Case study in Riyadh: Analyzing Urban Heat Island (UHI) by Object Detection from High-Resolution Satellite Imagery

Austin Nguyen
Institute for Applied
Computational Science (IACS)
Harvard University
150 Western Avenue
Allston, 02134
austinnguyen@g.harvard.edu

Faisal Alnasser
Department of Civil
Environmental Engineering
Massachusetts Institute for Technology
77 Massachusetts Ave
Cambridge, MA 02139
alnasser@mit.edu

Abstract

The use of machine learning on satellite imagery has the potential to address and anticipate adverse climate events by estimating climate impacts of urbanization, especially in quickly developing regions of the world. We demonstrate that satellite imagery in Riyadh, Saudi Arabia can support prediction of vehicle, asphalt, and building density. We seek to understand the impact of climate as measured through the Urban Heat Island (UHI) effect by examining the relationship between predictions and LandSat Temperature data (LST).

1 Introduction

In this work, we train a machine learning model on satellite image data with vehicle labels taken in a specific neighborhood of Riyadh, Saudi Arabia with the goal of detecting the presence of vehicle, building, and road density in satellite images taken in the city over the course of 2017. The goal is to understand the climate impacts in the urban environment. We aim to detect the urban infrastructure, analyze the effects of urban infrastructure, and consequently understand their correlation with the urban heat island (UHI) effect.

This research is motivated by the following research questions:

- What are the spatial and temporal relationships between the urban landscape density and pollution?
- What are the spatial and temporal relationships between urban landscape and urban heat island effect?
- What are the spatial and temporal relationships between vehicle density and emissions?
- How does this vary by roads with different types of vehicles?

2 Background

Other research has been done on object detection with satellite image data (García-González, et al. 2021), but our work adds value by fusing high-resolution images with land surface temperature. Earlier research has been done on analyzing Urban Heat Islands by land type from Landsat images, we extend on that work by using the higher resolution satellite, WorldView-3. Previous work on machine learning on global satellite imagery demonstrated methods that were generalizable to diverse

Preprint. Under review.

prediction tasks such as estimating forest cover and housing prices (Rolf, et al, 2021). Our work adds value by examining the use of machine learning techniques to analyze physical parameters such as heat. Our work utilizes U-NET as an architecture choice for semantic segmentation on object detection. Originally built for biomedical image segmentation, U-Net uses a contracting network by successive layers and replaces typical pooling operations with upsampling operators which ultimately increases the resolution of the output. The contracting path or encoder consists of a series of repeated convolutions in which spatial information is reduced while feature information is increased. The architecture takes advantage of combining spatial information through a sequence of up-convolutions and concatenation of high-resolution features from the encoder portion to the decoder portion. This architecture helps propagate valuable information from the encoder portion of the architecture to the decoder portion to improve semantic segmentation.

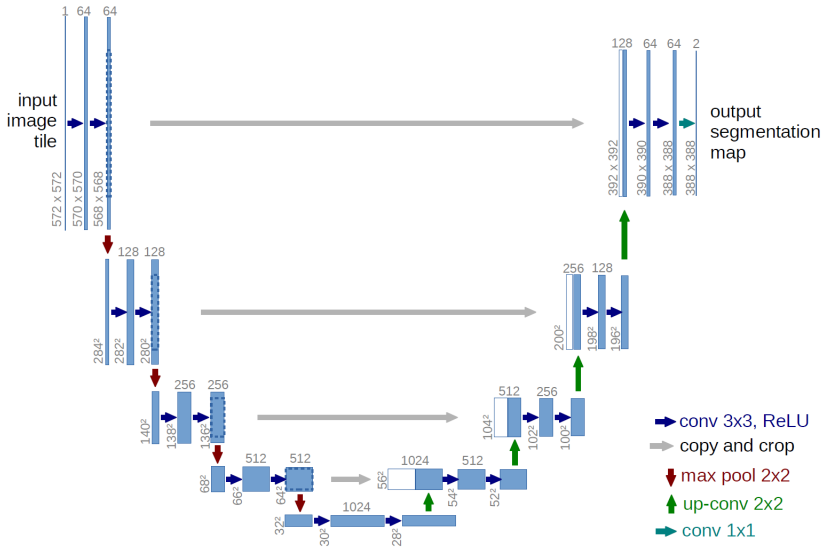

Figure 1: U-Net Architecture. Source: By Mehrdad Yazdani , CC BY-SA 4.0

The typical CNN architecture involves increasing the number of feature maps following each max pooling operation. However, the U-NET model we chose for our implementation instead uses a constant number (64 feature maps) throughout the entire network. This is because we are willing to trade-off the loss of information after the down-sampling encoder portion because the U-NET architecture permits access to low-level features through skip connections. Furthermore, satellite images do not have depth or high-level 3D objects to process, so an increasing number of feature maps may not necessarily lead to better performance.

3 Data Description

3.1 Data Collection

We obtain high resolution satellite imagery from the WorldView3 (WV3) satellite for the city of Riyadh with a spatial resolution of 0.3 m which we use for object detection. The images were provided by the Center of Complex Systems (CCS) in Riyadh.

We are working with labeled data. For training and testing purposes, we have a single neighborhood in the city that was manually labeled and includes 5k vehicles, 2 square km of asphalt and 3 square km of built area (Figure 1). Additionally, utilized a huge labeled dataset published by SpaceNet (a public and free repository of precision-labeled, high-resolution satellite imagery). The labels contain buildings and roads for 5 cities around the globe. Among these cities, Khartoum and Shanghai were used to pre-train our models, the cities were chosen based on their similarity with Riyadh.

We also procure land surface temperature (LST) provided by Landsat-8, a satellite which captures each area in the world with a temporal resolution of 16 days at 10am local time. The dataset is

publicly available through Google Earth Engine. (LST) is captured by satellite through infrared thermal imaging (see Figure 2).

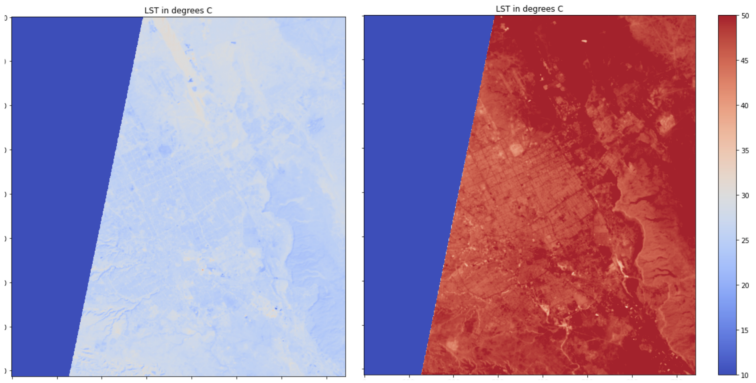

Figure 2: Land Surface Temperature (LST) captured by Landsat 8 in Riyadh during winter 2016 (left) and summer 2016 (right)

3.2 Preprocessing

Landsat imagery comes with several noise issues such as clouds and missing data/gaps. To overcome the noise produced by cloud contamination, we set a filter for cloud cover at less than 0.5%. However, the pre-processing technique is not perfect, as some cloudy images persist even after applying this threshold (Figure 3). By empirically looking at the data, a threshold of 0 Celsius was used to eliminate clouds. Sub-zero temperatures are rare in the region so the threshold is reasonable. Additionally, we take the annual mean across all images (2016-2018) to overcome coverage differences and gaps in the data (as shown in Figure 2).

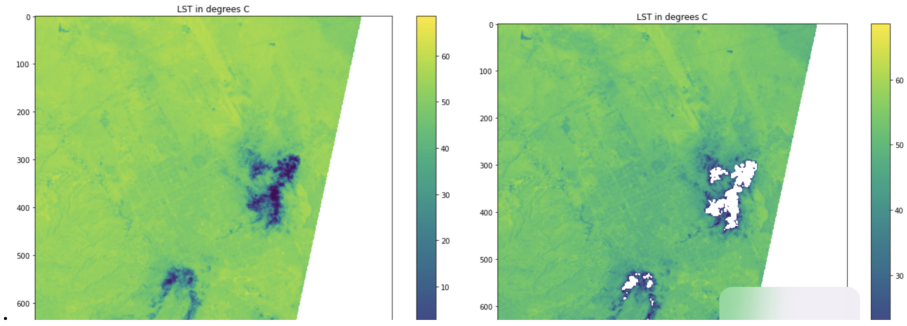

Figure 3: Cloud contaminated image (left), after performing temperature threshold (right)

The full city image was captured on a different time than our labeled neighborhood. In order to optimize the predictions on the full-sized image, a histogram matching [5], a method to match color histograms, was applied to our data before training to match the histogram distributions between the full city image and our labeled neighborhood (Figure 4).

3.3 Data Augmentation

We implemented a series of data augmentation steps to our image data in order to improve our model's generalizability to unseen data. We applied the following transformations: Gaussian noise, brightening, darkening, rotation, horizontal flipping, and vertical flipping. We implement these changes because the city of Riyadh does not follow a perfect grid. Vehicles, buildings, and roads can be present in an image at different orientations and rotations from the x-axis and y-axis that would follow latitude and longitude, respectively. Thus, data augmentation techniques such as rotation and

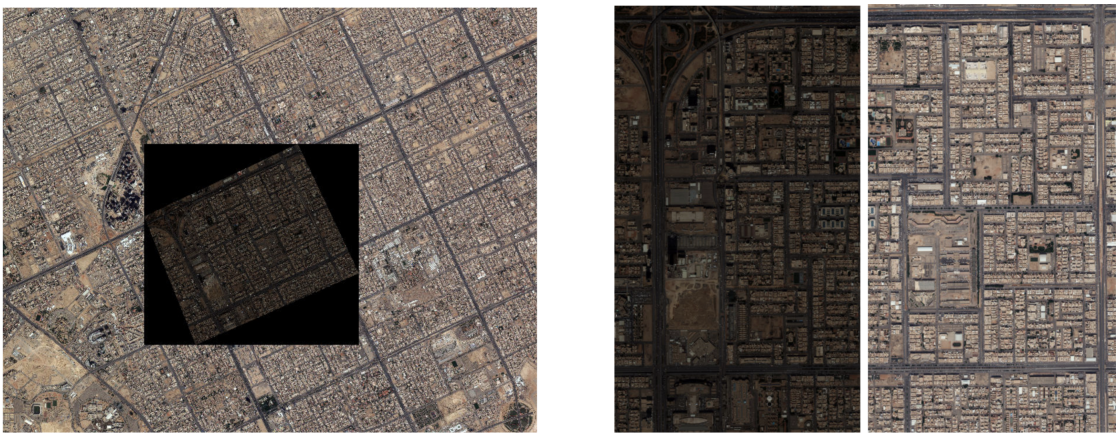

Figure 4: Histogram matching, the full image with the neighborhood overlaid on top (left), before-after histogram matching (right).

flipping may make it more challenging to for our model, but this is in service of improving model generalizability.

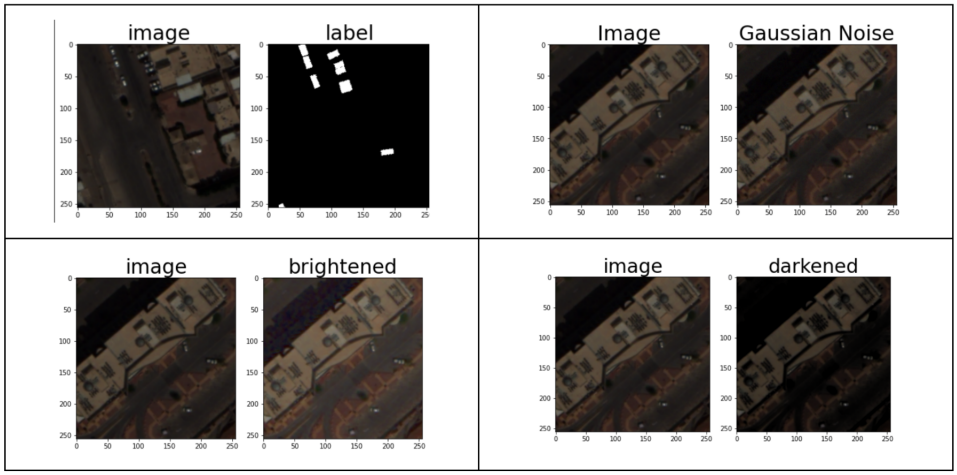

Figure 5: (a) Cropped 256px image with a mask that represents our label of interest in white (vehicles) and everything else in black. (b) Cropped 256px image with Gaussian noise applied (c) Cropped 256 image with brightened transformation and (d) Cropped 256px image with darkened transformation.

4 Methods

We have hand-labeled vehicles, buildings, and roads labels for image data from high-resolution satellites capturing Riyadh, Saudi Arabia. We cropped this high-resolution image into smaller images to ease computational processing at two different crop-dimensions: 64 pixels and 256 pixels. We evaluated the model based on predicting the vehicle label in our validation set.

We trained four different models: (a) model trained from scratch on 64px images, (b) model trained from scratch on 256px images, (c) model with pre-trained weights from SpaceNet data, re-trained on our 64px images, and (d) a model with pre-trained weights from SpaceNet data, re-trained on 256px images. We ran our U-NET model with the following set of hyperparameters: 10 epochs, binary cross entropy for loss, and Adam for optimizer. We choose binary cross entropy because our labels is a mask overlaid on our input images in which 1 represents the object of interest (vehicles, roads, buildings) and 0 represents the rest. We choose Adam as an optimizer because it adapts the learning

rate as well as stores keeps an exponentially decaying average of past gradients in a way that mirrors momentum and has been shown to lead to faster convergence. We chose to experiment with different crop sizes to better understand the trade-off between providing more spatial information with larger crop sizes and computational cost.

5 Results

We retrained models (c) and (d) with pre-trained weights from a model trained on SpaceNet. The motivation for re-training a model with pre-trained weights is to allow comparable performance. To evaluate the performance of our models, we examined precision, recall, F1-score and accuracy. Our goal metric is F1-score because it accounts for both false positives and false negatives. Given that the imbalance of pixels in our image (i.e. vehicles often account for a small portion of each image), F1 score is appropriate because it accounts for class imbalance. Lastly, F1 score is the harmonic mean between precision and recall. Precision is a metric that ensures the correctness of our predictions whereas recall is a metric that ensures we capture all objects of interest in the image. Instead of prioritizing one or the other, we prioritize F1 score which balances the trade-off between precision and recall.

Table 1: Performance on test set across models

	Metric			
Model	Precision	Recall	F1 score	Accuracy
(a) UNet from scratch 64px	0.76	0.85	0.80	0.99
(b) UNet from scratch 256px	0.82	0.81	0.81	0.99
(c) UNet pre-trained 64px	0.83	0.81	**0.82**	0.99
(d) UNet pre-trained 256px	0.82	0.79	0.81	0.99

For model (a), we observed that the model's prediction for vehicles in our test set had a 76% precision, 85% recall, and 80% F1-score. This model did not use pre-trained weights and used a 64px crop.

The predictions of model (b) shows a 82% precision, 81% recall, and 81% F1-score. This model was trained from scratch and uses a larger crop size of 256px and we thus see a slight improvement in the F1 score metric. This may be a function of providing the model more data per image and so it is better at capturing the spatial relationship between pixels to locate the object of interest.

The predictions of model (c) for vehicles shows 83% precision, 81% recall, and 82% F1 score on the test data. This model was trained on pre-trained weights obtained from training on SpaceNet data. This is an improvement from model (b) which was the model with weights trained from scratch (76% precision, 85% recall, and 80% F1-score). This demonstrates that the model with pre-trained weights on SpaceNet data for 64px cropped images performed better than model trained from scratch for 64px cropped images. This suggests the value of using a model with pre-trained weights.

The predictions of model (d) for vehicles shows a 82% precision, 79% recall, and 82% F1 score on the test set. This model also uses pre-trained weights but takes advantage of a larger crop size of 256px. This is comparable to the performance of model (b) which is similar to model (d) in that both were implemented on images with 256px crop, however model (d) suffers from a lower recall score compared to model (b) (0.79 vs 0.81).

We observed that train accuracy and validation accuracy increase with each epoch, with 99% for validation accuracy and 98% training accuracy by the 10th epoch. We observe that validation accuracy is higher than train accuracy because we carried out a series of data augmentation steps to training data and not to our validation data. While the validation set was sampled from the same distribution as the training set, the final validation observations are not as challenging as our training examples because we did not carry out data augmentation on the validation set. We defined binary cross entropy as our loss. Training loss decreases as epochs increase. Best training loss is 0.48 and best validation loss is 0.28. We observe that training loss is higher than validation loss due to lack of data augmentation applied to the validation set.

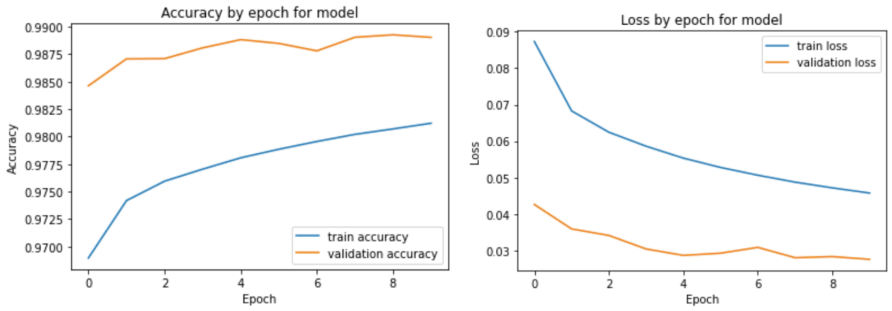

Figure 6: Accuracy and loss by epoch for the best model — model (c) — which is the the model initialized with weight trained on SatNet data.

6 Discussion

The performance of the best model yielded a precision score of 0.83, recall score of 0.81, F1 score of 0.82, and accuracy of 0.99. We empirically validate this by comparing the predictions across all three objects for vehicles, roads, and buildings. The predictions for vehicles are highly precise and are able to capture vehicles on main roads, residential streets, and parked alongside buildings. The predictions for the roads are able to properly label the pixels associated with main roads as well as residential roads. The predictions for buildings are able to capture explicit buildings and exclude regions such as empty plots of land adjacent to buildings as well as the shadows associated with buildings. This illustrates that the predictions we obtained for our model are fairly robust.

Figure 7: Input image (left) and predictions for vehicles (right) in blue

To do this, we defined a high vehicle region as a region where predictions exceed a threshold of 0.30 and low vehicle region as a region where predictions are below a threshold of 0.30. We then compare the LST between the two regions based on a threshold and find differences in the mean: high vehicle density region is 44.5 Celsius and low vehicle density region has an average of 43 degrees Celsius. We observed a difference of 1.5 degrees Celsius in the mean LST between high vehicle density and low vehicle density area. This is consistent with our expectations with the UHI effect, as we would expect regions with more vehicles to emit emissions and exhaust in a way that would increase LST even if marginally.

We additionally wanted to examine the relationship between LST and our model's prediction of buildings throughout the city of Riyadh. We define the urban area of Riyadh and rural area of Riyadh based on vehicle density and generate a mask to extract only the urban or rural region. For each urban and rural region, we then plot a boxplot of LST against deciles of building prediction. For city regions, contrary to what we expected, we observe a decrease in LST and buildings density increases. This suggests that buildings have a cooling effect. This may be due to the light color of buildings in

Figure 8: Input image (left) and predictions for vehicles in blue and roads in purple (right)

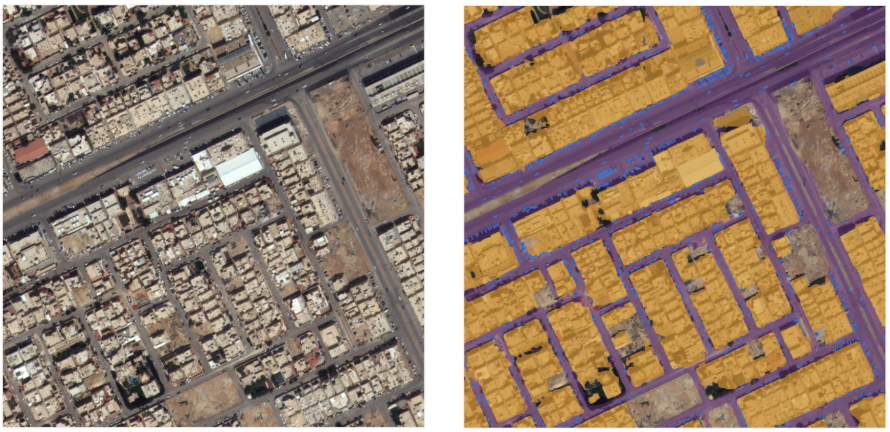

Figure 9: Input image (left) and predictions for vehicles in blue, roads in purple, and buildings in orange (right)

Figure 10: Low vehicle density area (left) and high vehicle density area (right)

the city, additionally, another explanation could be the function of wind speed between buildings, which is supported by previous research that found that wind speed was a key factor in providing a cooling effect in cities [5]. For rural regions, we find that as building density increases we also find a decrease in LST, but this decrease is less pronounced compared to urban areas. Our original hypothesis is that because buildings have a lighter color than its surroundings, urban settings are more effective at reflecting heat so we would expect urban areas to have higher heat. However, in our

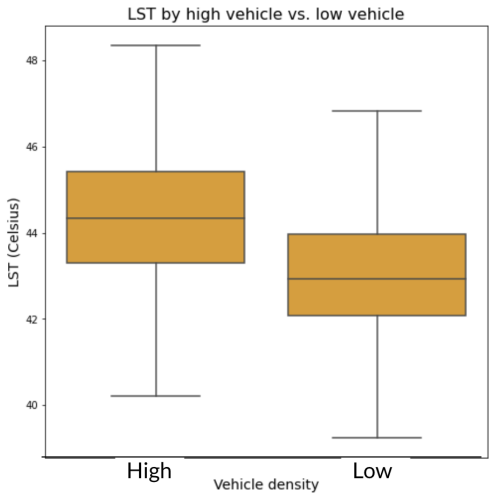

Figure 11: LandSat Temperature (LST) by high vehicle density vs. low vehicle density

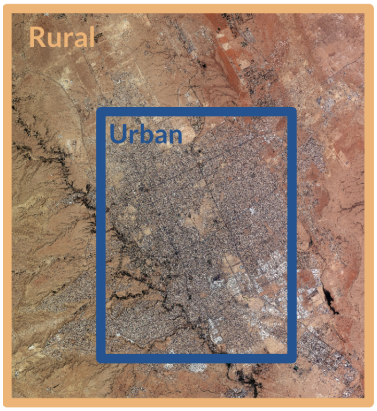

Figure 12: Rural vs. urban region of Riyadh, Saudi Arabia

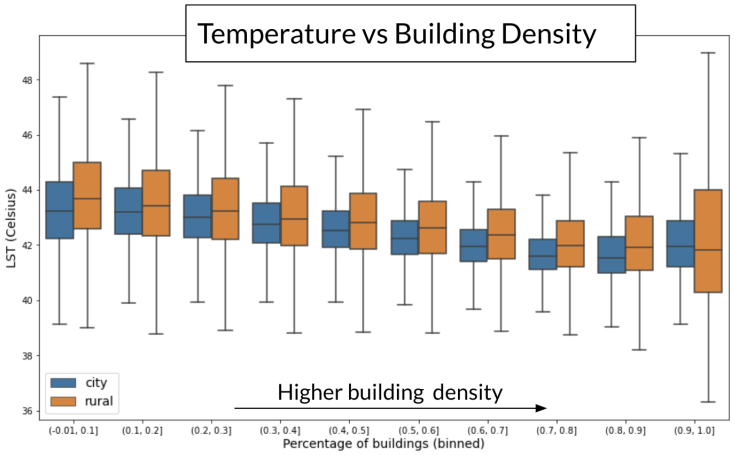

Figure 13: Rural vs. urban region of Riyadh, Saudi Arabia

case, we do not see the urban heat island effect in our case because the temperatures in rural areas are higher on average. This may be a function of rural areas not having any form of vegetation that would help cool the area in a way that would result in lower LST.

Conclusions

In this case study, we demonstrate the value of applying machine learning techniques to satellite image data in service of predicting elements of urban infrastructure such as vehicles, roads, and buildings. We do this specifically in the city of Riyadh, Saudi Arabia, the first of its kind in the academic literature. This is a unique case study as the city of Riyadh is rapidly developing. We also demonstrate the value of using the U-Net infrastructure to help with semantic segmentation of satellite image data, re-purposing it from its original use for biomedical images. We additionally observe a relationship between the urban infrastructure and LST. However, the relationship is not consistent across all types of urban infrastructure. For example, we observe a positive relationship between vehicle and road density and LST, which is consistent with our hypothesis. However, we observe a negative relationship between buildings and LST in both urban and rural areas, suggesting that buildings may offer some protective cooling effect. We hypothesize this may be a function of higher wind speed that occurs between buildings in areas with high building density.

Future research can perform an analysis on potential confounding variables for the urban heat island effect. For example, regions with roads that are highly connected in the urban core of Riyadh may have very different relationship to LST compared to regions with roads on the periphery of town. Additionally, future research can examine the temporal trends in LST to understand if the amplitudes of LST are progressively increased in a way that may be indicative of climate change. Other research can examine, predict, and analyze the presence of vegetation in the city of Riyadh to understand whether vegetation or other forms of landscaping have a cooling impact on LST.

Broader Impact

The goal of this research is to understand the merits of ML methods to help with detection of urban infrastructure to correlate and quantify the relationship between urban infrastructure and climate change. This work can help with other applications of object detection using satellite image data which have a broad range of societal implications. Of particular concern is the use of these techniques to carry out mass surveillance or target specific households or neighborhoods for malicious or nefarious purposes. We would encourage further work to name, acknowledge, and understand the potential biases and uses this technology can be used for. We do not stand, support, or condone unethical uses of the techniques implemented in this research paper.

Acknowledgments and Disclosure of Funding

We thank Saurabh Amin for his mentorship, support, and guidance as part of MIT's Machine Learning for Sustainable Systems course (MIT 1.C01/1.C51). We also thank CCS for providing the data.

References

[1] Rolf, E., Proctor, J., Carleton, T. et al. A generalizable and accessible approach to machine learning with global satellite imagery. Nat Commun 12, 4392 (2021). https://doi.org/10.1038/s41467-021-24638-z.

[2] Q. Jiang, L. Cao, M. Cheng, C. Wang and J. Li, "Deep neural networks-based vehicle detection in satellite images," 2015 International Symposium on Bioelectronics and Bioinformatics (ISBB), 2015, pp. 184-187, doi: 10.1109/ISBB.2015.7344954.

[3] X. Chen, S. Xiang, C. -L. Liu and C. -H. Pan, "Vehicle Detection in Satellite Images by Parallel Deep Convolutional Neural Networks," 2013 2nd IAPR Asian Conference on Pattern Recognition, 2013, pp. 181-185, doi: 10.1109/ACPR.2013.33.

[4] Jorge García-González, Miguel A. Molina-Cabello, Rafael M. Luque-Baena, Juan M. Ortiz-de-Lazcano-Lobato, Ezequiel López-Rubio, Road pollution estimation from vehicle tracking in surveillance videos by deep convolutional neural networks, Applied Soft Computing, Volume 113, Part B, 2021, 107950, ISSN 1568-4946.

[5] Chen, Wei, Jianjun Zhang, Xuelian Shi, and Shidong Liu. "Impacts of Building Features on the Cooling Effect of Vegetation in Community-Based MicroClimate: Recognition, Measurement and Simulation from a Case Study of Beijing." International Journal of Environmental Research and Public Health 17, no. 23 (December 2020): 8915. https://doi.org/10.3390/ijerph17238915.