GEOSPATIAL ALGORITHM FOR MULTI-MODAL APPLICATIONS

Martin Weiss

1 INTRODUCTION

Multi-modal neural models trained on web-scale datasets of natural images and text have produced impressive results in applications ranging from novel image synthesis to image-and-text dialogue. While these models are more flexible and general than existing ImageNet models, the training data is not indexed on a spatiotemporal basis. This limits the ability of the model to contextualize new data, and to understand the physical processes that enable life and civilization on earth.

Though the amount of geospatial data is growing exponentially, far fewer artificial intelligence models have been trained on this data. One reason that so few neural models have been designed for geospatial data is that geospatial data is heterogenous. For example, multispectral sensors measure electromagnetic radiation in a small number of spectral bands, with large variance in the exact number of bands, spectral resolution, and spatial resolution. Hyperspectral sensors can capture hundreds of spectral bands. Both are passive electromagnetic sensors, but the differing number of spectral channels creates a major challenge for those wishing to use a convolutional neural network pre-trained on data with only three spectral channels (RGB) such as ImageNet (Deng et al., 2009). Furthermore, orbiting sensors for earth observation also include active sensors like Synthetic-aperture radar (SAR), and highly specialized sensors used for a narrow range of tasks, such as the methane sensing interferometer developed by GHGSat (Jervis et al., 2021). And geospatial datasets are not restricted to orbiting sensors, there are huge crowd-sourced geospatial text databases (e.g., parts of Wikipedia) and land cover segmentation masks from Open Street Map (OpenStreetMap contributors, 2017).

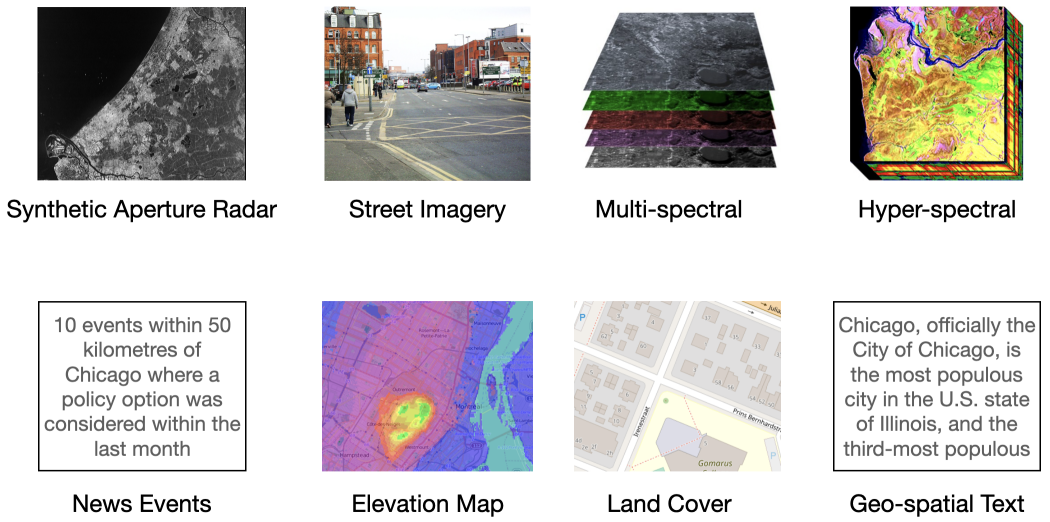

Figure 1: **Heterogenous Geospatial Data Modalities**. Geospatial data is very diverse, and includes active sensors (SAR), passive sensors (natural images, multispectral, and hyperspectral), derived data (elevation maps), and human annotated information (land cover labels, geo-spatial wikipedia, news events).

In this work, we propose the Geospatial Algorithm for Multi-Modal Applications (GAMMA), an architecture that incorporates geospatial inductive biases while preserving the flexibility to process heterogenous data. The geospatial inductive bias is the assumption that every data sample was captured at some time and place, and therefore describe various observations of the underlying

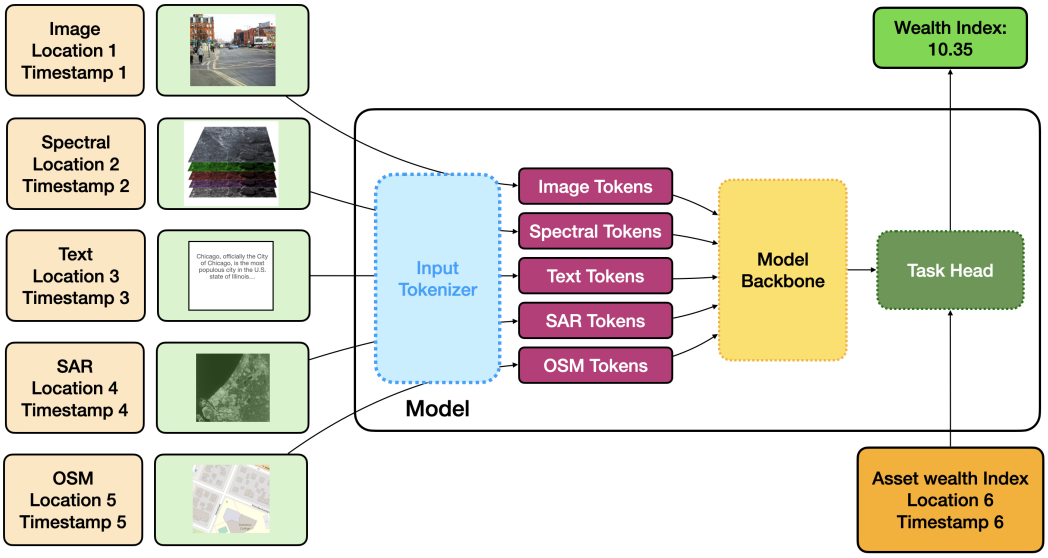

Figure 2: GAMMA Architectural Overview. GAMMA is composed of three primary components, the input tokenizer which transforms a variety of input keys (modality type, spatial and temporal information) and values (the captured data) into a common token encoding. These tokens are then processed and fused by the model backbone. Finally, an input query is provided to the task head with the backbone's output representation to make a prediction.

partially observable system, earth. Data samples processed by GAMMA include spatial and temporal information, spectral information with arbitrary bands and resolutions, text, natural images, SAR data, segmentation masks, and bounding boxes. When coupled, these diverse data modalities and geospatial inductive biases will enable GAMMA to ground new data into a global representation that has qualitatively different biases than a model trained on web data. Further, by strongly indexing on location and time, the model will be able to make predictions about the real world. We intend to pre-train GAMMA with a masked prediction approach to a wide variety of geospatial data, and to finetune to SustainBench and Climate Change Benchmark tasks. We anticipate that the pre-training will demonstrate strong low-shot adaptation performance when compared to state-of-the-art models.

Figure 2 shows the architecture with its three core components: the **tokenizer**, **backbone**, and **task head**. The tokenizer takes in geospatial data and projects it into a common encoding format for the backbone to read in. The backbone is a typical neural network that can process spatial or tokenized data, such as a ConvNeXt (Liu et al., 2022) or Perceiver (Jaegle et al., 2021b). The task head takes in processed representations from the backbone and a query to produce an output. During self-supervised training, we use a modality-specific decoder conditioned on mask tokens and the backbone representation, then compare the output with the true data.

Our contributions are as follows. **(a)** We propose GAMMA, a general-purpose neural architecture with geospatial inductive biases that processes the data modalities commonly found in geospatial datasets. **(b)** We will quantitatively evaluate the model on a masked reconstruction pretext task and perform downstream evaluations on SustainBench (Yeh et al., 2021) and (unreleased) Climate Change Benchmark from ElementAI. We hypothesize that pre-training the model on a diverse set of self-supervised learning tasks will increase its performance on these important benchmarks. **(c)** We propose a data-pruning approach to make the self-supervised learning procedure more computationally efficient.

2 Geospatial Algorithm for Multi-Modal Applications

Modality Encoding and Tokenization. Given the diverse modalities of geo-spatial data, a key design decision is how to encode it as input for the neural network. For example, how should we encode a

multi-spectral satellite image with its associated location? For the vast majority of data, the associated spatial location is represented as a latitude and longitude along with some uncertainty. To encode this as a **geospatial positional encoding** we cannot directly use a Fourier Positional Encoding scheme (similar to Vaswani et al. (2017a)) because longitude has a discontinuity along a line on the earth at the +/-180 degree meridian. This can cause issues for optimization algorithms and when calculating displacement. One solution is to first produce a non-singular n-vector encoding (Gade, 2010) of the latitude and longitude, then Fourier encode the resulting 3-vector. This representation is still biased in some ways, and many alternatives should be considered (Mai et al., 2022). Additionally, we want to encode spatial and spectral resolution in such a way that samples with different resolutions lie in the same space and can be easily fused. To achieve this, we intend to integrate over the Fourier positional encoding based on the input resolution. The **multispectral encoding** will be constructed by first reshaping the multispectral image and passing each spectral channel through a Fourier Positional Encoding based on its spectral band location. To encode the spectral resolution, we integrate over this positional encoding. We show the general idea in Figure 3.

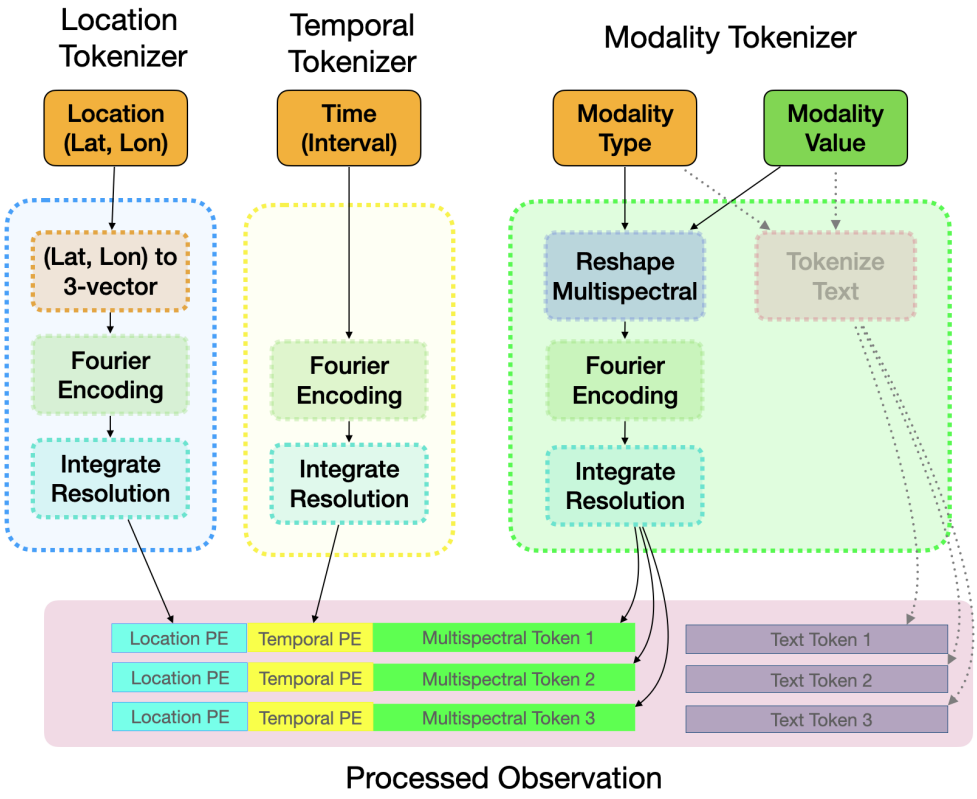

Figure 3: **Multispectral Modality Encoding**. The processed observation of a multispectral observation is a sequence (in raster order) of tokens. Each token is a concatenation of a Fourier encoded location and time with a normalized multi-spectral patch. In this example, the modality type would be a multispectral object type defining the number of spectral channels and their radiometric resolution as well as the image dimension. This type information determines which modality tokenizer to apply, and parameterizes its Fourier positional encoding. We grey out the text tokenizer given that this observation does not contain text.

Each geospatial datum is associated with a non-uniformly sampled spatiotemporal region. Simply, all the data that we consider is captured at some time and place. We assert that we know the time and place (with some uncertainty) that this data describes, but that it is not necessarily sampled on a grid. One might assume that a satellite image would correspond to a spatial uniform grid-sampling of electromagnetic radiation from the ground, but in reality orbital push-broom and whisker-broom sensors capture the pixels of an image sequentially and each pixel has a corresponding ground-sample distance (GSD) that depends on its distance from the flight path of the satellite and the local ground

topology. As a result, assuming that learned features should be translationally equivariant (as is done in convolutional neural networks) would be sub-optimal in many cases. For another example, some UAVs capture both hyperspectral and LIDAR data to construct hyperspectral point clouds (López et al., 2022), and the (geospatially located) Wikipedia article that describes the city of Chicago in many ways is also describing a non-convex 227.73 square mile area. We can treat images as a special case of point clouds (with a grid-like arrangement), so we can process the information associated with each point and its spatiotemporal region. However, given the extremely high dimensionality of satellite imagery, processing each spectral pixel as its own token through the backbone quickly becomes computationally intractable. In order to process all these data types in a single model, we will investigate mechanisms to downsample the input in an off-grid way while preserving important information. One necessary component is a model for processing point clouds. One inspiration is the PointNet family of models which includes PointNet (Qi et al., 2016), PointNet++ (Qi et al., 2017), and PointNeXt (Qian et al., 2022). These model the permutation invariance of points with shared MLPs that restrict feature extraction to be pointwise and use hierarchy to capture local geometric structures. Therefore, we could use PointNeXt to process the collection of points for a spatiotemporal region in a permutation invariant way. In Figure 4, we represent these spatiotemporal "point patches" with blue circles and show the embedding procedure for two images that have different spatial resolutions.

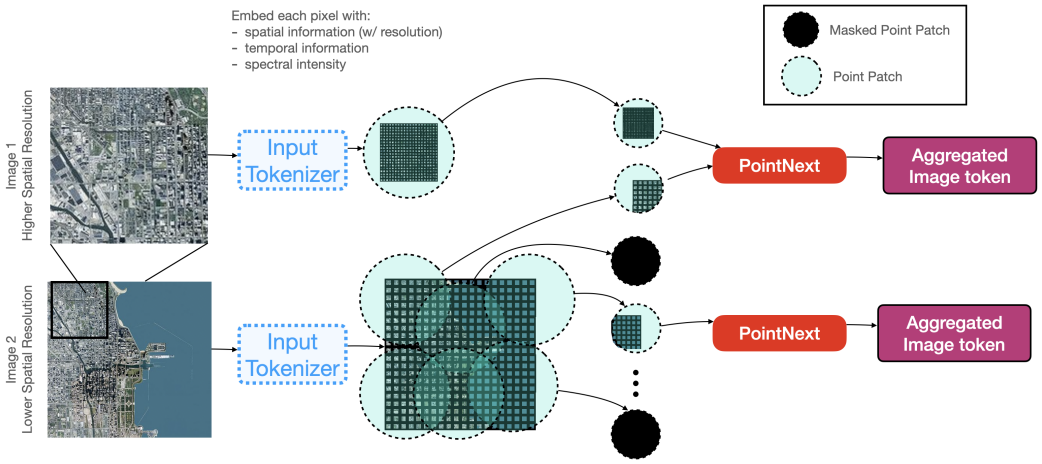

Figure 4: **Image point patch encoding**. In this example, we are tokenizing two multi-spectral images of the same region with different spatial resolutions (though they could have different spectral channels, different spectral resolutions). We take a fixed-size point patch and sample different locations. All the pixels of the higher resolution image, covering less ground cell area, is entirely covered by a single point patch. The point patch corresponding to the same area on the lower resolution image captures fewer pixels. Each pixel is tokenized by the input tokenizer following Figure 3, then the encoded pixels are processed in a permutation invariant way by PointNext to produce a single spectral token for the area covered by that point patch. This point patch has information about its spatiotemporal centroid in addition to the aggregated spectral information.

Backbones. Given the broad diversity of input modalities, and sparsity of spatio-temporal data coverage, the number of processed observation tokens will be very large to accommodate every modality. To manage this large number of tokens, the backbone cross attends to the processed observation tokens, re-sampling them to a smaller and more compact cardinality. Following this, the backbone can process this latent representation with self-attention (Vaswani et al., 2017b). In previous sections, we have mentioned the possibility of using a ConvNeXt (Liu et al., 2022) backbone. This is interesting as a baseline model, but mostly raises problems. For example, the convolutional backbone will struggle to fuse multi-modal information such as text or multi-resolution imagery.

Task Head. The task head takes as input the encoded representation produced by the backbone, and an output query. The output query determines which information to extract and how to format it as output. The simple way to implement the task head is that the input query indexes into an array of 'sub task heads' and this custom head is used to produce an output of the requested dimensionality. While a necessary baseline, this approach suffers from several issues. First, if there are two classification

tasks with different numbers of classes, they live in different spaces and it will be more difficult to have cross-task interference. An alternative approach would be to develop a unified input-output space. For example, Pix2Seq proposes a method of training a model to bounding boxes as sequences of tokenized outputs using a discretized output space (Chen et al., 2021). Similarly, Unified-IO uses a SentencePiece tokenizer (Kudo & Richardson, 2018) and patch embedded natural images with a VQ-VAE discretizing bottleneck to produce a tokenized input sequence for a T5 transformer encoder and decoder model (Raffel et al., 2019). The decoded output tokens are then passed to either the SentencePiece tokenizer to produce a text output, or the VQ-VAE decoder to produce an image output. Flamingo (Alayrac et al., 2022) offers yet another approach by using a language model as a sort of language API that can be modulated with a Perceiver re-sampler (Jaegle et al., 2021a) and frozen image vision encoder.

Training and Prioritized Sampling. Many earth observation tasks can be easily parallelized. However, the amount of computation required to extract insight on a national or global scale is considerable. Further, self-supervised pre-training requires massive amounts of data. Recent work has shown that the use of a high-quality data pruning metric can reduce the power law scaling of self-supervised pre-training methods to exponential scaling (Sorscher et al., 2022). If this result applies to geospatial data, then the question becomes: how do we determine which observations to use for self-supervised training in geo-spatial datasets? One approach which we can try is influenced by prioritized experience replay (Schaul et al., 2015). By maintaining a running average of the SSL loss we can determine which areas are more complex (having a higher density of challenging-to-predict data) and should be resampled.

More broadly, we intend that our pre-training pipeline to roughly follow Point-MAE (Pang et al., 2022) which proposes a self-supervised method of point cloud patch reconstructions. For our work, we will mask spatiotemporal point regions and predict these based on unmasked regions. More specifically, the decoder ingests query vectors that contain the spatial, temporal, and modality information for the masked regions. Figure 5 shows this process.

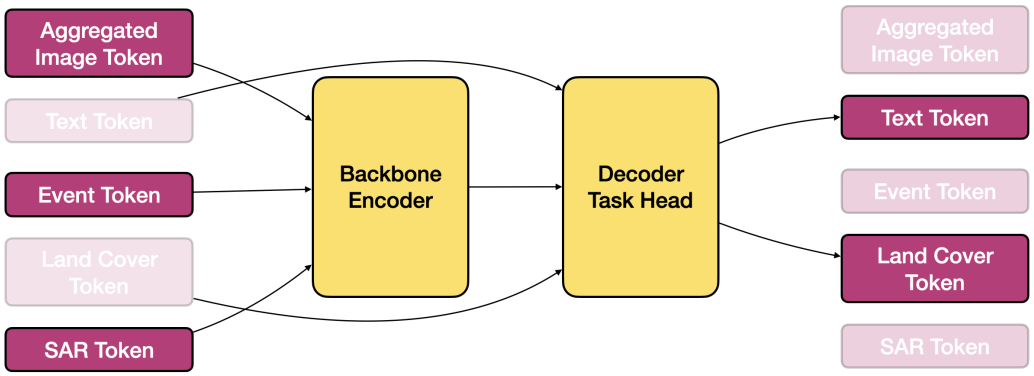

Figure 5: **Geospatial Reconstruction Task**. The reconstruction task takes input tokens (each with its own spatial and temporal positional information) of different modalities, and feeds an unmasked subset of these to the backbone encoder. The decoder task head then takes in the spatiotemporal positional embedding and modality type information (e.g., text, land cover segmentation) to predict the masked tokens.

3 Related Work

Self-supervised Learning. Self-supervised pre-training on web-scale text became popularized with models such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018). More recently, large-scale image-captioning datasets like LAION-400M (Schuhmann et al., 2021) have been collected and used to train multi-modal models such as CLIP (Radford et al., 2021) and DALL·E-2 (Ramesh et al., 2022). These models are built using self-attention (Vaswani et al., 2017a) and learn a joint representation space for text and images. These efforts have yielded large-scale foundation models (Bommasani et al., 2021) with joint text and natural image representations. However, these models

do little to address geospatial data and tasks (Lacoste et al., 2021). And existing models do not use well-aligned multi-modal representations. For example, the model (Sheehan et al., 2019) trained on a Demographic and Health Surveys (DHS) prediction task, where the model takes as input a location, nightlights imagery, and nearby encoded Wikipedia articles. The encoder is simply Doc2Vec (Le & Mikolov, 2014) pre-trained on the global geolocated Wikipedia article corpus and fixed during training.

Multi-task Learning. There are several advantages of multi-task learning with a unified output space, including **(a)** reducing the need for hand crafting models with appropriate inductive biases for each domain, **(b)** increasing the amount and diversity of training data, **(c)** and allowing for a more rigorous understanding and comparison between methods. Many unified frameworks for modelling text have been proposed (Radford et al., 2019; McCann et al., 2018; Keskar et al., 2019) including T5 (Raffel et al., 2019) which proposed to treat every text processing problem as a "text-to-text" problem, rather than text to classification, regression, or text. More recently, models like GPV-2 (Kamath et al., 2022) and Unified-IO (Lu et al., 2022) have combined a pre-trained T5 with an object detector (or VQ-VAE in the case of Unified-IO) to construct a single model which can be applied to tasks including visual question answering, captioning, object localization, image classification and region classification. Similarly, DeepMind's Gato (Reed et al., 2022) unifies observations from a diverse set of reinforcement learning environments that include text, images, discrete action values, and continuous values (such as proprioceptive inputs) into a homogenous set of tokens.

4 EXPERIMENTS

Self-Supervised Pretraining. Masked auto-encoders (MAE) (He et al., 2021) are a simple method of denoising auto-encoding (Bengio et al., 2013) that is an effective method of self-supervised pre-training for neural image models. By masking and inferring the masked patches of the natural images in ImageNet (Deng et al., 2009), MAEs learn a strong image representation. While this method of pre-training should be used as a baseline method, we do not believe that it will be effective without modification in the multi-modal geospatial setting for several reasons. **(1)** The data distribution of RGB satellite image data is different than the distribution of ground-based RGB images, therefore, the masking ratio is likely to be different. **(2)** Most satellite-based image sensors capture a broader range of spectral data than RGB cameras, and as a result channel masking may be more effective. **(3)** The geospatial inductive bias must be taken into account, given that some regions (e.g. U.S. energy facilities) will have much higher data coverage (e.g. task satellite imagery) than other regions. **(4)** Some data modalities will be local (e.g., patches of a satellite image will be informative of nearby patches) while other data modalities are non-local (such as the war in Ukraine affecting energy prices in Germany).

In order to handle the biased data coverage issue, we could introduce a form of importance sampling based on the inverse of coverage. For example, if the data gathered about Rwanda can be compressed into fewer bits of data than NYC, we can sample more spatially to compensate for the fact we don't have as much data there. This approach is based on the assumption that there is uniform coverage of information across the world (i.e., it should be possible to predict the wealth index for any area on earth).

Multi-task Finetuning. We intend to fine-tune GAMMA on a diverse set of geospatial tasks from two primary benchmarks. First, we will use the upcoming Climate Change Benchmark developed by Alex Lacoste's team at ElementAI. Second, **SustainBench** (Yeh et al., 2021) provides a collection of 15 benchmark tasks across 7 United Nations Sustainable Development Goals (SDGs). These include tasks related to economic development, agriculture, health, education, water and sanitation, climate action, and life on land. Collectively, the SustainBench datasets include data covering 119 countries and 24 years (from 1996-2019), with satellite images, street-level images, and time series data. We reproduce Table 2 from the SustainBench paper to underscore the diversity of the tasks.

In conclusion, we intend to build GAMMA, a multi-modal neural architecture for geospatial tasks. We believe that through large-scale self-supervised pre-training on multi-modal datasets with geospatial inductive biases, GAMMA can dramatically outperform existing machine learning models across many geospatial tasks including SustainBench (Yeh et al., 2021) and the Climate Change Benchmark. However, there are several significant risks with this project. First and foremost, this approach will require a very large amount of compute. Second, by opting for general-purpose models we should be

SDG	Task	Countries	Metric	Benchmark Value	Model Type	Ref
No Poverty	Poverty prediction over space Poverty prediction over time	48 countries 5 African countries	r^2 r^2	0.63 0.35*	kNN ResNet-18	[109] [109]
Zero Hunger	Weakly supervised cropland classification	United States	F1 score	0.88 (pixel label) 0.80 (image label)	U-Net	[102]
	Crop type classification	Ghana, South Sudan Kenya	Macro F1 Macro F1	0.57, 0.70 0.30	LSTM Random forest	[83] [58]
	Crop yield prediction	United States Argentina, Brazil	RMSE	0.37 t/ha 0.62 t/ha, 0.42 t/ha	CNN+GP LSTM	[110] [101]
	Field delineation	France	Dice score	0.61 0.87	U-Net FracTAL Res-UNet	[9] [99]
Good Health & Well-Being	Child mortality rate Women BMI	56 countries 53 countries	r^2 r^2	0.01 0.42	kNN kNN	– –
Quality Education	Women education	53 countries	r^2	0.26	kNN	–
Clean Water and Sanitation	Water index Sanitation index	49 countries 49 countries	r^2 r^2	0.40 0.36	kNN kNN	– –
Climate Action	Brick kiln detection	Bangladesh	Accuracy	0.94*	ResNet-50	[63]
Life on Land	Representation learning for land cover	United States	Accuracy	0.55 ($n = 1,000$) 0.58 ($n = 10,000$)	Tile2Vec with ResNet-50	[53]
	Out-of-domain land cover classification	Global	Kappa	0.32 (1-shot, 2-way)	MAML with shallow 1D CNN	[104]

Figure 6: SustainBench Benchmark performance on 15 tasks across 7 SDGs. The asterisk (*) indicates a result on a similar datset, but not the exact SustainBench test set.

mindful that we are losing some useful inductive image biases (e.g., translational equivariance). As a result, the model will either require significant pre-training or architectural modification to recover these useful biases. Third, we must collect a large pre-training dataset.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. *CoRR*, abs/1305.6663, 2013. URL http://arxiv.org/abs/1305.6663.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey E. Hinton. Pix2seq: A language modeling framework for object detection. *CoRR*, abs/2109.10852, 2021. URL https://arxiv.org/abs/2109.10852.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Kenneth Gade. A non-singular horizontal position representation. *Journal of Navigation*, 63(3): 395–417, 2010. doi: 10.1017/S0373463309990415.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. URL https://arxiv.org/abs/2111.06377.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs and outputs, 2021a. URL https://arxiv.org/abs/2107.14795.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. *CoRR*, abs/2103.03206, 2021b. URL https://arxiv.org/abs/2103.03206.

D. Jervis, J. McKeever, B. O. A. Durak, J. J. Sloan, D. Gains, D. J. Varon, A. Ramier, M. Strupler, and E. Tarrant. The ghgsat-d imaging spectrometer. *Atmospheric Measurement Techniques*, 14 (3):2127–2140, 2021. doi: 10.5194/amt-14-2127-2021. URL https://amt.copernicus.org/articles/14/2127/2021/.

Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised concept expansion for general purpose vision models. *CoRR*, abs/2202.02317, 2022. URL https://arxiv.org/abs/2202.02317.

Nitish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. Unifying question answering and text classification via span extraction. *arXiv*, 04 2019.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226, 2018. URL http://arxiv.org/abs/1808.06226.

Alexandre Lacoste, Evan David Sherwin, Hannah Kerner, Hamed Alemohammad, Björn Lütjens, Jeremy Irvin, David Dao, Alex Chang, Mehmet Gunturkun, Alexandre Drouin, Pau Rodríguez, and David Vázquez. Toward foundation models for earth monitoring: Proposal for a climate change benchmark. *CoRR*, abs/2112.00570, 2021. URL https://arxiv.org/abs/2112.00570.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–1188–II–1196. JMLR.org, 2014.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks, 2022. URL https://arxiv.org/abs/2206.08916.

Alfonso López, Juan M. Jurado, J. Roberto Jiménez-Pérez, and Francisco R. Feito. Generation of hyperspectral point clouds: Mapping, compression and rendering. *Computers and Graphics*, 106: 267–276, 2022. ISSN 0097-8493. doi: https://doi.org/10.1016/j.cag.2022.06.011. URL https://www.sciencedirect.com/science/article/pii/S0097849322001145.

Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Stefano Ermon, Jiaming Song, Krzysztof Janowicz, and Ni Lao. Sphere2vec: Self-supervised location representation learning on spherical surfaces, 2022. URL https://openreview.net/forum?id=FS0XKbpkdOu.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730, 2018. URL http://arxiv.org/abs/1806.08730.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning, 2022. URL https://arxiv.org/abs/2203.06604.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2016. URL http://arxiv.org/abs/1612.00593. cite arxiv:1612.00593.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5105–5114, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies, 2022. URL https://arxiv.org/abs/2206.04670.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *arXiv*, 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *arXiv*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL http://arxiv.org/abs/1910.10683.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent, 2022. URL https://arxiv.org/abs/2205.06175.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay, 2015. URL http://arxiv.org/abs/1511.05952. cite arxiv:1511.05952Comment: Published at ICLR 2016.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. URL https://arxiv.org/abs/2111.02114.

Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, David B. Lobell, Marshall Burke, and Stefano Ermon. Predicting economic development using geolocated wikipedia articles. *CoRR*, abs/1905.01627, 2019. URL http://arxiv.org/abs/1905.01627.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2022. URL https://arxiv.org/abs/2206.14486.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017a. URL http://arxiv.org/abs/1706.03762.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017b.

Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David Lobell, and Stefano Ermon. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems, Datasets and Benchmarks Track (Round 2)*, 12 2021. URL https://openreview.net/forum?id=5HR3vCylqD.