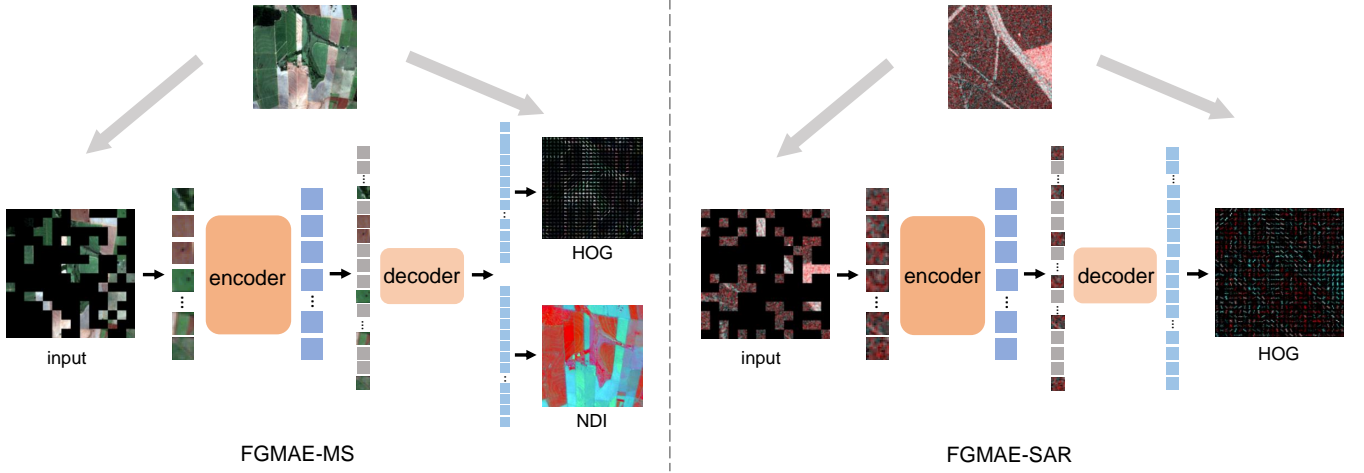


Fig. 2. The general structure of the proposed FG-MAE method. We replace the reconstruction target of MAE [4] by remote sensing image features.

the magnitudes and orientations of gradients at each pixel using gradient filtering. Then, the gradients within each small local window are accumulated into normalized orientation histogram vectors voted by gradient magnitudes.

HOG is able to capture local shapes and appearances while being partially invariant to geometric changes. HOG is also invariant to photometric changes, as image gradients and local contrast normalization absorb brightness and foreground-background contrast variation. Unlike CannyEdge, HOG does not focus solely on edges but provides information about the magnitude and orientation of edge gradients.

HOG can be implemented similarly to CannyEdge as a two-channel convolution to generate gradients, followed by histogramming and normalization. We follow the implementation of MaskFeat [20] that writes HOG as (weight-fixed) neural network modules. Each channel of the raw image provides one HOG feature. The histograms of masked patches are then flattened and concatenated into a 1-D vector as the target feature.

SIFT Scale-invariant feature transform (SIFT) [10] is a feature descriptor that is used to extract distinctive and invariant local features from images. It works by detecting key points in an image that are invariant to scale, rotation, and illumination changes. Once the key points are detected, SIFT computes a descriptor for each key point by extracting the local image gradient orientations and magnitudes. These gradients are then transformed into a histogram of orientations, which is used to create a feature vector that describes the local image patch around the key point.

The SIFT descriptor is robust against scale, rotation, illumination, and noise, making it applicable for a wide range of applications like image registration [30]. However, the complicated workflow of key point detectors and feature descriptors make it difficult for the model to learn. Another specific issue is that instead of region-based features, SIFT provides point-based features that do not align well with a standard ViT model design. Accordingly, it is tricky to integrate the famous SAR-SIFT [31] algorithm for SAR images. How to efficiently deal with the dynamic key points and the model’s learning

capacity remains a challenging task for future research. As a preliminary showcase in this work, we simplify the key point detection process by computing SIFT descriptor densely over the image. We utilize the feature toolbox of kornia [29] to calculate dense SIFT features. Due to memory constraints, we perform the calculation using grayscale images.

NDI Normalized Difference Indices (NDI) is a technique used to identify one type of ground objects by quantifying the differences between two spectral bands. It is often used in remote sensing applications such as changes in vegetation health or soil moisture levels. NDI works by calculating the ratio of the difference between two feature-sensitive spectral bands to their sum. This ratio is then normalized to a range between -1 and 1, where values closer to 1 indicate an increase in the feature of interest.

NDI is a simple and effective way to detect changes in vegetation health or soil moisture levels, as it is sensitive to changes in the reflectance of different spectral bands. Three most popular NDIs are normalized difference vegetation index (NDVI), normalized difference water index (NDWI), and normalized built-up index (NDBI):

$$NDVI = \frac{NIR - R}{NIR + R} \quad (1)$$

$$NDWI = \frac{G - NIR}{G + NIR} \quad (2)$$

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} \quad (3)$$

where NIR represents near infrared, R represents red, G represents green, and $SWIR$ represents short wave infrared. In this work, we calculate the three indices for each pixel and concatenate them into a three-channel target image.

In our experiments, we demonstrate that all above features serve as good reconstruction targets to replace raw images. Results will be discussed in V-A, where we perform a study on separately reconstructing the above features and evaluate corresponding downstream performances.

B. FGMAE-MS / SAR

We then develop our proposed self-supervised methods, FGMAE, based on the feature study. We consider two popular modalities in RS, multispectral imagery and polarimetric SAR imagery. For multispectral imagery, we combine spatial feature HOG and spectral feature NDI to complement each other; for SAR, we select HOG for its computational efficiency and noise robustness.

As is shown in Figure 2, we retain the asymmetric encoder-decoder structure of MAE while modifying the reconstruction targets. Specifically, for FGMAE-SAR, the augmented raw images with shape $(B, 2, W, H)$ are divided into L non-overlapping patches with shape (B, L, w, h) , of which L_m random patches are masked out. The remaining visible patches with shape $(B, L - L_m, w, h)$ are flattened to $(B, L - L_m, w * h)$, processed with a linear embedding layer to $(B, L - L_m, K_{en})$ and passed through the ViT encoder. The encoded visible patches have shape $(B, L - L_m, K_{en})$. At the beginning of the decoding process, a linear layer is used to embed encoded patches to $(B, L - L_m, K_{de})$. They are then combined with mask tokens to (B, L, K_{de}) as input to a lightweight ViT decoder. The last layer of the decoder is a linear layer that converts the decoded patches to HOG predictions with shape (B, L, K_{out}) , where K_{out} is defined by HOG window size, number of bins and input channel numbers.

While mostly similar for FGMAE-MS, the last layer of the decoder is replaced by two parallel linear layers, one outputting HOG and the other NDI. Note that for both modalities the outputs cover all patches, and only the masked ones are counted in the L2 loss calculation.

IV. EXPERIMENTAL SETUP

A. Self-supervised pretraining

Dataset We pretrain vision transformers on Sentinel-1 GRD and Sentinel-2 L1C products of SSL4EO-S12 dataset [32]. The dataset is sampled from 250K locations around the world. Each location has four images from four seasons with size 264×264 and ground sampling distance 10m. The multispectral images have 13 channels, and the SAR images have 2 channels.

Data augmentation One image from a random season is selected for one location, followed by RandomResizedCrop to 224×224 and RandomHorizontalFlip as the data augmentations.

Model architecture We adopt the architecture design of MAE [4], which includes a regular ViT encoder (by default ViT-S/16 unless specifically noted) and a lightweight ViT decoder. Only the encoder is transferred to downstream tasks. The masking ratio is set to 70% as recommended in [32].

Optimization We pretrain ViTs with batchsize 256 for 100 epochs. We use the AdamW optimizer [33] with weight decay 0.05 and a basic learning rate $1.5e-4$. The learning rate is warmed up for 10 epochs, and then decayed with a cosine schedule. Training is distributed across four NVIDIA A100 GPUs and takes about 7 hours for multispectral and 4 hours for SAR.

B. Transfer learning

Dataset The pretrained models are transferred to scene classification and semantic segmentation downstream tasks for both multispectral and SAR imagery. For

- *scene classification*, we evaluate EuroSAT [15] (single-label land cover classification) and BigEarthNet-MM [14] (multi-label land cover classification) via linear probing (freeze encoder) and end-to-end fine tuning.
- *semantic segmentation*, we evaluate DFC2020 [16] (land cover segmentation) via fine tuning.

BigEarthNet-MM and DFC2020 have both multispectral and SAR images available. For BigEarthNet-MM, we use the 19-class labels, and follow the official train/val/test splits. For DFC2020, we use the 10-class high-resolution segmentation labels, and adjust the official test/validation data for 5128 training and 986 testing images. For EuroSAT, we perform a random 80%/20% train/test split.

Since EuroSAT has only RGB and multispectral images, we collected EuroSAT-SAR by pairing the published EuroSAT-MS from Sentinel-1 GRD products. This is done by matching the geocoordinates of EuroSAT-MS images and downloading the corresponding patches with Google Earth Engine [34]. Because EuroSAT-MS has no exact collection time information, we performed a rough year-level match based on the publication time. In the end, we performed a manual check on random patches for the semantic correctness.

Data augmentation We follow a common practice to use RandomResizedCrop (scale 0.2 to 1.0, resized to 224×224) and RandomHorizontalFlip as data augmentations for all linear probing experiments. For BigEarthNet-MM, we set the smallest crop scale as 0.8 to avoid cutting out too many objects for the multilabel task. For DFC2020, we set the resized image size 256×256 following MAE [4]. For fine tuning experiments, we add mixup [35] augmentation. The multispectral images of BigEarthNet-MM (Sentinel-2 L2A) are zero-padded to 13 channels to match the pretrained models.

Model architecture We use standard ViTs for scene classification on BigEarthNet-MM and EuroSAT. For semantic segmentation on DFC2020, we use UperNet [36] with ViT backbones following MAE [4].

Optimization For BigEarthNet-MM, we minimize MultiLabelSoftMargin loss. The batchsize is set to 256. For linear probing, we train SGD optimizer without weight decay for 50 epochs. For fine tuning, we train AdamW optimizer with weight and layer decay for 20 epochs. The learning rate is 0.5 with cosine decay for linear probing, and $1e-3$ with cosine decay and 3-epoch warm-up for fine tuning.

For EuroSAT, we minimize cross entropy loss. The batchsize is set to 256. For linear probing, we train SGD optimizer with weight decay 0.001 for 50 epochs. For fine tuning, we train AdamW optimizer with weight and layer decay for 20 epochs. The learning rate is 0.1 with cosine decay for linear probing, and $1e-3$ with cosine decay and 3-epoch warm-up for fine tuning.

For DFC2020, we use the RSI-Segmentation library [37] for fine tuning. We minimize cross entropy loss for 40k iterations

with batchsize 8. We use AdamW optimizer with layer decay. The basic learning rate is 1e-4, which is warmed up for 500 iterations and then polynomial-decayed.

Evaluation metrics We use mean average precision (mAP) and F1 score for the evaluation of BigEarthNet-MM. Overall accuracy (OA) and class-wise average accuracy (AA) are used for EuroSAT. For DFC2020, we evaluate overage accuracy (OA), average accuracy (AA) and mean intersection over union (mIoU).

V. RESULTS

A. FG-MAE: target features

We first conduct a study on replacing raw image with different target features in MAE for both multispectral and SAR imagery. We pretrain ViTs on SSL4EO-S12 and transfer them to a 10% subset of BigEarthNet-MM. As shown in Table I, all features perform comparably well to the raw image (MAE) under both linear probing and fine tuning settings in multispectral imagery. HOG is even better than the raw image for both settings. This proves the effectiveness of reconstructing image features as a new variant of MAE.

TABLE I
A STUDY OF THE FEATURES ON BIGEARTHNET-10% – MS.

	Linear probing	Fine tuning
Rand. Init.	70.3	-
Supervised	-	81.3
Raw image (MAE)	77.8	84.8
CannyEdge	77.9	84.8
HOG	77.9	85.0
Dense SIFT	77.8	84.9
NDI	77.3	84.6
HOG&NDI (ours)	78.1	85.2

Among the individual features, both CannyEdge and HOG show an advantage over NDI in linear probing. This is due to the fact that spatial feature descriptors capture better image-level semantics from e.g. shape information, while the spectral feature NDI does not consider pixel relationships. In addition, while HOG performs best among individual features, NDI provides a good complement that combining both pushes the performances further.

A similar but more interesting behavior is shown in Table II for SAR imagery. We can observe that both SIFT and HOG perform better than raw image (MAE), and HOG provides a remarkable boost. This can be attributed to the fact that MAE reconstructs every pixel and thus strongly disturbed by the speckle noise, while feature descriptors provide natural noise filtering. Furthermore, Dense SIFT performs worse than HOG. This is due to the coarse setting that we consider each pixel as one key point and thus have too many false positives. In fact, this inspires a promising research direction to better integrate scale-invariant features into MAE structure.

Qualitative examples of feature reconstruction can be seen in Figure 1 and 3. Despite masking out 70% of the input patches, the reconstruction results remain impressive for multi-spectral images. For SAR images, the ground truth themselves are very noisy, but interestingly, the reconstructed features

TABLE II
A STUDY OF THE FEATURES ON BIGEARTHNET-10% – SAR.

	Linear classification	Fine tuning
Rand. Init.	58.1	-
Supervised	-	72.7
Raw image (MAE)	69.8	74.9
CannyEdge	69.9	74.9
Dense SIFT	69.8	75.8
HOG (ours)	71.7	78.0

appear clearer than the ground truth. This observation suggests another exciting research direction for better low-level feature extraction algorithms [31].

B. FGMAE-MS

We then benchmark the performance of the proposed FGMAE-MS (HOG+NDI) and FGMAE-SAR (HOG) on extensive downstream datasets. Table III shows the transfer results on the full set of the multi-label scene classification dataset BigEarthNet. The proposed FGMAE-MS outperforms MAE consistently on both linear probing and fine tuning, with improvements up to 0.9%.

TABLE III
FGMAE-MS ON BIGEARTHNET-100%.

	Linear classification		Fine tuning	
	mAP	F1	mAP	F1
Rand. Init.	72.0	60.0	-	-
Supervised	-	-	87.8	78.9
MAE	78.0	68.0	88.6	79.9
FG-MAE (ours)	78.5	68.7	89.3	80.8

Table IV presents the transfer learning results on the single-label scene classification dataset EuroSAT. Similar to BigEarthNet, slight but consistent improvements can be observed in all scenarios.

TABLE IV
FGMAE-MS ON EUROSAT.

	Linear classification		Fine tuning	
	OA	AA	OA	AA
Rand. Init.	79.3	79.5	-	-
Supervised	-	-	96.7	96.3
MAE	94.2	94.0	98.5	98.2
FG-MAE (ours)	94.8	94.8	98.7	98.5

Finally, Table V demonstrates the transfer learning results on the semantic segmentation dataset DFC2020, where FGMAE-MS outperforms MAE by noticeable margins on all metrics (e.g. 3.4% increase in mIoU). This underscores the promising benefits of FG-MAE on dense prediction tasks.

TABLE V
FGMAE-MS ON DFC2020.

	OA	mIoU	AA
Supervised	63.3	46.2	59.2
MAE	66.9	48.0	63.5
FG-MAE (ours)	69.6	51.4	66.4

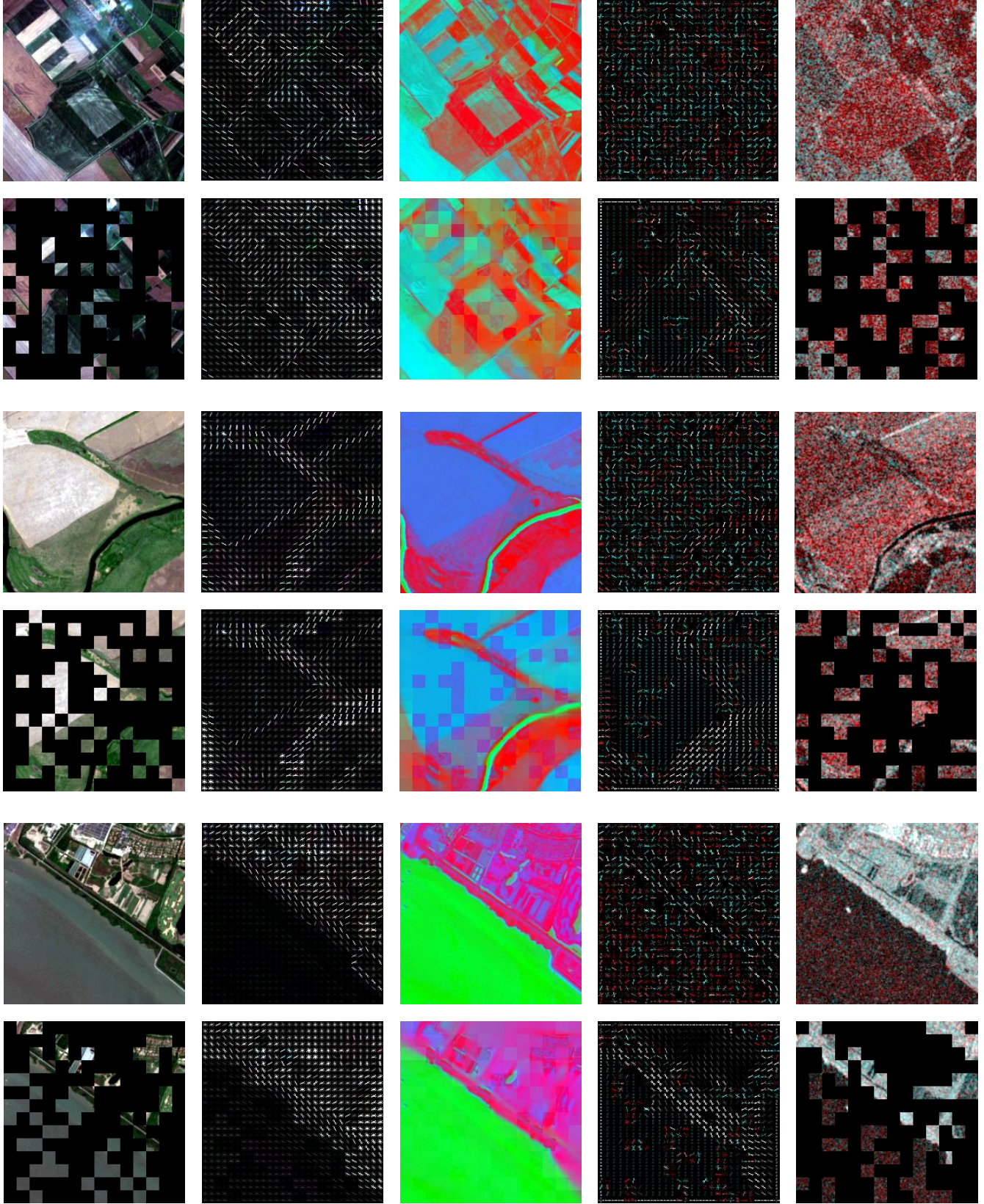


Fig. 3. Examples of FG-MAE reconstructed features. Every two rows represent one MS-SAR pair. From left to right, first row: MS image, MS HOG target, MS NDI target, SAR HOG target, SAR image; second row: MS image masked, MS HOG prediction, MS NDI reconstruction, SAR HOG prediction, SAR image masked.

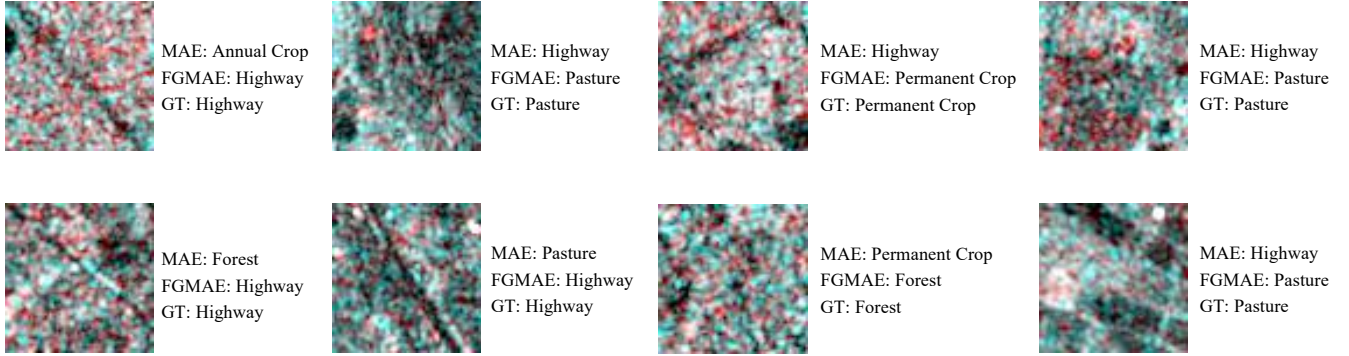


Fig. 4. Examples of EuroSAT-SAR prediction results where FG-MAE gives the correct label while MAE doesn't. FG-MAE better captures semantics that are more distinguishable from the HOG features (e.g. a highway).

TABLE VI
PER-CLASS BENCHMARK RESULTS ON EUROSAT-SAR. FG-MAE OUTPERFORMS MAE BY LARGE MARGINS ON MOST OF THE CLASSES.

	Annual Crop	Forest	Herbaceous Vegetation	Highway	Industrial	Pasture	Permanent Crop	Residential	River	Sea/Lake
Supervised	76.4	77.7	66.4	66.0	90.7	58.4	59.1	90.2	89.3	98.1
MAE	79.6	79.1	70.2	72.7	92.0	64.6	59.2	91.8	92.4	98.8
FG-MAE (ours)	84.1 (+3.5)	85.4 (+6.3)	78.1 (+7.9)	82.4 (+9.7)	93.5 (+1.5)	75.7 (+11.1)	67.7 (+8.5)	94.3 (+2.5)	94.2 (+1.8)	98.8

C. FGMAE-SAR

Likewise, we benchmark the transfer learning results on SAR imagery of the aforementioned datasets. As can be seen from Table VII, FGMAE-SAR demonstrates remarkable improvements compared to MAE on BigEarthNet. Especially when compared to the multispectral scenario (0.5% to 0.9% improvements), the benefit of FG-MAE is much more significant (up to 3.1%). This again highlights the advantage of implicit noise filtering with HOG features.

TABLE VII
FGMAE-SAR ON BIGEARTHNET-100%.

	Linear classification		Fine tuning	
	mAP	F1	mAP	F1
Rand. Init.	59.0	40.4	-	-
Supervised	-	-	79.5	71.1
MAE	70.4	59.1	81.3	72.8
FG-MAE (ours)	72.3	62.2	82.7	74.0

Table VIII presents the results on our collected EuroSAT-SAR dataset. Similar to BigEarthNet results, substantial performance boosts can be observed with FGMAE-SAR. While FGMAE-MS gives 0.2% to 0.8% improvements compared to MAE, FGMAE-SAR provides up to 5.0% improvement. Detailed per-class benchmarks are shown in Table VI, where FGMAE-SAR outperforms MAE by a large margin for most of the classes (e.g. as much as 11.1% for the pasture class). Figure 4 presents some patch examples, which MAE misclassifies while FG-MAE predicts the correct label. We can observe from the figure that FG-MAE helps the model better capture the semantics that are easier to recognize with HOG features (e.g. a highway image).

Finally on DFC2020, consistent improvements compared to MAE can be seen from Table IX. Though the improvements compared to FGMAE-MS here are not as much as the previous two scene classification datasets, they are still noteworthy

TABLE VIII
FGMAE-SAR ON EUROSAT.

	Linear classification		Fine tuning	
	OA	AA	OA	AA
Rand. Init.	61.9	61.3	-	-
Supervised	-	-	78.4	77.7
MAE	79.3	78.6	81.0	80.4
FG-MAE (ours)	80.7	79.9	85.9	85.4

compared to supervised learning. This is also shown in Figure 6, where the segmentation results of two example image pairs are presented. The limited benefits can be attributed to the characteristics of SAR imagery, where interpreting fine grained pixel details is very challenging.

TABLE IX
FGMAE-SAR ON DFC2020.

	OA	mIoU	AA
Supervised	61.4	37.3	56.1
MAE	62.1	38.9	56.9
FG-MAE (ours)	62.3	39.3	57.0

D. Scaling ViTs

The efficiency of MAE is well-preserved in the proposed FG-MAE, thus we are able to scale-up the pretrained models to a series of ViTs with up to 658 million parameters: ViT-Small, ViT-Base, ViT-Large and ViT-Huge. We evaluate linear classification and fine tuning results on both multispectral and SAR imagery of the BigEarthNet-MM dataset. As is shown in Figure 5, scaling up ViTs provides consistent improvements for both modalities under linear classification protocol. This supports the potential benefits of even larger foundation models [38]. However, we also observe significant overfitting

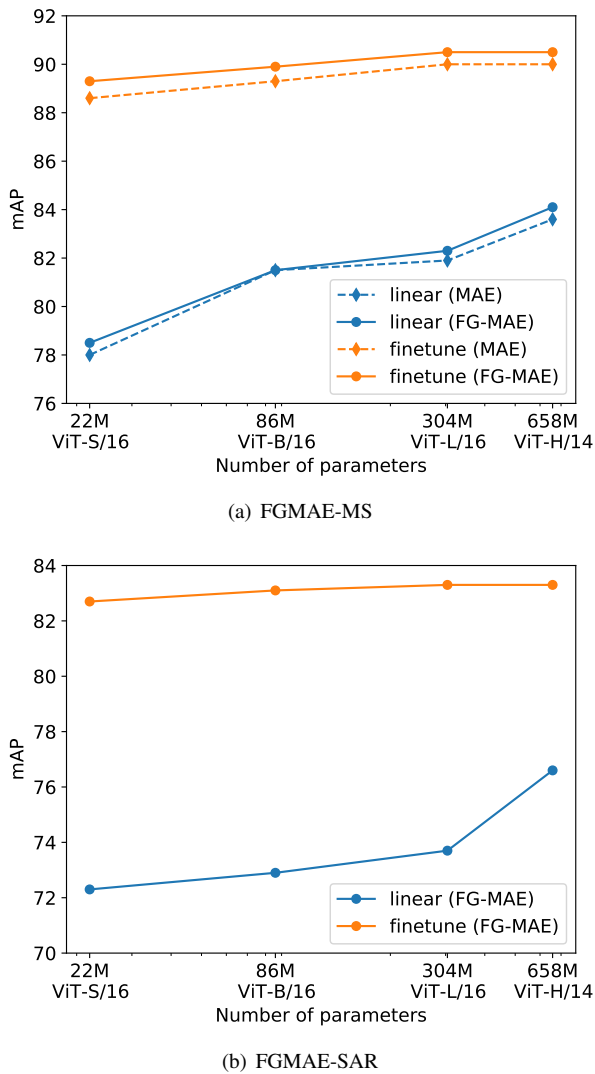


Fig. 5. Similar to MAE, FG-MAE scales well on BigEarthNet linear evaluation for both multispectral and SAR imagery.

phenomenon under fine tuning protocol, as reflected by the saturation trend in Figure 5. This indicates the need for further research on how to effectively fine-tune big foundation models.

VI. CONCLUSION

In this study, we demonstrated that image features are comparable or superior reconstruction targets for masked image modelling based pretraining in remote sensing, particularly for SAR imagery. We proposed a novel variant of MAE, called feature guided masked autoencoder (FG-MAE), which modifies the reconstruction targets. For multispectral imagery, we combined HOG and NDI, while for SAR imagery, we used HOG alone. Experimental results on three downstream tasks verify the effectiveness of FG-MAE. In addition, we demonstrated the scalability of FG-MAE, and released a series of pretrained vision transformers with size up to 0.7B parameters for multispectral and SAR imagery.

Though the proof of concept has been made clear, one limitation of this work is that we can not make the best use

of scale-invariant features such as SIFT / SAR-SIFT out-of-the-box. However, we believe these features are of great value with proper and more sophisticated design. Scale-MAE [25], for example, though not directly inspired by SIFT, shares a similar idea and provides promising insights.

Another limitation, as we have mentioned, is that both MAE and FG-MAE scale well to larger backbones in linear probing, but not in fine tuning. As we are entering the era of big EO foundation models, how to effectively transfer the foundation knowledge remains an important but not yet well-studied problem.

There are also two interesting thoughts that we believe deserve further investigation. First, we have shown a relatively poor performance in reconstructing raw SAR images because of the effect of speckle noise. However, what if we reconstruct the despeckled images instead? SAR-despeckling has been widely studied and there are many well-developed algorithms. If integrated into MAE pretraining, would it help the model prevent confusion due to noise? Second, the reconstructed SAR features sometimes seem to be clearer than the corresponding noisy ground truth. This may inspire a promising direction for low-level tasks, including the aforementioned SAR-despeckling.

ACKNOWLEDGEMENT

This work was funded by the Helmholtz Association through the Framework of Helmholtz AI, grant ID: ZT-I-PF-5-01 – *Local Unit Munich Unit @Aeronautics, Space and Transport (MASTr)*. The compute was supported by the Helmholtz Association’s Initiative and Networking Fund on the HAICORE@FZJ partition. The work of X. Zhu is supported by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab ”AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond” (grant number: 01DD20001).

REFERENCES

- [1] Yi Wang et al. “Self-Supervised Learning in Remote Sensing: A Review”. In: *IEEE Geoscience and Remote Sensing Magazine* (2022).
- [2] Neal Jean et al. “Tile2vec: Unsupervised representation learning for spatially distributed data”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 3967–3974.
- [3] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *International Conference on Learning Representations*. 2020.
- [4] Kaiming He et al. “Masked autoencoders are scalable vision learners”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.
- [5] Yezhen Cong et al. “Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 197–211.
- [6] Mohamed Ali and David Clausi. “Using the Canny edge detector for feature extraction and enhancement of remote sensing images”. In: *IEEE International Geoscience and Remote Sensing Symposium*. Vol. 5. Ieee. 2001, pp. 2298–2300.

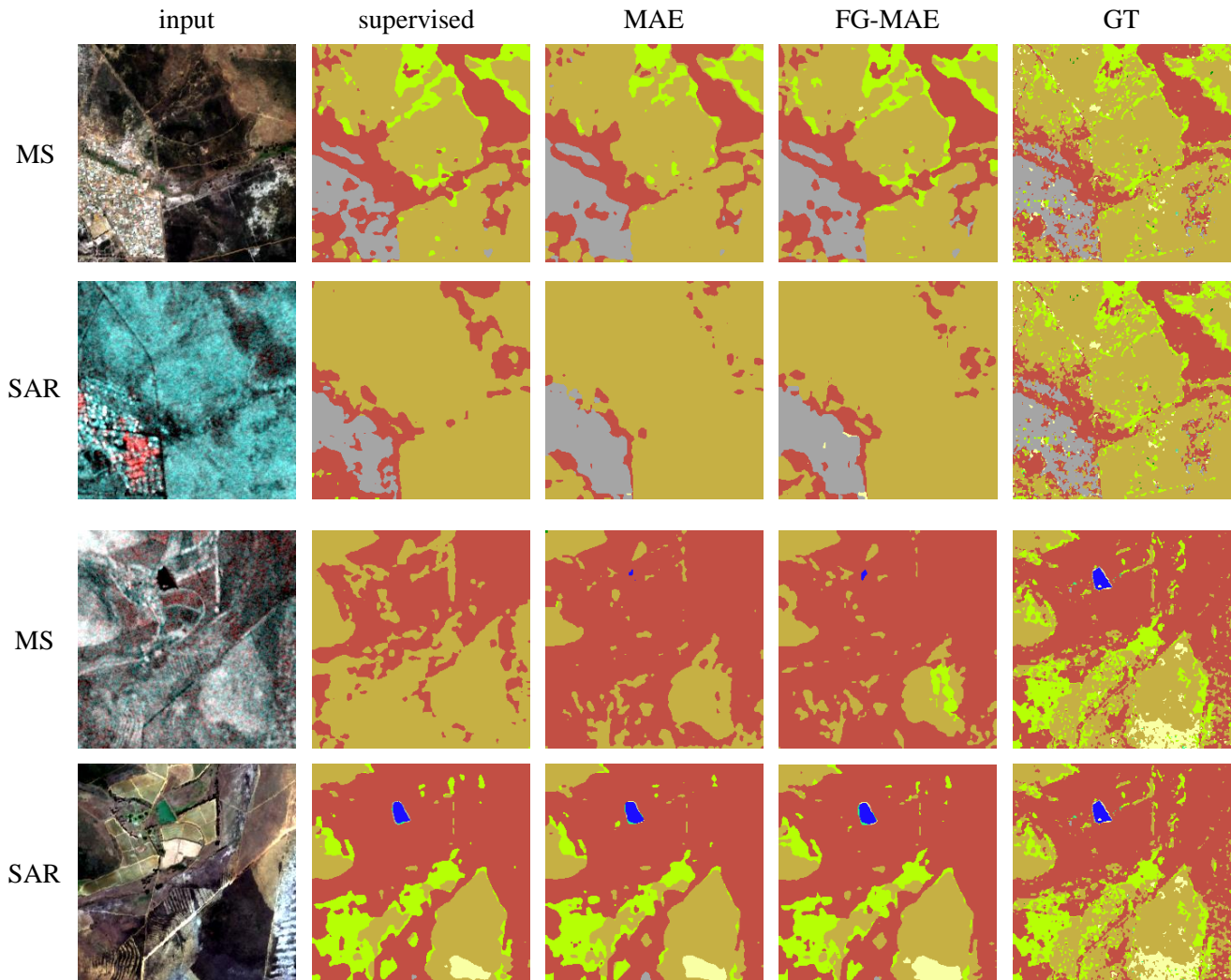


Fig. 6. Examples of DFC2020 segmentation maps. Every two rows represent one MS-SAR pair. From left to right, first row: MS image, MS prediction supervised, MS prediction MAE, MS prediction FG-MAE, ground truth mask; second row: SAR image, SAR prediction supervised, SAR prediction MAE, SAR prediction FG-MAE, ground truth mask.

- [7] Ross S Lunetta et al. “Land-cover change detection using multi-temporal MODIS NDVI data”. In: *Remote sensing of environment* 105.2 (2006), pp. 142–154.
- [8] John Canny. “A computational approach to edge detection”. In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), pp. 679–698.
- [9] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Vol. 1. Ieee. 2005, pp. 886–893.
- [10] David G Lowe. “Distinctive image features from scale-invariant keypoints”. In: *International journal of computer vision* 60 (2004), pp. 91–110.
- [11] Nathalie Pettorelli. *The normalized difference vegetation index*. Oxford University Press, 2013.
- [12] Bo-Cai Gao. “NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space”. In: *Remote sensing of environment* 58.3 (1996), pp. 257–266.
- [13] Yong Zha, Jay Gao, and Shaoxiang Ni. “Use of normalized difference built-up index in automatically mapping urban areas from TM imagery”. In: *International journal of remote sensing* 24.3 (2003), pp. 583–594.
- [14] Gencer Sumbul et al. “BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]”. In: *IEEE Geoscience and Remote Sensing Magazine* 9.3 (2021), pp. 174–180.
- [15] Patrick Helber et al. “Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2019).
- [16] Michael Schmitt et al. *IEEE GRSS Data Fusion Contest*. 2020.
- [17] Mark Chen et al. “Generative pretraining from pixels”. In: *International conference on machine learning*. PMLR. 2020, pp. 1691–1703.
- [18] Hangbo Bao et al. “BEiT: BERT Pre-Training of Image Transformers”. In: *International Conference on Learning Representations*. 2021.
- [19] Zhenda Xie et al. “Simim: A simple framework for masked image modeling”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 9653–9663.
- [20] Chen Wei et al. “Masked feature prediction for self-supervised visual pre-training”. In: *Proceedings of the IEEE/CVF Con-*

- ference on Computer Vision and Pattern Recognition. 2022, pp. 14668–14678.
- [21] Anthony Fuller, Koreen Millard, and James R Green. “SatViT: Pretraining Transformers for Earth Observation”. In: *IEEE Geoscience and Remote Sensing Letters* 19 (2022), pp. 1–5.
- [22] Xian Sun et al. “Ringmo: A remote sensing foundation model with masked image modeling”. In: *IEEE Transactions on Geoscience and Remote Sensing* (2022).
- [23] Hongmiao Wang et al. “Land Cover Classification for Polarimetric SAR Images Based on Vision Transformer”. In: *Remote Sensing* 14.18 (2022), p. 4656.
- [24] Damian Ibanez et al. “Masked Auto-Encoding Spectral–Spatial Transformer for Hyperspectral Image Classification”. In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–14.
- [25] Colorado J Reed et al. “Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4088–4099.
- [26] Toby N Carlson and David A Ripley. “On the relation between NDVI, fractional vegetation cover, and leaf area index”. In: *Remote sensing of Environment* 62.3 (1997), pp. 241–252.
- [27] Peter A Torrione et al. “Histograms of oriented gradients for landmine detection in ground-penetrating radar data”. In: *IEEE Transactions on Geoscience and Remote Sensing* (2013).
- [28] H Liu and KC Jezek. “Automated extraction of coastline from satellite imagery by integrating Canny edge detection and locally adaptive thresholding methods”. In: *International journal of remote sensing* 25.5 (2004), pp. 937–958.
- [29] Edgar Riba et al. “Kornia: an open source differentiable computer vision library for pytorch”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 3674–3683.
- [30] Wenping Ma et al. “Remote sensing image registration with modified SIFT and enhanced feature matching”. In: *IEEE Geoscience and Remote Sensing Letters* 14.1 (2016), pp. 3–7.
- [31] Flora Dellinger et al. “SAR-SIFT: a SIFT-like algorithm for SAR images”. In: *IEEE Transactions on Geoscience and Remote Sensing* 53.1 (2014), pp. 453–466.
- [32] Yi Wang et al. “SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation”. In: *IEEE Geoscience and Remote Sensing Magazine* (2023).
- [33] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations*. 2018.
- [34] Noel Gorelick et al. “Google Earth Engine: Planetary-scale geospatial analysis for everyone”. In: *Remote sensing of Environment* (2017).
- [35] Hongyi Zhang et al. “mixup: Beyond Empirical Risk Minimization”. In: *International Conference on Learning Representations*. 2018.
- [36] Tete Xiao et al. “Unified perceptual parsing for scene understanding”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 418–434.
- [37] Zhitong Xiong et al. “EarthNets: Empowering AI in Earth Observation”. In: *arXiv:2210.04936* (2022).
- [38] Keumgang Cha, Junghoon Seo, and Taekyung Lee. “A billion-scale foundation model for remote sensing images”. In: *arXiv:2304.05215* (2023).
- [39] Timnit Gebru et al. “Datasheets for datasets”. In: *Communications of the ACM* (2021).

Appendix: EuroSAT-SAR Dataset

Below we provide extensive information about the collected EuroSAT-SAR dataset, which is a SAR version of EuroSAT. As a side contribution of this paper, we believe this simple, clean and well-balanced SAR dataset (which surprisingly rarely exists yet) is of great value to further machine-learning research and education on SAR imagery.

A. Data collection

To create EuroSAT-SAR, we match the published EuroSAT-MS (Sentinel-2 L1C) dataset with dual-pol Sentinel-1 GRD images from Google Earth Engine. Specifically, for each geotiff image in EuroSAT-MS, we extract the corresponding coordinate system and bounding box coordinates. We then build a geo-referenced rectangle region for the patch. Meanwhile, we build a temporal period between the years 2016 and 2017. Next, we filter available SAR images based on the region and period, and download a random qualified patch with bands VV and VH. Since no cloud filtering is needed for SAR imagery, the data collection is very fast (within one hour). In the end, we match the whole EuroSAT-MS dataset and download 27,000 SAR images, each assigned with the same class label as the corresponding MS image. Figure 7 illustrates the data collection process.

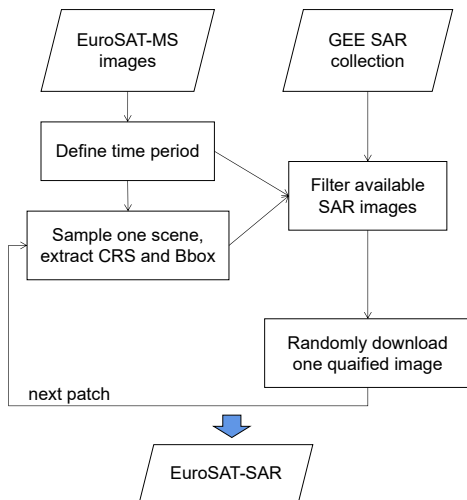


Fig. 7. EuroSAT-SAR creation pipeline.

B. Dataset characteristics

EuroSAT-SAR dataset has 27,000 dual-pol Sentinel-1 GRD images with size 64×64 and two channels VV and VH. There are 10 land cover land use classes, each containing 2000 to 3000 images. To complement the EuroSAT paper, Table X presents the detailed class distribution. Also, sample images are shown in Figure 8.

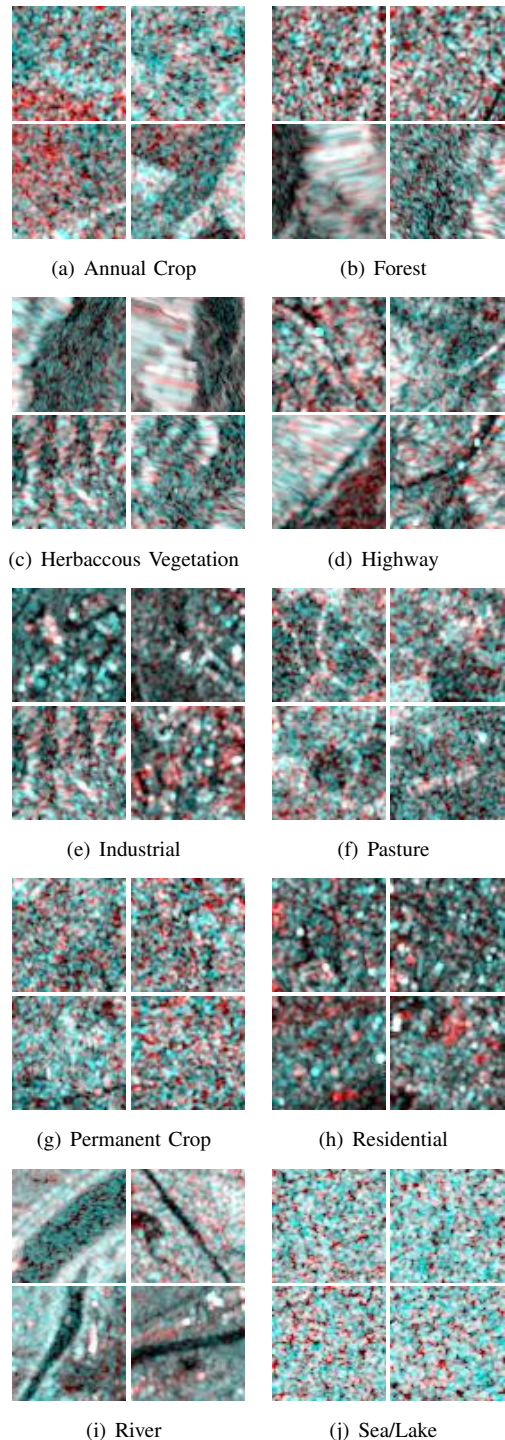


Fig. 8. Sample image patches of all 10 classes covered in the collected EuroSAT-SAR dataset.

TABLE X
EUROSAT-SAR CLASS DISTRIBUTION.

	Annual Crop	Forest	Herbaceous Vegetation	Highway	Industrial	Pasture	Permanent Crop	Residential	River	Sea/Lake
Number of images	3000	3000	3000	2500	2500	2000	2500	3000	2500	3000

Datasheets for EuroSAT-SAR

Here we answer the questions outlined in the datasheets for datasets paper by Gebru et al. [39].

A. Motivation

For what purpose was the dataset created? The dataset was created as a SAR version of the popular EuroSAT dataset to evaluate SAR foundation models.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The dataset was created by the lab "Data Science in Earth Observation" at Technical University of Munich and German Aerospace Center.

Who funded the creation of the dataset? The creation of the dataset was funded by the Helmholtz Association through the Framework of Helmholtz AI.

B. Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? This dataset contains satellite images.

How many instances are there in total (of each type, if appropriate)? The dataset contains 27,000 dual-pol SAR images with size 64×64 .

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The dataset is a sample of all Sentinel-1 satellite images to match the EuroSAT dataset.

What data does each instance consist of? A Sentinel-1 GRD image.

Is there a label or target associated with each instance? Yes, the images are stored in different folders and labels are indicated by the folder names.

Is any information missing from individual instances? No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? Not applicable, geographic location can be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? No. Following EuroSAT dataset which doesn't have an official split, we provide the full dataset in a whole as well. We use our random splits in the benchmarks.

Are there any errors, sources of noise, or redundancies in the dataset? Yes, since we don't have the exact acquisition dates of EuroSAT images, we match them with SAR images in a rough time period assuming no change happened. Though the data looks good with some manual check, we didn't check all the images.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

Does the dataset identify any subpopulations (e.g., by age, gender)? No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? No.

C. Collection process

How was the data associated with each instance acquired? The data was collected from the publicly available Sentinel-1/2 database.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? Google Earth Engine with Python was used to collect the data.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? We sample Sentinel-1 images by matching the geocoordinates and rough acquisition time of the published EuroSAT dataset.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? The data was automatically collected and verified by the authors.

Over what timeframe was the data collected? The data was collected by the authors between February and March 2022. The images within the dataset were captured in the year 2016/2017.

Were any ethical review processes conducted (e.g., by an institutional review board)? No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? The data was collected from open sources.

Were the individuals in question notified about the data collection? N/A.

Did the individuals in question consent to the collection and use of their data? N/A.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? N/A.

D. Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? The data was pre-processed by GEE internally during the collection/downloading process. No further pre-processing was done.

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? No, not necessary.

Is the software used to preprocess/clean/label the instances available? Yes, we use Google Earth Engine with Python which is freely available.

E. Uses

Has the dataset been used for any tasks already? In this paper we use the dataset as a downstream task to evaluate our proposed pretraining algorithms.

Is there a repository that links to any or all papers or systems that use the dataset? Yes we will organize and maintain all related information at <https://huggingface.co/datasets/wangyi111/EuroSAT-SAR>.

What (other) tasks could the dataset be used for? The dataset can be used as a simple, clean SAR scene classification dataset for the remote sensing community, matching the popular multispectral EuroSAT dataset.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? We do not unify the orbiting (ascending/descending) of Sentinel-1 data, which should be taken into consideration for SAR related applications.

Are there tasks for which the dataset should not be used? The authors are not aware of any specific task that should be avoided.

F. Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes, the dataset is publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The dataset will be distributed as tarball. Access to the dataset can be found at <https://huggingface.co/datasets/wangyi111/EuroSAT-SAR>.

When will the dataset be distributed? Starting from July 2023.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? MIT license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

G. Maintenance

Who is supporting/hosting/maintaining the dataset? The dataset is supported and maintained by the authors.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The manager of the dataset can be reached at the email addresses: yi4.wang@tum.de or yi.wang@dlr.de.

Is there an erratum? If errors are found an erratum will be added.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Any updates will be posted and the dataset will be versioned.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? Depending on the updates (if there are), we will either continue hosting the older versions or make a clear update log that older versions can be generated from the newest version.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, please feel free to reach out to us.

H. Author statement of responsibility

The authors confirm all responsibility in case of violation of rights and confirm the licence associated with the dataset.