NeighborMAE: Exploiting Spatial Dependencies between Neighboring Earth Observation Images in Masked Autoencoders Pretraining

Supplementary Material

1. Pretraining Details

We show the pretraining hyperparameters in Table 6, which are based on the original MAE. The training time of the main fMoW experiments on 4 H100s is given in Table 7.

Pretrain Data	fMoW	Satellogic
input size	224	
base lr (batch 256)	1.5e-4	
actual lr	1.2e-3	
lr schedule	cosine decay	
weight decay	0.05	
optimizer	AdamW	
AdamW betas	$\beta_1, \beta_2 = 0.9, 0.95$	
batch size	2048	
augmentation	RandomResizedCrop	
crop scale	(0.2, 1)	
epochs	800	50
warmup epochs	40	2.5
IoU threshold α	0.1	0.0
Lower mask ratio m_1	0.75	
Upper Mask ratio m_2	0.85	

Table 6. Pre-training settings

Model	batch time	wall time
MAE -Large	0.298s	11h 31m 16s
NeighborMAE -Large	0.364s	14h 01m 50s

Table 7. Training time of the main fMoW experiments on 4 H100s.

2. Evaluation Protocols

2.1. Image Classification Details

We show the training settings for image classification in the Table 8, which are based on the evaluation script of the original MAE. We do not use auto augmentation and color jitter as we find them suboptimal for EO classification tasks.

2.2. Semantic Segmentation Details

We show the training settings for semantic segmentation in the Table 9, which are based on the configuration to fine-tune MAE models with UperNet [44] head in MMSegmentation [10].

Experiment setting	linear probing	fine-tuning
input size	224	
base lr (batch 256)	1e-3	1e-3
layer-wise lr decay	N/A	0.75
lr schedule	cosine decay	
weight decay	0	0.05
optimizer	AdamW	
AdamW betas	$\beta_1, \beta_2 = 0.9, 0.95$	
batch size	depends on datasets	
warmup epochs	1	1
epochs	20	20
augmentation	RandomResizedCrop	
crop scale	(0.08, 1)	
label smoothing	0	0.1
mixup	0	0.8
cutmix	0	1.0
drop path	0	0.2
global pooling	cls token	average

Table 8. Image classification settings used in evaluation protocols.

Experiment setting	frozen backbone	fine-tuning
input size	depends on datasets	
base lr (batch 256)	1e-4	
layer-wise lr decay	N/A	0.75
lr schedule	cosine decay	
weight decay	0	0.05
optimizer	AdamW	
momentum	$\beta_1, \beta_2 = 0.9, 0.999$	
batch size	16	
warmup iterations	1000	
iterations	20000	
augmentation	RandomScale, RandomCrop	
scale range	(0.5, 2.0)	
layers for feature pyramid	7, 11, 15, 23	
inference	sliding windows on original images	

Table 9. Semantic segmentation settings used in evaluation protocol.

3. Adaptations for Baseline Models

We use the publicly available model weights of baseline models, except for our reproduced MAE on EO datasets. All models are based on ViT-Large-16 with potential subtle differences. We upsample the positional embedding of

Cross-Scale MAE [35] to an input size of 224 since it was trained on an input size 128. For the temporal variant of SatMAE [9], we replace its timestamp embedding with a learnable positional embedding as timestamps are unavailable in downstream tasks. We use B-G-R input according to the pretraining of SatMAE++ [29]. The input for all baseline models is normalized by the $mean$ and std used in their original pretraining.

4. Visualization

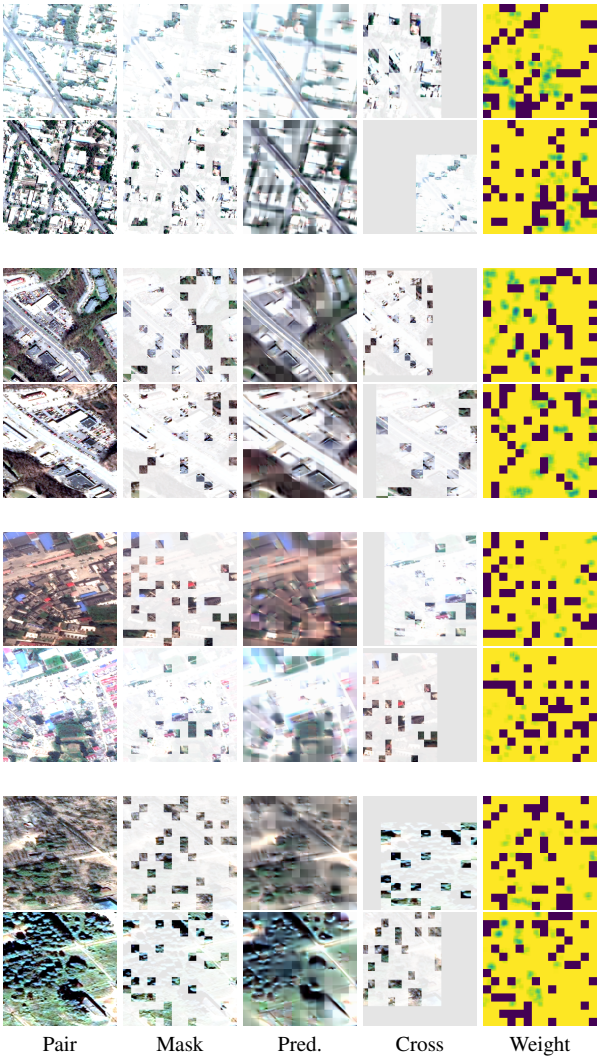

Figure 5. Visualization of the reconstruction of neighboring images from fMoW-RGB. From left to right, we show pairs of neighboring images, masked images, prediction, cross-visible pixels, and the loss weight. Neighboring images from fMoW-RGB usually exhibit significant temporal changes and therefore our loss weighting by cross visibility has less impact.

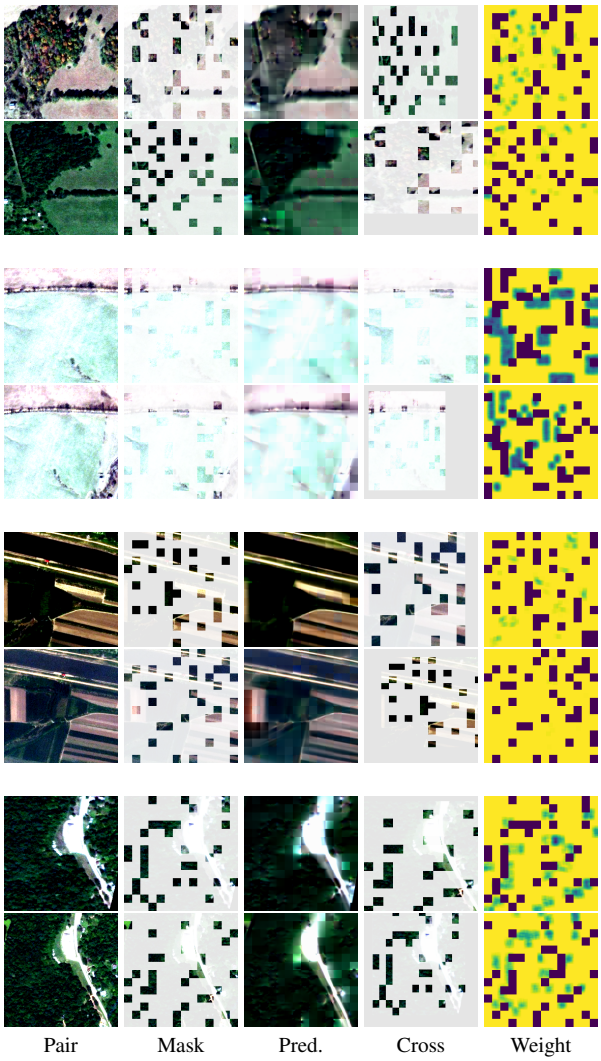

Figure 6. Visualization of the reconstruction of neighboring images from Satellogic-RGB. From left to right, we show pairs of neighboring images, masked images, prediction, cross-visible pixels, and the loss weight. Neighboring images from Satellogic-RGB present fewer changes (fewer and even no revisits from the data). are even identical images. Our loss weighting by input visibility are more effective on Satellogic-RGB to avoid short learning shortcuts.

4.1. Reconstruction

We visualize the reconstruction of neighboring images from fMoW-RGB [8] and Satellogic-RGB [41] in Figure 5 and 6. fMoW-RGB shows more temporal changes and semantic contents in images and Satellogic-RGB has fewer revisits and its contents are not semantic-aware. Therefore, our loss weighting by cross visibility can alleviate the information leak when training on Satellogic-RGB and improve performance.

4.2. Attention Map

We show the attention map associated with a specific patch when unmasked neighboring images are fed to the learned models in 7. The corresponding regions from neighboring images receive high attention scores, which indicates that NeighborMAE can learn spatial dependency from neighboring images.

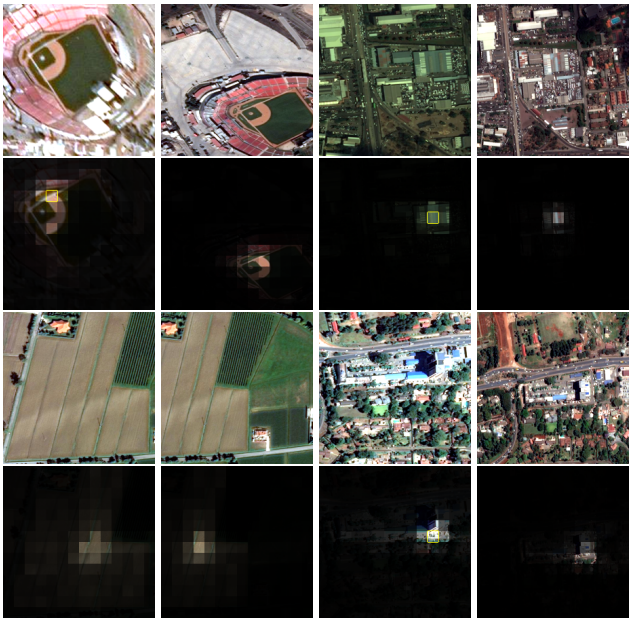

Figure 7. Visualization of the attention of neighboring images. The first row shows the input neighboring images and the second row shows the attention scores with respect the to the patch marked by yellow boundary in the first image. The visualization indicates that NeighborMAE can build spatial relations across neighboring images and learn correlated representations.

4.3. Spatial and Temporal Distributions of Datasets

Examples of images from the used fMoW [8] and Satellogic [41] datasets for pretraining are displayed in Figure 8 and Figure 9. fMoW and Satellogic present different spatial and temporal characteristics, and NeighborMAE is robust and achieves significant improvement on both datasets.

5. Statistical Significance

To assess whether the performance improvement by NeighborMAE is statistically significant, we perform 5 independent runs of consecutive SSL pretraining and evaluation on fMoW, following the same setting as the main experiment in 1. From table 10, we can see that the standard deviations are relatively small for both methods, and the mean difference between the baseline MAE and our NeighborMAE is large enough. This suggests that the observed performance gain is unlikely to be due to random variation and can be attributed to the effectiveness of our method.

Figure 8. We show spatial and temporal distributions of fMoW. Pixel values are averaged over overlapping regions. The image sizes are mostly different in fMoW and the images represent specific areas of interest. fMoW features organized temporal sequences captured by multiple EO projects.

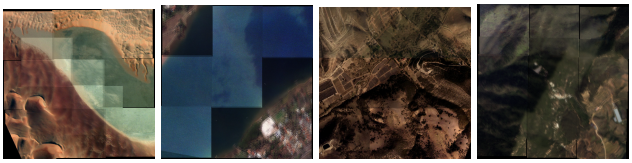

Figure 9. We show spatial and temporal distributions of Satellogic. Pixel values are averaged over overlapping regions. The image sizes are consistent in Satellogic and the image contents are not semantic-aware. Few or no revisits are provided for specific locations.

Method	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	mean±std
MAE	67.19	66.99	66.71	66.56	65.99	66.69±0.41
NeighborMAE	69.42	69.33	68.90	68.61	68.08	68.87±0.49
MAE	78.39	78.22	78.13	78.07	77.91	78.14±0.16
NeighborMAE	79.53	79.53	79.26	79.24	79.11	79.33±0.17

Table 10. Performance of 5 independent runs of consecutive SSL pretraining and evaluation on fMoW, using baseline MAE and our NeighborMAE with the main experiment setting in 1. The upper and lower part shows the accuracy of linear probing and finetuning, respectively. We rank the performance from left to right. The results demonstrate that the improvement achieved by NeighborMAE is beyond the deviation caused by randomness.