Scale-MAE: Supplementary Material

1. Datasets

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

1.1. Diversity in classes

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

Dataset	Resolution (px)	GSD (m)	Number of Images	Number of Classes	Task Type
AiRound [9]	500	0.3 - 4800	11,753	11	C
CV-BrCT [9]	500	0.3 - 4800	24,000	9	C
EuroSAT [8]	64	10	27,000	10	C
MLRSNet [10]	256	0.1 - 10	109,161	46	C
Optimal-31 [12]	256	0.5 - 8	1,860	31	C
RESISC-45 [3]	256	0.2 - 30	31,500	45	C
UC Merced [14]	256	0.3	2,100	21	C
WHU-RS19 [6]	256	0.5	1050	19	C
fMoW [4]	Various	0.3	1,047,691	62	C
SpaceNet v1 [11]	Various	0.5	6,940	2	SS

Table 1. Statistics of all datasets used in our experiments. Task types are classification (C) and semantic segmentation (SS).

1.2. Spectrum of GSDs

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

1.3. Quality control

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

		$k = 20$			$k = 100$			$k = 5$		
Dataset	**Res**	**Scale.**	**Sat.**	**Conv.**	**Scale.**	**Sat.**	**Conv.**	**Scale.**	**Sat.**	**Conv.**
AiRound	16	0.401	0.375	**0.423**	0.396	0.367	**0.401**	0.370	0.355	**0.403**
	32	**0.561**	0.510	0.539	**0.536**	0.491	0.517	**0.541**	0.492	0.539
	64	**0.689**	0.607	0.658	**0.643**	0.579	0.621	**0.692**	0.604	0.666
	128	**0.743**	0.650	0.681	**0.690**	0.600	0.622	**0.749**	0.660	0.690
	256	**0.729**	0.662	0.658	**0.678**	0.621	0.602	**0.731**	0.663	0.676
	496	**0.670**	0.664	0.620	0.609	**0.613**	0.566	**0.685**	0.669	0.632
CV-BrCT	16	0.522	0.478	**0.567**	0.485	0.443	**0.513**	0.524	0.475	**0.585**
	32	0.653	0.615	**0.656**	0.588	0.560	**0.592**	0.695	0.644	**0.699**
	64	**0.744**	0.701	0.711	**0.674**	0.635	0.644	**0.780**	0.727	0.754
	128	**0.763**	0.725	0.732	**0.710**	0.662	0.667	**0.805**	0.758	0.782
	256	**0.761**	0.725	0.727	**0.694**	0.666	0.664	**0.802**	0.770	0.771
	496	**0.737**	0.727	0.709	0.656	**0.657**	0.631	**0.792**	0.771	0.765
EuroSAT	16	0.744	0.727	**0.826**	0.699	0.695	**0.788**	0.751	0.729	**0.835**
	32	**0.901**	0.876	0.898	**0.869**	0.854	0.863	**0.912**	0.871	0.909
	64	**0.956**	0.931	0.940	**0.935**	0.913	0.914	**0.960**	0.934	0.947
MLRSNet	16	0.563	0.491	**0.607**	0.535	0.461	**0.549**	0.551	0.479	**0.617**
	32	**0.772**	0.677	0.744	**0.726**	0.625	0.688	**0.772**	0.684	0.762
	64	**0.893**	0.815	0.851	**0.849**	0.754	0.792	**0.911**	0.839	0.876
	128	**0.936**	0.875	0.894	**0.892**	0.814	0.834	**0.950**	0.899	0.918
	256	**0.918**	0.892	0.882	**0.862**	0.840	0.817	**0.940**	0.913	0.910
OPTIMAL-31	16	0.354	0.322	**0.439**	0.312	0.298	**0.370**	0.317	0.319	**0.418**
	32	0.574	0.500	**0.587**	**0.567**	0.508	0.545	**0.565**	0.519	0.561
	64	**0.793**	0.609	0.698	**0.742**	0.561	0.598	**0.782**	0.646	0.688
	128	**0.816**	0.670	0.714	**0.731**	0.646	0.595	**0.809**	0.694	0.725
	256	**0.739**	0.681	0.646	**0.653**	0.638	0.550	**0.761**	0.731	0.693
RESISC	16	0.382	0.347	**0.458**	0.370	0.327	**0.428**	0.353	0.323	**0.435**
	32	**0.628**	0.527	0.601	**0.597**	0.505	0.568	**0.609**	0.508	0.592
	64	**0.798**	0.667	0.731	**0.754**	0.631	0.677	**0.803**	0.667	0.734
	128	**0.864**	0.748	0.798	**0.819**	0.699	0.743	**0.882**	0.762	0.817
	256	**0.826**	0.758	0.762	**0.761**	0.708	0.690	**0.850**	0.771	0.788
UC Merced	16	0.524	0.472	**0.598**	0.400	0.370	**0.462**	0.512	0.488	**0.617**
	32	**0.767**	0.670	0.683	**0.605**	0.535	0.593	**0.828**	0.682	0.726
	64	**0.842**	0.795	0.771	0.719	**0.729**	0.652	**0.884**	0.842	0.845
	128	**0.858**	0.788	0.750	0.662	**0.738**	0.655	**0.884**	0.847	0.838
	256	0.762	**0.802**	0.700	0.595	**0.757**	0.590	**0.851**	0.842	0.817
WHU-RS19	16	0.545	0.445	**0.576**	0.400	0.380	**0.562**	0.525	0.490	**0.631**
	32	0.650	**0.729**	0.670	0.610	**0.675**	0.576	**0.760**	0.690	0.754
	64	**0.850**	0.805	0.833	**0.770**	0.730	0.680	**0.920**	0.840	0.837
	128	**0.970**	0.910	0.882	**0.890**	**0.890**	0.685	**0.985**	0.895	0.941
	256	**0.960**	0.940	0.892	0.880	**0.925**	0.709	**0.975**	0.945	0.931

Table 2. ***Scale-MAE* outperforms SatMAE and ConvMAE on kNN classification across a variety of k, across a variety of resolutions.** kNN Classification results for *Scale-MAE*, SatMAE and ConvMAE across a variety of k. Resolution is reported in pixels.

2. Laplacian and Upsampling Block Architectures

Figure 1 illustrates the architecture of Laplacian and Upsampling block architectures described below.

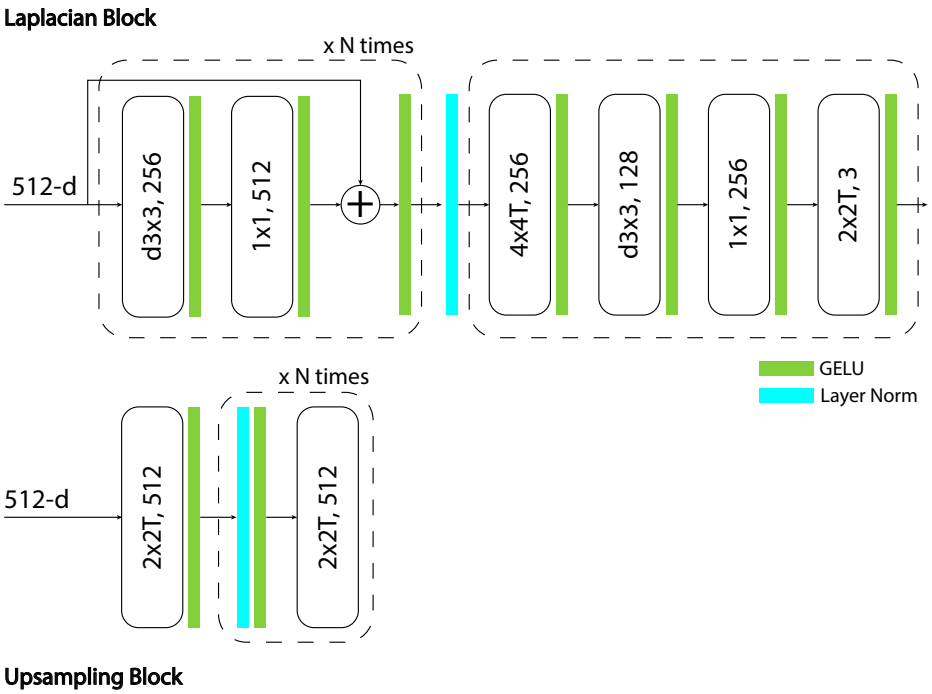

Figure 1. (top) The Laplacian Block (LB) is a fully convolutional architecture consists of a chain of Feature Mapping Block followed by one final Reconstruction Block. (bottom) The UpSampling Block (UB) consists of a series of transpose convolution layers separated by LayerNorm and GELU activation.

2.1. Laplacian Block

Laplacian Blocks are used to reconstruct the target at a specific resolution and frequency. A Laplacian Block consists of a chain of Feature Mapping Block, which distills information at a specific frequency, followed by one final Reconstruction Block, which generates the final output. A Feature Mapping Block consists of a 3x3 depth-wise convolution layer with GELU activation, followed by 1x1 convolution. A Reconstruction Block consists of a 4x4 transpose convolution layer followed by a 3x3 depth-wise convolution layer, a 1x1 convolution layer, and a 2x2 transpose convolution layer. In our experiments, we have two Feature Mapping Blocks per Laplacian Block.

2.2. Upsampling Block

Upsampling Blocks are used to upsample the feature map to a higher resolution. It consists of a series of 2x2 transpose convolution layers with LayerNorm and GELU activation between them. The number of such transposed convolution layers are a function of the output and input resolution. This is a progressive process in which we repetitively upsample the feature map by a factor of 2 until we reach the desired target resolution. Figure 1 illustrates the architecture of these two blocks.

3. Evaluation Details

As discussed in the main experimental section, we investigated the quality of representations learned from *Scale-MAE* pretraining through a set of experiments that explore their robustness to scale as well as their transfer performance to additional tasks. We provide more information and details on these evaluations here. In order to compare with SatMAE [5] and ConvMAE [7], for our main experiments, we pretrained *Scale-MAE* with a ViT-Large model using the Functional Map of the World (FMoW) RGB training set, which consists of 363.6k images of varying image resolution and GSD. The initial higher resolution image I_{hr} is taken as a random 448px^2 crop of the input image, and the input image I is then a downsampled 224px^2 from I_{hr}. The low frequency groundtruth is obtained by downscaling I_{hr} to 14px^2 and then upscaling to 224px^2, while the high frequency groundtruth is obtained by downscaling I_{hr} to 56px^2 and then upscaling to 448px^2 and subtracting this image from I_{hr}. This is a common method for band pass filtering used in several super resolution works, where a high to low to high resolution interpolation is used to obtain only low frequency results, and then high frequency results are obtained by subtracting the low frequency image.

As further discussed in the main experimental section, we evaluate the quality of representations from *Scale-MAE* by freezing the encoder and performing a nonparametric k-nearest-neighbor (kNN) classification with eight different remote sensing imagery classification datasets with different GSDs, none of which were encountered during pretraining. All kNN evaluations were conducted on 4 GPUs. Results are in Table 2. The kNN classifier operates by encoding all train and validation instances, where each embedded instance in the validation set computes the cosine distance with each embedded instance in the training set, where the instance is classified correctly if the majority of its k-nearest-neighbors are in the same class as the validation instance. The justification for a kNN classifier evaluation is that a strong pretrained network will output semantically grouped representation for unseen data of the same class. This evaluation for the quality of representations occurs in other notable works [1, 2, 13].

4. Visualization of SpaceNet Segmentation

Figure 2 shows an additional set of segmentation examples comparing *Scale-MAE* and vanilla MAE pre-trained on FMoW and finetuned on SpaceNet v1. The left, center, right columns are ground truth labels, *Scale-MAE* and vanilla MAE respectively. The top row shows a 0.3m GSD image and the bottom row shows a 3.0m GSD image. As shown in the figure, *Scale-MAE* performs better at both higher and lower GSDs.

5. Glossary

5.1. Ground sample distance

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

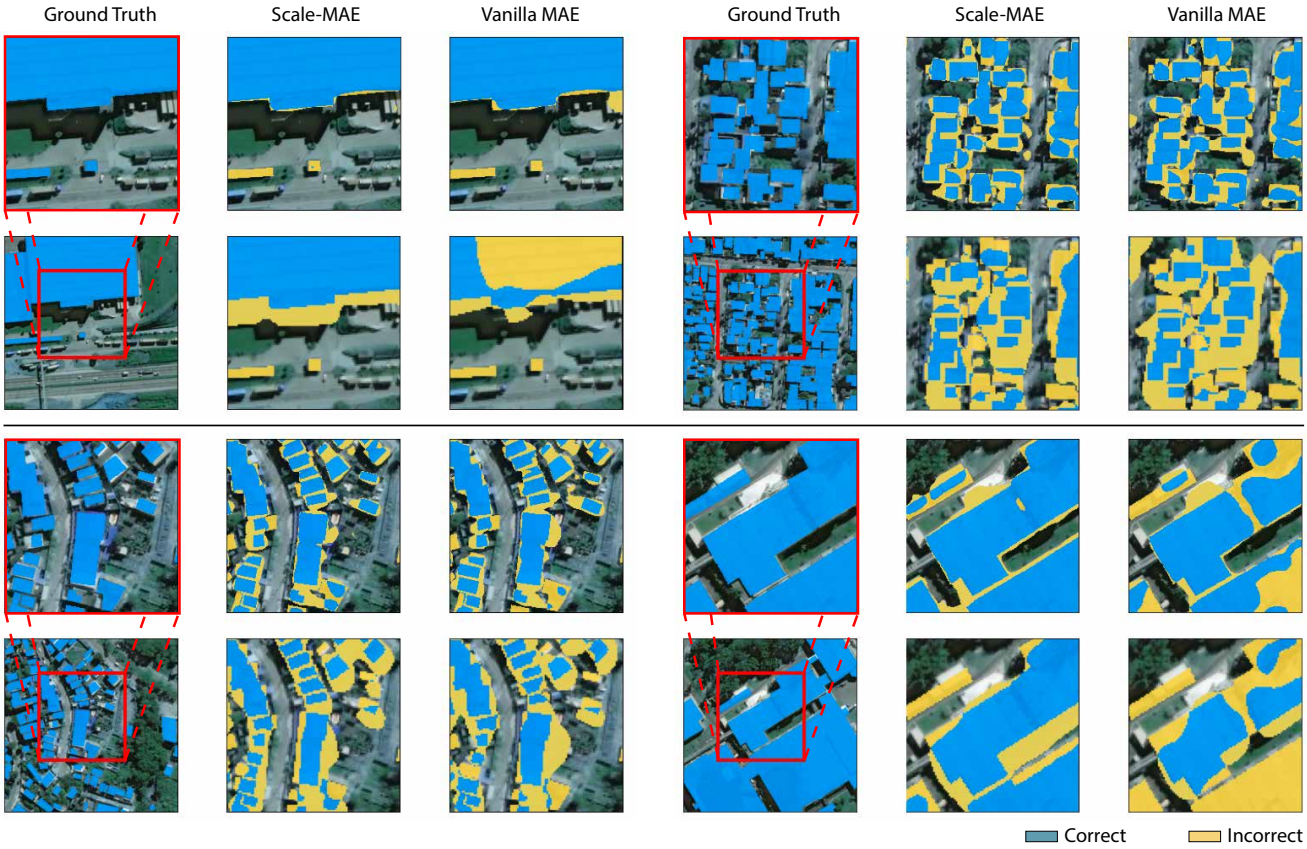

Figure 2. Visualization of Segmentation Results on SpaceNet. The left, center, right columns are ground truth labels, *Scale-MAE* and vanilla MAE, respectively. The top row shows a 0.3m GSD image and the bottom row shows a 3.0m GSD image. As shown in the figure, *Scale-MAE* performs better at both higher and lower GSDs.

References

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

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

[3] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10):1865–1883, Oct. 2017.

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

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

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

[7] Peng Gao, Teli Ma, Hongsheng Li, Ziyi Lin, Jifeng Dai, and Yu Qiao. ConvMAE: Masked Convolution Meets Masked Autoencoders, May 2022.

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

[9] 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.

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

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

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

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

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