Satellite Imagery-Informed Deep Learning for Road Speed Prediction in Rural Mozambique

George Hu
gehu@stanford.edu

Vivek Vajipey
vvajipey@stanford.edu

Akayla Hackson
akayla@stanford.edu

Tillmann von Carnap
tcarnap@stanford.edu

Abstract

Many regions of Mozambique can be characterized by high poverty rates and frequent extreme weather events, and an important development strategy is to target road improvements to better serve infrastructure needs and to boost the agricultural productivity of impoverished areas. However, accurately determining travel speeds and understanding local climate impacts is challenging due to the lack of comprehensive, real-time ground data. This study addresses the challenge of evaluating road quality over time in these provinces. We utilized publicly available high-resolution (5-meter) NICFI Planet satellite imagery, combined with GPS tracking and survey data collected in three phases from April 2021 to August 2023. Our methodology predicts average road speeds more accurately than conventional road quality assessments or seasonal baselines. This contributes valuable insights into rural road development, guiding investments towards creating a more resilient road network.

1. Introduction

The provinces of Nampula and Zambézia in Mozambique are characterized by a striking contrast: they possess significant agricultural capabilities yet suffer from some of the world's highest poverty rates [1]. This dilemma is further complicated by regular extreme weather incidents that damage infrastructure and impede economic progress. In an effort to stimulate agricultural growth, the Mozambican government, supported by the World Bank, has initiated a significant effort to enhance the rural road network. However, questions remain on the durability of these roads given the local climate conditions and their actual effectiveness in improving transportation [1].

The challenges of physically surveying these roads due to their poor condition and associated risks necessitate an alternative approach [1]. This research utilizes advanced satellite imagery with high resolution and frequency to track and assess road conditions over time. This approach not only helps to determine the direct impact of road construction on agricultural development, but also offers valuable insights for the upkeep and design of more robust road systems.

This project benefits from exclusive primary road survey data from these provinces, enriched by comprehensive archives of 5-meter multispectral PlanetScope imagery via the NICFI program. These surveys have been carried out over three phases, ensuring that each road segment was surveyed at least three times between April 2021 and August 2023.

The primary goal of this study is to leverage the GPS data to predict travel speeds, including both the average and variability, on these road segments. This predictive analysis is aimed at gaining a deeper understanding of how effective the road network is in these areas, providing a concrete gauge of the road infrastructure's influence on local travel. The outcomes of this research are expected to inform sound policy decisions and infrastructure planning, contributing towards the economic and social advancement of the Nampula and Zambézia regions.

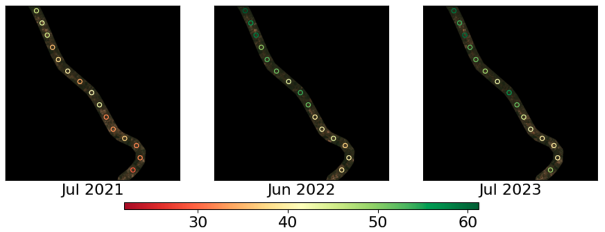


Figure 1. Road segment speed (in km/h) predictions over time.

Moreover, this study is contextualized within broader African infrastructure challenges: less than 50 percent of people in Africa live within two kilometers of an all-season road [13], high transport costs significantly hamper agricultural productivity [17], and despite substantial development

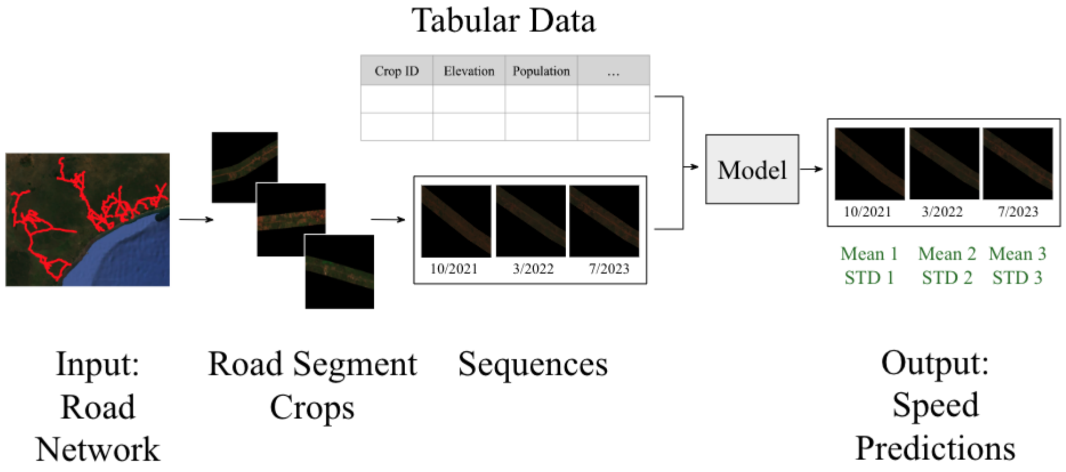


Figure 2. Full modeling approach to predict road segment speeds over time in Mozambique.

financing for roads, there's limited evidence of their effects on market access and agricultural intensification, especially considering road quality.

2. Related Work

Predicting road conditions and speed from satellite imagery is a relatively recent domain, as higher resolution satellite imagery needed to distinguish road features have only recently become publicly accessible [2]. Parallel to these advancements on the data side, advancements in image processing have found Convolutional Neural Networks (CNNs) to perform well across a wide array of applications. Most of the literature in satellite road quality analysis has focused on CNN-based methods, and Kalpoma et al. (2022) [10] show that deep learning methods generally outperform combining hand-engineered features with traditional statistical methods for road quality analysis.

Steinen et al. (2023) [16] perform a similar task to ours by using a custom CNN architecture to encodings of satellite images, which are used as inputs, along with additional tabular weather features, into fully connected layers in order to predict mean speed along road segments in Indonesia. Using 10-meter resolution RGB satellite imagery from Sentinel-2 in 10 pixel by 10 pixel crops, along with short wave infrared statistics within the bounds, rainfall information, road type, and time of day, they create a model that generalizes to unseen roads with a prediction error of 8.5 km/hr RMSE. The processing of satellite image crops along a segment of road is notable, aggregating across the whole segment by averaging the crops for a segment-wide prediction. And while we prefer greater granularity and less smoothing of image inputs, we emulate the approach here of integrating important tabular features in later stage linear layers.

Our modeling approach also focuses on two more complex but distinct areas: leveraging in-domain, i.e. within satellite imagery, pretraining through self-supervised methods, and time-sequence prediction.

Recent trends in machine learning in the language domain have revealed the importance of self-supervised methods such as next token prediction and span-masking that pretrain large language models across all areas of human language [19]. And in the vision domain, many approaches employing self-supervision and representation learning of large natural image datasets have proliferated, such as simply pretraining on ImageNet [8] or masked autoencoding [6]. However, given that satellite imagery features are inherently different, Wang et al. (SSL4EO-S12, 2023) [18] use 12-band Sentinel-2 imagery across various biomes to train models for the satellite image domain, using momentum-contrastive learning [7] that can act as a generalizable satellite image encoder for various applications.

A particular area we decided to focus on is time-sequence prediction so that our methods could inform road quality across multiple time snapshots. In machine learning incorporating sequential biases is not new, and the time-dependent domain of video analysis has spearheaded numerous sequential prediction methods. In the domain of action recognition and video description, Donahue et al. (2016) [5] show that using a CNN as an image encoder and then passing that sequence of embeddings through a Recurrent Neural Network (RNN), in this case a Long Short-Term Memory (LSTM) network, is an effective approach. We find this work specifically useful because the vision encoder is separate from the sequence prediction, so that we can still use the structure from Steinen et al. [16] and Wang et al. [18].

3. Data

3.1. Tabular Data

We use a dataset from the Nampula and Zambézia regions, assisted by local organizations and researchers from the World Bank, Stockholm University and Stanford. They employed regional survey firms to drive along specified routes of the road network, recording GPS tracks and qualitative statements regarding the road condition. Each road segment has been visited at least three times between April 2021 and August 2023 over three rounds of surveying, by one of five drivers (referred to as enumerators).

The road network traversed by the enumerators was provided with the dataset as a shapefile. The dataset consists of 3,312,184 total GPS tracks, with 44% of observations from Round 1, 41% from Round 2, and 15% from Round 3. Each GPS track is associated with a timestamp, coordinates, speed (in km/h), elevation, and information regarding data collection, such as the ID of the enumerator and the round number. The GPS tracks were recorded at varying times of the year, with May 2021 having over 300,000 observations and several months having a few thousands observations (December, 2021) or none at all (e.g. September 2022 through May 2023). Moreover, the majority of GPS observations were collected during the morning.

In addition to our collected GPS tracks, we incorporated population density into our analysis, using Meta Data For Good's high resolution population density maps [11]. By combining granular population density information with other attributes like type of road of surrounding land use, we could achieve better speed predictions. In figure 3 the relationship between population density and GPS track road speed is shown.

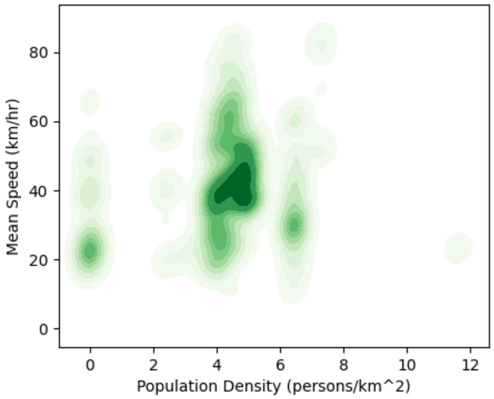


Figure 3. Density plot of population density and mean road speed. We can observe how various clusters have important correlations.

3.2. Satellite Imagery

To obtain real-time visual information on road conditions, we use satellite imagery of monthly composites from Planet NICFI basemaps [14] , which have a resolution of 5 meters for visible (R, G, B) and near-infrared bands. The satellite images were generated with 100 meters of buffer on each side around the corresponding portion of the road in the road network shapefile. Figure 1 illustrates a representative example of these GPS tracks overlaid onto the Planet NICFI satellite imagery.

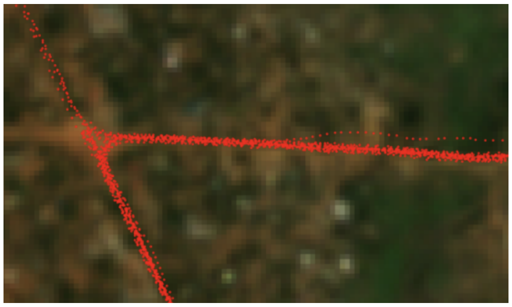


Figure 4. GPS Tracks overlaid on satellite image of a road segment.

3.3. Data Preprocessing

Provided the NICFI imagery as monthly composites over three imaging rounds corresponding to the survey data in 2021, 2022, and 2023, we aimed to provide speed predictions along a road segment for each time step. Thus, we first grouped all images by road segment and sequenced them along the time axis into any valid (round 1, round 2, round 3) sequence. Then, to get spatial predictions within that road segment, we generated 224 pixel by 224 pixel crops along the road path every 56 pixel span for all image sequences. This leaves us with crop sequences of shape (3, 224, 224, 4) for sequence length 3 and channels RGB+NIR. We demonstrate the cropping proccess in figure 5

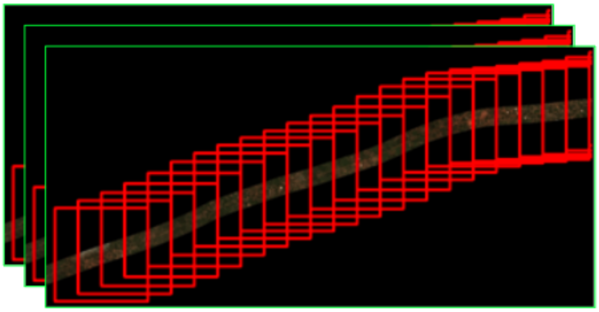


Figure 5. Crops (224x224) of road segment from satellite imagery data.

For tabular features, our approach was inspired by Yin et al. (2021) [20] in their creation of a road feature dataset. We decided to include the following:

- Elevation mean and standard deviation - The elevation within a crop can provide information on the type of terrain and the slope, which can affect how fast cars tend to drive in an area.
- Population density - The density of residents can affect the type of usage for the road and congestion patterns.
- Date and time features - We one-hot encode the year, corresponding to the round number, and month to allow the model to learn desirable seasonal and general improvement over time biases. We also include a histogram for the number of GPS observations within each hour to provide information on the proportion of speed measurements at day and night or with respect to traffic patterns.

We normalized all of our inputs and outputs, including all input image channels, all tabular features, and speed information, to have mean 0 and variance 1. This approach was taken to prevent any feature from having a greater impact on the model simply because of its scale. For our results, the outputs of the model in terms of speed are denormalized back to an interpretable km/hr.

4. Method

4.1. Baselines

Although the Steinen at al. [16] paper for road speed prediction in Indonesia has a similar approach and domain of rural roads, we cannot directly use their reported error rate of 8.5 km/hr as a baseline because their evaluation dataset is not the same. Instead, we use baselines involving coarse and/or sparse data that are common used for rural road speed prediction in developing nations.

1. Absent any modeling, it would be reasonable to use the latest speed recorded by a driver for a specific road segment. Thus, we use our GPS tracks to find the most recent set of speeds observed within a road crop before the desired prediction month. This is the **Previous Observation Baseline**.
2. In 2016, the government of Mozambique created an comprehensive survey categorizing roads into poor, fair, good, and unavailable [3]. We use a simple regression derived from these categories as the **Road Quality Baseline**. In figure 6, we show the distribution of categories of our road crops.

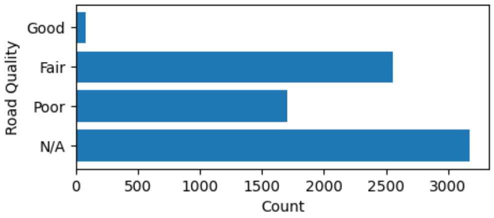


Figure 6. Distribution of road quality for all data (train and val)

4.2. Modeling Approaches

Given our input data consisting of sequences of RGB and NIR crops along with our chosen tabular features providing seasonal, time-of-day, and geographical context, we combined the approaches of the CNN with concatenated features in Steinen et al. [16] and the CNN-RNN sequential modeling approach in Donahue et al. [5]. That is, our architecture for embedding an image and its corresponding categorical features involved passing the image through a CNN and then concatenating the output embedding with the tabular features. In order to learn sequential information between the three image crops, we passed those features through an RNN to get the mean speed prediction at each time step. A diagram of the approach can be seen in figure 7.

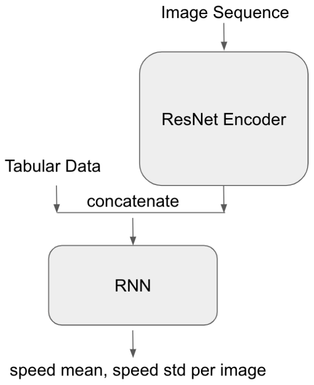


Figure 7. Depiction of our sequence modeling approach

We decided on using a ResNet encoder as our choice of CNN for its ubiquity in research and adaptive pooling in the last layer to ensure consistency if we wanted to explore changing the crop size. Furthermore, we wanted to employ the same usage of self-supervision on satellite imagery in SSL4EO-S12 [18], to ensure the encoder weights have a very good initialization. They provide momentum-contrastive self-supervised models pretrained on 12-band Sentinel-2 for ResNet18 and ResNet50, so we use that family of pretrained models. To initialize the first-layer weights of our 4 channel model, we discarded the other 8 channels in the pretrained model that we do not use.

We explored the vanilla RNN, GRU [4], and LSTM [9] architectures, each just using one layer and having hidden state(s) initialized with zeros for the sequence modeling. Since our sequence length was only 3, we supposed that it would not be mandatory to use a complex RNN architecture, so we decided to experiment with just a few simple versions of common approaches. We chose not to incorporate bidirectionality, as we wanted to keep our method as look-forward and in theory applicable without future knowledge.

All of our methods were implemented in PyTorch on a Google Compute Engine VM with a NVIDIA T4 GPU. We used batch size 32, learning rate 0.003, dropout rate 0.1 [15], the AdamW optimizer [12] with weight decay 0.0001, and hidden dimension 128. For all models, we trained for 50 epochs and selected the best checkpoint based on validation RMSE.

5. Experiments

5.1. Results

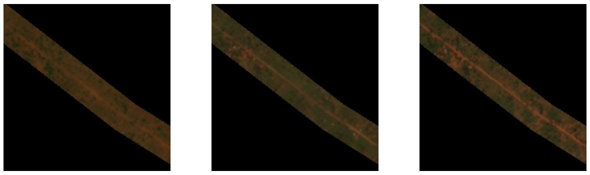
Our results found that the ResNet50 encoder with Vanilla RNN performed the best and greatly outperformed our seasonal baseline and road quality baseline. We evaluated our models using root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (r^2).

Model	RMSE	MAE	r^2
Previous Obs Baseline	15.186	11.432	-0.162
Road Quality Baseline	13.532	11.213	-0.060
Ours	**11.240**	**9.083**	**0.407**

Table 1. Results of baselines and our method on the validation set.

In Table 1, we can observe the divergence between the naive baselines currently used in the economic development space, and our more complex model, particularly in the r^2 value. The seasonal baseline performs marginally better than the road quality baseline, likely due to being able to account better for the wet and dry seasons.

In figure 8, we can see an example prediction of our model compared to the ground truth. Visual inspection can see how the road quality has improved over the three rounds, which our model captures, but it is unable to fully ascertain the magnitude of the speed increase, indicating an anchoring effect on the first prediction and a tendency to not predict at the extremes of the speed distribution.



Nov 2021
Label: 30.10 km/hr
Model: 32.48 km/hr

Apr 2022
Label: 52.71 km/hr
Model: 43.15 km/hr

Jun 2023
Label: 69.32 km/hr
Model: 48.90 km/hr

Figure 8. Comparison of model prediction and label for image sequence over 3 years. Brightness enhanced by 3x.

5.1.1 Ablation Studies

In the results shown in Table 2, we found that the effect of pretraining type and the addition of sequence modeling helped our results significantly.

Pretraining Method	RMSE
No Pretraining	13.961
Supervised ImageNet Pretraining	12.773
Ours (Sentinel-2 MOCO Self-Supervised)	**11.240**

Table 2. Effect of pretraining type and data source on the validation set.

For the no pretraining experiment, we initialized weights with the default Xavier uniform initialization, and the supervised ImageNet pretraining is from the torchvision default weights. The Sentinel-2 pretraining seems to help our generalization by being in-domain applied to a diverse set of satellite data, and also trained in a effective way to map similar representations together and dissimilar ones apart.

Sequence Modeling Method	RMSE
No Sequence Modeling (FC)	12.430
GRU	11.843
Ours (Vanilla RNN)	**11.240**

Table 3. Effect of sequence modeling approach on the validation set.

In Table 3 we can see how not having sequence modeling and instead separate linear layers gives poor performance, as the model does not incorporate any relevant temporal information. But it also seems that too much complexity in the sequence model has negative effects, as the GRU performs worse than the vanilla RNN. Many of these more complex sequence models have in mind long-term recurrence, and since our sequence length is three, it's not surprising the vanilla RNN does very well.

5.2. Analysis

In order to better understand the model's performance, we examined how the best model (ResNet50 encoder with Vanilla RNN) performed various sequence categories. Since each sequence consists of three images, the mean speed for the sequence can either monotonically increase, monotonically decrease, increase then decrease, or decrease then increase. These results are presented in Table 4.

Speed Mean Trend	# Seqs	RMSE	MAE
Monotonic Incr	246	12.494	10.383
Incr then Decr	98	13.439	11.290
Decr then Incr	126	**12.019**	**10.045**
Monotonic Decr	24	19.125	15.793

Table 4. Performance of best model (ResNet50 encoder with Vanilla RNN) on subsets of the validation set based on the trend in mean speed for the sequence.

Based on this analysis, it is clear that the best model underperforms on sequences with monotonically decreasing mean speeds compared to the other mean speed trend categories. This may be due to the fact that monotonically decreasing mean speeds are relatively underrepresented in both the training and validation set, meaning that the model is disproportionately trained to identify speed increases.

In order to understand the spatial distribution of the model's performance, we plotted the quartile of MAE per sequence in Figure 9. In areas such as Niasaca in the southwest portion of Figure 9, the model systematically predicts incorrectly. Conversely, there are regions such as the northeast portion of Figure 9 for which the model has varying performances, even along the same road. Overall, there does not seem to be a clearly identifiable geographic trend in the model's performance, meaning that subtler differences in road characteristics could be responsible.

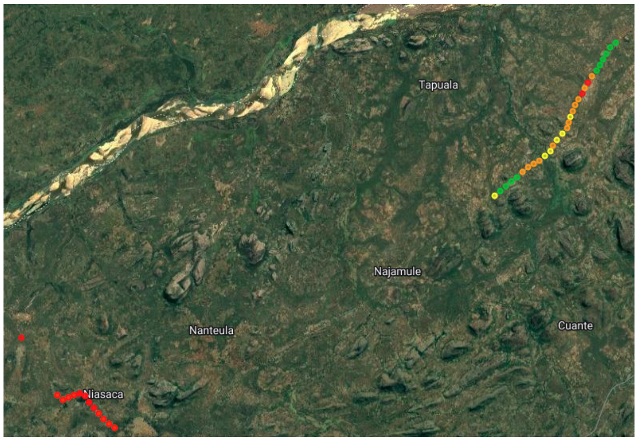


Figure 9. Map of best model (ResNet50 encoder with Vanilla RNN) mean speed performance in Niasaca, a northern portion of the Nampula region. Color represents the quartile of mean absolute error for the sequence: green, yellow, orange, and red correspond to 1st, 2nd, 3rd, and 4th quartiles, respectively.

6. Conclusion

6.1. Future Work

Our deep learning methodology has demonstrated the capacity to outperform baseline models with a relatively modest dataset, highlighting the efficiency of our approach not just in terms of predictive performance but also in cost and time, compared to traditional survey methods. However, it is important to acknowledge certain limitations of our current approach. For instance, the reliance on available datasets may introduce biases or overlook subtler variances not captured by the data. Additionally, satellite image quality and the frequency of updates can impose constraints on the timeliness and relevance of predictions. Considering the context of using the speed detection model to identify roads needing attention and improvement, it is crucial that the solution is particularly effective for these examples. However, the best model's relatively poor performance on sequences with monotonically decreasing speed indicates that future iterations will need to develop methods to improve performance in these underrepresented yet crucial categories of roads.

To further advance our methodology, we plan to pretrain our model on higher resolution satellite imagery. This approach is crucial as current imagery, such as that from Sentinel-2, does not resolve roads adequately for our purposes. By using higher resolution data, our model can potentially offer more precise and reliable results. Despite these challenges, we remain optimistic that future advancements in technology and data collection will enable us to overcome these hurdles. As higher quality datasets become available and machine learning algorithms grow more sophisticated, we anticipate that the drawbacks of today's models will give way to tomorrow's solutions, further enhancing the strategic planning and maintenance of rural road networks. This progress, in turn, will contribute to the broader goals of economic development and poverty alleviation in regions like Nampula and Zambézia.

6.2. Takeaways

Our study paves the way for a series of enhancements that promise to refine the predictive capabilities of machine learning in road infrastructure analysis. Based on the comparisons between our model and the baselines from the literature, it is evident that our work shows promise for quantitative characterizations of rural roads in developing regions.

We show that satellite image based deep learning techniques greatly improve from proper in-domain pretraining, and that suitable sequential time predictions increase accuracy when road segments change over time.

Current methodologies for determining road quality and road speed over time in rural Mozambique rely upon data that is often sparse, unreliable, and/or expensive, so our relatively accurate remote sensing approach or similar approaches can provide magnitudes better analyses. In the application area of determining connectivity-based metrics of underutilized economically productive land, this far more accurate picture of complete road segments over time can allow for significant on-the-ground impacts. In the picture of economic development for the world's most impoverished and data-scarce regions, we find very strong potential in using satellite imagery and deep learning methods as a first step.

References

[1] World Bank, Jun 2023. 1

[2] Ethan Brewer, Jason Lin, Peter Kemper, John Hennin, and Dan Runfola. Predicting road quality using high resolution satellite imagery: A transfer learning approach. *Plos one*, 16(7):e0253370, 2021. 2

[3] Paul Christian and Anna Tompsett. Route to development: Impacts of road network improvements on agricultural intensification in mozambique. World Bank-IDA with co-financing from the Government of Mozambique, 2020. Sponsored by DFID, United Kingdom, and implemented in partnership with Administração Nacional de Estradas (ANE) – National Road Authority, Mozambique. 4

[4] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. 5

[5] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description, 2016. 2, 4

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

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

[8] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training, 2018. 2

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997. 5

[10] Kazi A Kalpoma, Dipesh Shome, Anas Sikder, and Abrar Jahin. A comprehensive study on road quality measurement from high resolution satellite imagery. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 2594–2597, 2022. 2

[11] Facebook Connectivity Lab and Center for International Earth Science Information Network CIESIN Columbia University. Mozambique: High resolution population density maps + demographic estimates, Oct 2022. 3

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 5

[13] Mehdi Mikou, Julie Rozenberg, Elco Koks, Charles Fox, and Tatiana Quiros. *Assessing Rural Accessibility and Rural Roads Investment Needs Using Open Source Data*. 1

[14] Planet API. Nicfi basemaps. `https://api.planet.com/basemaps/v1/mosaics`, 2023. 3

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. 5

[16] Valentijn Stienen, Dick den Hertog, J.C. Wagenaar, and J.F. Zegher. Better routing in developing regions: Weather and satellite-informed road speed prediction. Workingpaper, CentER, Center for Economic Research, Sept. 2023. CentER Discussion Paper Nr. 2023-025. 2, 4

[17] Tavneet Suri. Selection and comparative advantage in technology adoption. *Econometrica*, 79(1):159–209, 2011. 1

[18] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation, 2023. 2, 4

[19] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022. 2

[20] Yifang Yin, An Tran, Ying Zhang, Wenmiao Hu, Guanfeng Wang, Jagannadan Varadarajan, Roger Zimmermann, and See-Kiong Ng. Multimodal fusion of satellite images and crowdsourced gps traces for robust road attribute detection. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '21, page 107–116, New York, NY, USA, 2021. Association for Computing Machinery. 4