The Effect of Image Resolution on Remote Sensing of Energy Infrastructure.*

Johannes M. Halkenhäußer
j.halkenhaeusser@students.hertie-school.org

Dinah Rabe
d.rabe@students.hertie-school.org

Hertie School, Berlin (GER)

May 2, 2023

Abstract

Accurate data about the structure of the electricity grid is crucial for all stakeholders working on the Seventh Sustainable Development Goal which aims for clean and affordable energy for everyone, especially in developing countries. Conventional ways of collecting this data, mostly relying on manual work by humans, are cost-intensive, slow, and hard to scale. Meanwhile, recent advancements in deep learning models have enabled automated localisation and classification of objects in imagery, referred to as object detection. These models, such as *Faster-RCNN*, have proven their capability to detect even small objects like solar panels or electricity towers in satellite imagery. However, it remains unclear to what extent the spatial resolution of these images affects the performance of such an object detection model, since earlier studies have always relied on satellite imagery with the highest, yet costly, available resolution of 30 cm/pixel. Therefore, in this work, we explore three questions: (1) How does a decrease in image resolution affect the ability of an object detection model to find distribution and transmission towers in satellite imagery, (2) what is the effect of changing the composition of geographies within the training set, and (3) how do differences in electricity tower sizes impact the detection performance? In our experiments we observe that the networks' precision deteriorates as we gradually decrease the resolution. When focusing on subsets of geographies we find, besides a high in-sample precision, that the inclusion of a variety of biotopes can lead to marginal improvements in the otherwise poor out-of-sample performance. Finally, the results suggest that the detection performance of the model, when specifically trained to detect larger towers, is more resilient to a decrease in resolution, even when size variance is high.[1]

Word Count: 11,432

*Master Thesis submitted for the M.Sc. Data Science for Public Policy (Cohort 2023), supervised by Prof. Lynn Kaack, PhD (Hertie School) and in partnership with Lukas Franken (University of Edinburgh)

[1]This thesis project is open source and available on Github

Acknowledgements

We are immensely grateful for the constructive feedback, guidance, and encouragement from Lukas Franken and Lynn Kaack. Further, we would like to thank Kyle Bradbury and Jordan Malof for the exciting conversations about our topic as well as Simon, Adrian, Benedikt, and Rodrigo for their valuable comments. Another note of appreciation has to go to Huy Dang for his tireless technical support.

Contents

List of Figures

List of Tables

List of Abbreviations

AP Average Precision. 10

AP50 Average Precision with decision boundary 0.5. 11

Faster-RCNN Faster Region-Based Convolutional Neural Network. 10

FPN Feature Pyramid Network. 10

IoU Intersection over Union. 11

LGRS Location-Generic-Resolution-Specific. 12, 21, 33

LOO Leave-One-Out. 21–23, 25, 29

ResNet Residual Neural Network. 10

ROI Head Region of Interest Head. 10

RPN Region Proposal Network. 10

SDG Sustainable Development Goal. 1

UAV Unmanned Aerial Vehicle. 4

YOLOv5 You Only Look Once Model - Version 5. 5

1 Introduction

The United Nations' seventh Sustainable Development Goal (SDG) advances the provision of affordable and clean energy for all (United Nations, 2018). Electricity grids are a vital component of this vision. The electrification of industry and households is increasingly paralleled by more decentralised energy systems, in which a significant share of electricity is generated by intermittent power generation sources such as wind and solar. These developments are putting pressure on the world's electricity grids, which are in need of significant planning, expansion and refurbishment to cope with the zero- or low-carbon energy systems of the future (Medjroubi, Müller, Scharf, Matke, & Kleinhans, 2017; Ren, Malof, et al., 2022). Reliable and up-to-date data on the location of grid infrastructure, such as transmission lines or towers, is the very least that policymakers and transmission system operators need in order to create the (political) roadmap for achieving clean and, at the same time, affordable energy. In the Global South, however, this data is often sparse, outdated, or incomplete (Arderne, Zorn, Nicolas, & Koks, 2020; Huang et al., 2021). To add to the challenge, the process of generating or updating information on grid infrastructure is costly: Oftentimes, data is only available through cost-intensive surveys, especially in contexts with decentralised grid structures (Castello, Roquette, Esguerra, Guerra, & Scartezzini, 2019), rendering the data acquisition a highly cost-intensive, slow, and hard-to-scale exercise. The fact that many use cases require the most up-to-date information on grid infrastructure, and therefore constant iterations of the data collection process, only exacerbate this issue. This bottleneck motivates researchers, international organizations, and private actors to investigate and find more resource-effective approaches.

Detection and prediction of energy infrastructure in satellite imagery. The existing research and modelling approaches for detecting energy grid infrastructure operate at three levels of analysis: the micro-level (focused on individual installations), the macro-level (ultimately focused on a broader socio-economic outcome) and the meso-level (focused on the grid itself).

Most research on the individual-unit level investigates the automated detection of solar panels in overhead imagery. While (Kruitwagen et al., 2021) find that individ-

ual commercial-, industrial- and utility-scale solar-panel installations are detectable on overhead imagery of a resolution of 10 m/pixel,[2] Ren, Malof, et al. (2022) systematically investigate the impact of image resolution on the detectability of small solar home systems. They find that for small solar panels, a minimum resolution of 15 cm/pixel is needed, thus scaling automated detection through satellite imagery becomes impossible. In their approach termed *Gridfinder* Facebook researchers focus on the macro-outcome level, specifically on the notion of 'accessibility' (Arderne et al., 2020). Using open data and nightlight imagery they predict electricity grid locations and subsequently derive the percentage of the global population living within 10km of a grid line. In that, they only focus on the grid only as a means to better understand a broader macro-outcome. However, with only one estimate per square kilometre, the *Gridfinder* data is too imprecise for detailed grid management and planning. The World Bank in partnership with the private organisation Development Seed (*Mapping the electric grid*, 2018) and Huang et al. (2021) with *GridTracer* both focus on a meso-level of analysis: the structure of the electricity grid. While the approach employed by the World Bank and Development Seed requires a human in the loop, rendering scaling very cost-intensive, *GridTracer* is the first fully automated approach to power grid mapping. So far, this attempt is exclusively focused on regions where sufficiently granular data is available (namely the US and New Zealand), using imagery with a very high resolution of 30 cm/pixel.

Contribution of this work. Of the three existing strands of literature, work at the meso-level promises the greatest opportunity to produce data sufficiently granular for grid planning, expansion, and management. In this line of research, there is a gap in knowledge regarding the impact of image resolution on power grid tower detection. In this paper, we aim to help fill this gap by investigating the possible extension of *GridTracer* to other locations in the Global South. In a first step, employing state-of-the-art detection models (namely Wu, Kirillov, Massa, Lo, and Girshick (2019)'s *Detectron2*), we study the ability of a location-generic model and understand its sensitivity to decreasing spatial resolution. We find that the task of electricity tower detection is too context-dependent, to aim for one detection model that is able to generalize across locations.

[2]We use the terms "spatial resolution" and "resolution" interchangeably within this work referring to the concept of spatial resolution as centimetre or meter per pixel (cm/pixel, m/pixel). Other aspects of resolution, namely spectral, radiometric and temporal, are not the focus of this paper.

Next, we simulate a situation in which there is no training data available for a location of interest to examine a potential use case for data-sharing co-operations between regions. To do so, we conduct experiments to test the out-of-sample prediction performance of a detection model. We find that it is crucial to train the model with data from the respective location of interest and that when such data is not available the low model performance can be slightly improved when using data of locations with a similar geography. Finally, based on the existing knowledge about the impact of object size on detectability (Ren, Malof, et al., 2022), we investigate the impact of tower size on detection performance. We find that, independent of resolution, detection capability for larger towers is higher than for smaller towers. Crucially, we observe that the model can detect large towers even in imagery with resolutions so low that distribution towers are invisible to the human eye (2-3 m/pixel).

With our findings, we contribute to the approaches that aim to address the unique challenges of large-scale mapping of power grid infrastructure in overhead imagery. Drawing conclusion, we can attest (to the rather trivial intuition) that practitioners aiming to build models for a particular region of interest should seek out imagery of the highest possible resolution. They should also focus on compiling the training dataset from images of the same geographical environment as the region of interest, as general out-of-sample approaches are not promising.

Our results indicate that detection models are very sensitive to decreasing spatial resolutions. If one wanted to define a threshold at which it is impossible to detect towers (e.g. to be able to procure less costly imagery), such a threshold will depend heavily on the specifications of the location of interest, e.g. the size of the towers to be detected, the image noise level and geographical environment. We hope that our insights support further research addressing the problem of large-scale automated power grid detection to rapidly develop practical solutions that can provide policy and other decision-makers with reliable and up-to-date data on the electricity grid to achieve the seventh Sustainable Development Goal.

2 Related Work

There are two research areas that are relevant to our project. One is the more general area of energy infrastructure and remote sensing and the other is the question of the impact of image resolution in remote sensing.

Energy Infrastructure and Remote Sensing. The field of energy infrastructure and remote sensing has been an active research field in the last two decades. Ren, Hu, et al. (2022) systematically classify and analyze the research on automated extraction of energy systems information published between 2000 and 2020. They find that especially the research in the last five years has mainly relies on deep learning methods for analyzing remotely sensed data due to its size and complexity. These models are used to solve mostly object detection or image segmentation tasks. The authors find that for research on transmission infrastructure drone or aerial imagery (<30 cm/pixel) or very high-resolution satellite imagery (30 - 50 cm/pixel) is predominantly used. Besides a substantial number of research on electricity generation infrastructure, in particular, on solar panel detection (Golovko et al., 2017; Kruitwagen et al., 2021; Malof, Collins, Bradbury, & Newell, 2016; Malof, Rui Hou, Collins, Bradbury, & Newell, 2015), a smaller sub-field of research focuses on transmission and distribution infrastructure (Chen & Miao, 2020; Han & Wang, 2017; Hu et al., 2018). The emphasis of this subfield lies on the decentralised and individual unit of analysis.

While much of the research so far has used exclusively UAV imagery for tower detection Huang et al. (2021) developed and publicly released a dataset of satellite imagery including annotations for transmission towers and power grid lines with a resolution of 30 cm/pixel. Additionally, they develop and present a combination of deep learning models called *GridTracer* that tackle large-scale tower detection and power grid graph inference. Doing so, they move beyond the detections of individual units, towards the grid level of analysis. The existing research shows that machine learning techniques have proven to be very useful for analyzing energy systems, however, they almost exclusively employ very-high-resolution imagery. Ren, Hu, et al. (2022) find that specific research is needed to understand which minimum resolution of data is needed to adequately address a certain object detection task to allow for cost-effective scaling.

Impact of Spatial Resolution on Remote Sensing. We add a puzzle piece to the research exploring the effects of resolutions on object detection. There is no general research on the impact of image resolution on the performance of object detection models using remotely sensed data because the needed resolution of imagery heavily depends on the task at hand (Ren, Hu, et al., 2022). A variety of object characteristics (such as width, height, colour) as well as the context they are in (background colour, noise through other objects, etc.) all have an impact on the objects' detectability (Fig. 1).

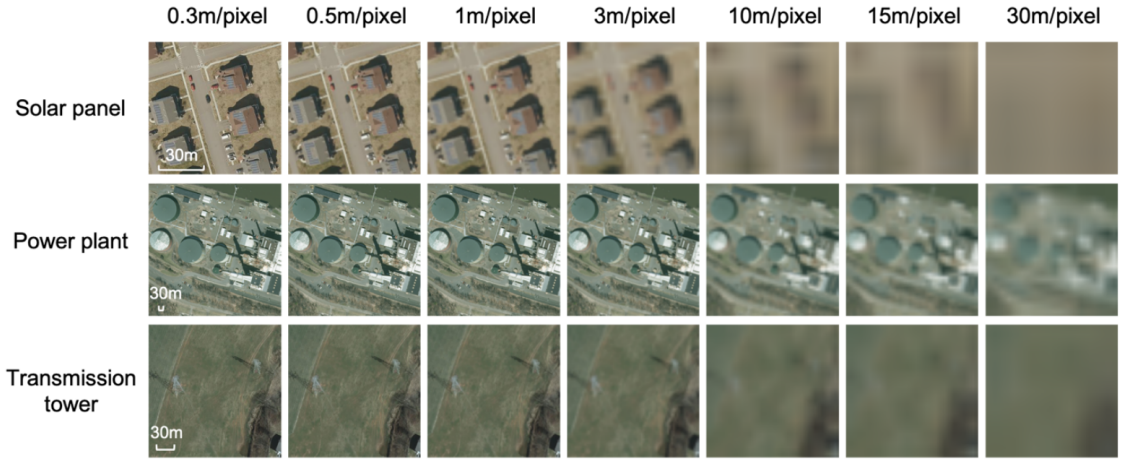


Figure 1: Energy infrastructure at varying resolutions. Images visualize the visibility of solar panels, a power plant and a transmission tower with decreasing resolution, adapted from (Ren, Hu, et al., 2022).

Given the task dependency, research on specific objects evaluates the model performance only for the resolution and task at hand. Thus, there is only an unsystematic overview of which resolutions are most likely needed to detect certain objects, e.g., ≤ 4.5 m/pixel for objects like swimming pools, vehicles, and planes according to Ding et al. (2022) or ≤ 10 m/pixel to detect and classify ships and vessels according to Ciocarlan and Stoian (2021). Kaack, Chen, and Morgan (2019) find that trucks can be identified in images with 31 cm/pixel resolution but report that obtaining viable results proved difficult at lower resolution. There are only few systematic analyses of varying spatial resolution. Brown, Qiao, et al. (2022) investigate the correlation between object detection performance and spatial degradation for animals, relevant in livestock management contexts. To investigate this relationship, they artificially downsample aerial images to satellite resolutions and repeatedly evaluate a YOLOv5 object detector on the produced

data. They find that detection performance drops steeply around 50 cm/pixel ground sampling distance. To investigate the detection performance of small solar home systems in drone imagery, Ren, Malof, et al. (2022) create a dataset of varying spatial resolution by collecting the same imagery from different altitudes. They evaluate object detection performance with a U-Net model and find that performance drops substantially for spatial resolutions below 15 cm/pixel.

With our work we build on the research of Brown, Qiao, et al. (2022); Huang et al. (2021); Ren, Hu, et al. (2022) seeking to fill the research gap for the impact of spatial resolution for the task of automated electricity tower detection.

3 Methodology

3.1 Data

We employ the data used in Huang et al. (2021), which is overhead imagery collected by drones and satellites. The images are taken in various locations across the world including New Zealand, the USA, China, Sudan, and Mexico. The dataset consists of 512 individual files that each contain some kind of electricity grid tower, either *Transmission*, *Distribution*, or *Other* towers. In contrast to Huang et al. (2021)[3] we include all locations for which data is available at a resolution $\leq$ 30 cm/pixel to ensure that our approach is applicable across various geographic locations. To ensure comparison with their results and to provide comparability between locations, all images are brought to a base resolution of 30 cm/pixel (see 3.2).

The dataset features a huge variety of geographical features (forest, desert, grassland) and building density (countryside, village, city) (Fig. 2). *Other* towers include cases for which annotators were uncertain (e.g., streetlights) and represent a negligible fraction of the total number of towers and are therefore excluded. Distribution towers and transmission towers are different in size and tower sizes vary between locations. Further descriptive details about the dataset can be found in Huang et al. (2021). The distribution towers outnumber the transmission towers by roughly 10:1 leading to a strongly unbalanced dataset.

[3]The authors only use data from Arizona (USA), Kansas (USA) and five cities in New Zealand

Figure 2: Example imagery. Images represent the variance across locations. Top left to bottom right: Arizona, Kansas, Rotorua, Clyde, Clyde, Sudan, Mexico, China.

During pre-processing, the large overhead images are cut into smaller 512x512 pixels sub-images. The towers are cut out using a random offset to create 16,189 individual sub-images each containing at least one tower. If by proximity a cut-out is made that contains further towers these are added to the annotation of the image. This means a tower may appear multiple times in the dataset (Fig. 3).

Next, we split the data into train, validation, and test sets on a rough 75/10/15 split stratifying on the underlying larger images, thereby avoiding any potential data leakage between the train and test sets. The split is not precise due to the stratification by original image and location, carrying with it the potential heterogeneity in tower count (as discussed above) that is propagated into the split (Fig. 4). Each sub-figure shows the distribution for one of the different tower types. In the last pre-processing step, the data is downsampled to lower spatial resolutions which is explained in the following section.

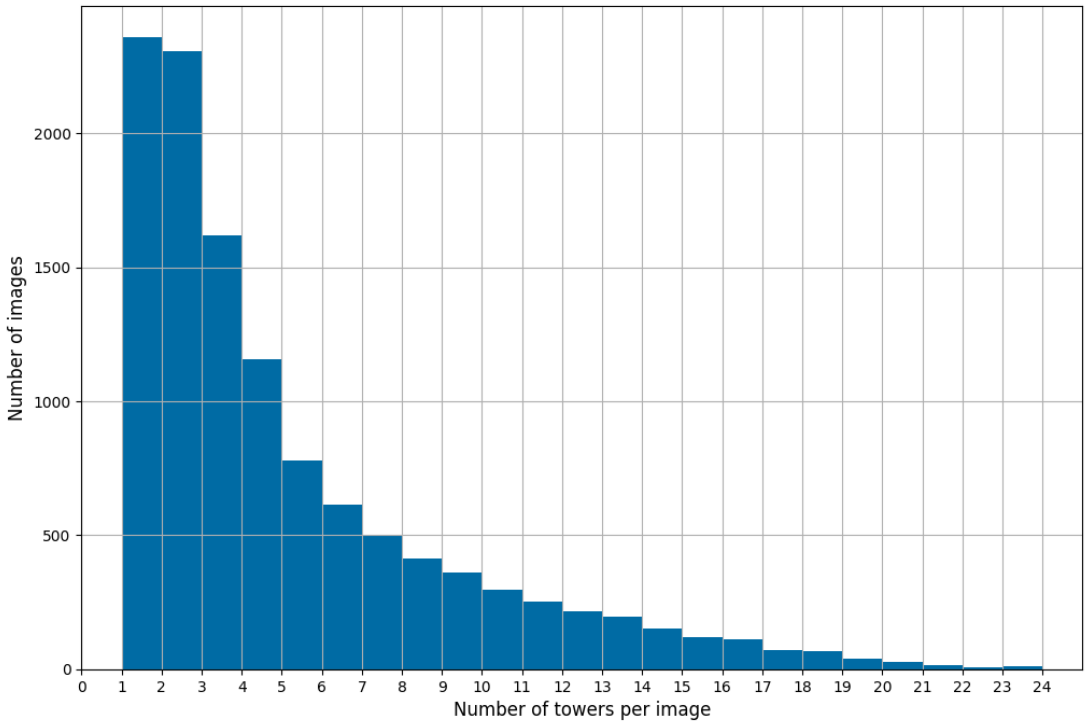
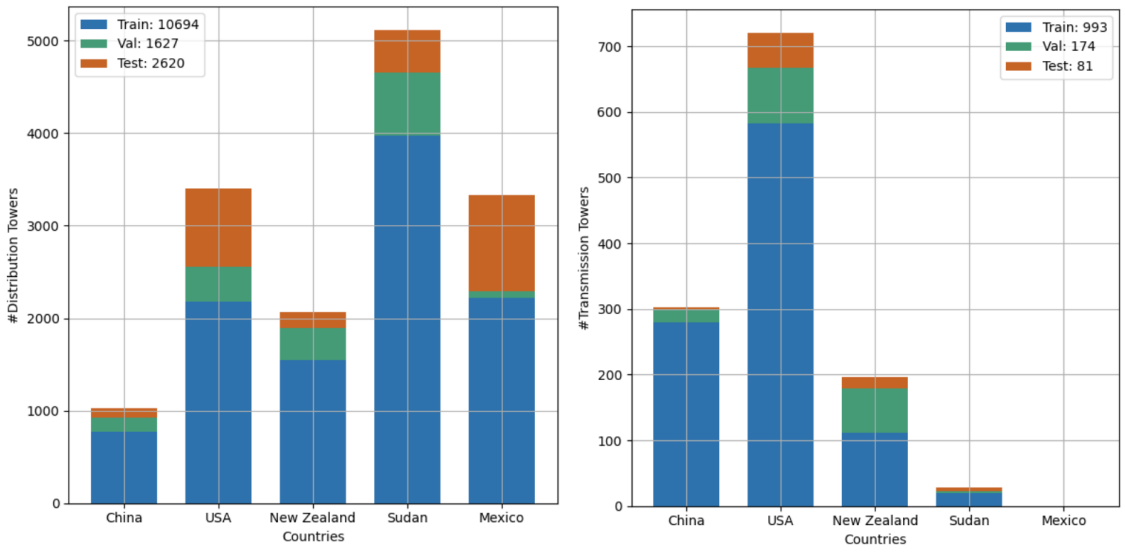


Figure 3: Tower count per image. The number of towers per image decreases exponentially with most images having only one tower. The upper limit is at 24 towers per image for a few images.



(a) Distribution Towers.

(b) Transmission Towers.

Figure 4: Tower count per tower type and region. There are many more transmissions than distribution towers and they appear in different frequencies within different locations.

3.2 Downsampling

Downsampling as defined in this work is the decrease of spatial resolution as measured in centimetres per pixel at constant image size. Even though our dataset contains imagery at a resolution of 15 cm/pixel, we decided to use 30 cm/pixel as our base resolution as it is the highest resolution commercially available for the entire world and widely used in research on object detection (Huang et al., 2021; Ren, Hu, et al., 2022). To keep our research relevant to practice we downsampled the image further to common resolutions of satellites: 30 cm/pixel, 50 cm/pixel, 70 cm/pixel, 100 cm/pixel (Ren, Hu, et al., 2022). We added smaller downsampling steps at the beginning, namely 35 cm/pixel, 40 cm/pixel and 45 cm/pixel and did not include resolutions $<$1 m/pixel due to the experience of fast performance decreases in similar experiments like (Brown, Qiao, et al., 2022) and Ren, Malof, et al. (2022). Through downsampling, we achieve that even though images were taken from different heights, as can be seen in Fig. 2, the spatial information stored in one pixel is the same across images.

Since the spatial resolution of the original images is available to us, we employ a downsampling approach which is based on the two-step process of reducing the size of an image to a lower number of pixels (e.g. 512x512 to 256x256) and then increasing the number of pixels back to the original size. If, for example, in the original image a pixel represented 15 cm and then the image size is reduced by half, the information from two pixels is combined into one pixel leading to a representation of 30 cm by one pixel. Increasing the image size afterwards again results in a reverted 512x512 pixels image that only contains the information of the 256x256 pixels image.

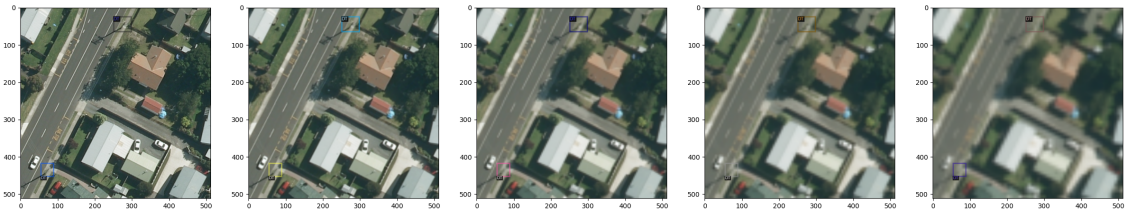


Figure 5: Visualization of downsampled images. An example image is downsampled from the 12 cm/pixel original resolution to 30 cm/pixel, 50 cm/pixel, 70 cm/pixel and 100 cm/pixel (left to right). Ground truth bounding boxes are shown for the two towers in the image.

To avoid aliasing artefacts in the downsampled images, which are common in sim-

ulated low-resolution imagery (Mitchell & Netravali, 1988), we use an approach which additionally smoothes over the imagery with a Gaussian kernel (Van der Walt et al., 2014). In line with the approach chosen by Huang et al. (2021), images are downsampled using bilinear interpolation (Fig. 5).

Potential challenges when downsampling a mix of drone and satellite images can be the differences in angles from which pictures are taken and the differences in lenses as discussed by Brown, Clark, Lomax, Rafique, and Sukkarieh (2022), but these potential issues are not relevant for the used dataset in this project (Huang et al., 2021).

3.3 Model

The development of deep learning towards ever more sophisticated models in size and architecture makes it both impractical and wasteful to rely on self-trained networks for complex modelling tasks. Hence, we take advantage of a publicly provided model and use transfer learning to adapt it to our needs. This involves loading an (often large) pre-trained model, locking its weights (freezing), and fine-tuning a number of top-level layers of the network. One framework that is able to perform strongly on a number of object detection tasks is *Detectron2* developed by Facebook AI Research (He, Gkioxari, Dollár, & Girshick, 2018; Wu et al., 2019). Among the models provided, we employ a Faster-RCNN Ren, He, Girshick, and Sun (2016) ResNet and Region Proposal Network (RPN) (R101-FPN) trained thrice over the Image Net data, achieving an Average Precision (AP) of 42 on the validation set. We refer to this pre-trained model specification as the *Detectron2* model.

A Faster Region-Based Convolutional Neural Network (Faster-RCNN) has two basic modules (Ren et al., 2016). The first is a fully connected convolutional network that outputs feature maps at different levels. In the case of detection, this backbone network is an FPN in which the input is passed through differently sized filters and the outputs of each layer are used to enrich the outputs of their predecessor (T.-Y. Lin et al., 2017). Our chosen *Detectron2* model is structured into a stem that has five Residual Neural Network (ResNet). The resultant feature maps are then passed into the second module which is broken down into a Region Proposal Network and a Region of Interest Head (ROI

Head). The first proposes a number of potential bounding boxes for objects based on the feature maps. The final ROI Head then takes these bounding boxes and the input from the feature mapping to make a classification into a class or *background*. From a model perspective, the goal is to minimize the losses in both the Region Proposal Network, as well as the ROI Head.

This two-step approach has proven to perform better than one-stage approaches on various benchmark tasks Ren et al. (2016) and was therefore adjusted for the particularly complex task of detecting small objects in satellite imagery by Azimi, Vig, Bahmanyar, Körner, and Reinartz (2018); Yang et al. (2018) and specifically for electricity tower detection by (Huang et al., 2021).

3.4 Metrics

We use the standard object detection metric mean average precision (mAP) (Padilla, Netto, & da Silva, 2020) for evaluation. The mean average precision is calculated by first finding the average precision (AP) for every class and then averaging over all classes. If only one class of objects is being detected (e.g., if all images only contain distribution towers) the AP of that class is equal to the mAP.

To calculate the average precision, we need a method to decide when detection is correct or incorrect. The standard approach for this is to calculate the Intersection over Union (IoU), which is the ratio of the area of overlap between the predicted box and the ground truth box to the area encompassed by both the predicted and the ground truth box. The IoU is then compared to a user-defined threshold to assign the label of correct (true positive) or incorrect (false positive) detection.[4]

The average precision for each class is subsequently the area under the precision-recall curve for a given threshold. To determine to which class the object in the predicted box belongs,[5] the model outputs a confidence score (softmax output) for each class. The highest confidence score determines the predicted class.

In our experiments, we use the so-called AP50 score (T. Lin et al., 2014). It represents

[4]False negatives in the object detection context are undetected ground-truth boxes. True negatives are not used in object detection because there is an infinite number of bounding boxes that should not be detected.

[5]if the dataset includes only one class the model predicts whether the object in the box belongs to the class or to the background.

a mAP score with an IoU threshold of 0.5 to determine whether a prediction is correct. We choose this low threshold as we are mainly concerned with the location of the towers rather than an accurate prediction of their size or shape. It is important to note that this metric ranges from 0 to 100 (instead of 0 to 1, which is more common).

4 Experiments

We conduct three experiments to investigate the impact of image resolution on the detection capability of electricity grid towers in satellite imagery. We begin by comparing the performance of fine-tuned *Detectron2* models on images with different resolutions (4.1). In 4.2, we explore the generalisability across regions and resolutions and in 4.3 we investigate the impact of tower size on model performance to test generalisability across object sizes. Each experiment section is further divided into a set-up section explaining the workflow and relevant parameters, a descriptive results section, and a discussion section, in which we analyse results and derive preliminary findings.

4.1 Experiment 1 - The Impact of Resolution

4.1.1 Set-Up

To investigate the impact of resolution on the capability to detect energy infrastructure using deep learning, we evaluate the performances of fine-tuned models on increasingly downsampled images of the same base data. First, we replicate the tower detection experiment from Huang et al. (2021), using the same model type, data subset and hyperparameters. Results for this pre-experiment are discussed in Appendix A.2. Further, we extend their findings by tuning the hyperparameters learning rate, weight decay, freezing depth, batch size, and gradient clipping and by training the model on a more diverse set of locations (Fig. 6). In the following, we will refer to the models from this experiment as Location-Generic-Resolution-Specific (LGRS) models.

We limit this experiment to the detection of distribution towers due to the fundamental differences in bounding box size, class imbalance, and distribution across locations (Fig. 4). Transmission towers are large power infrastructures, whose average size is 6491

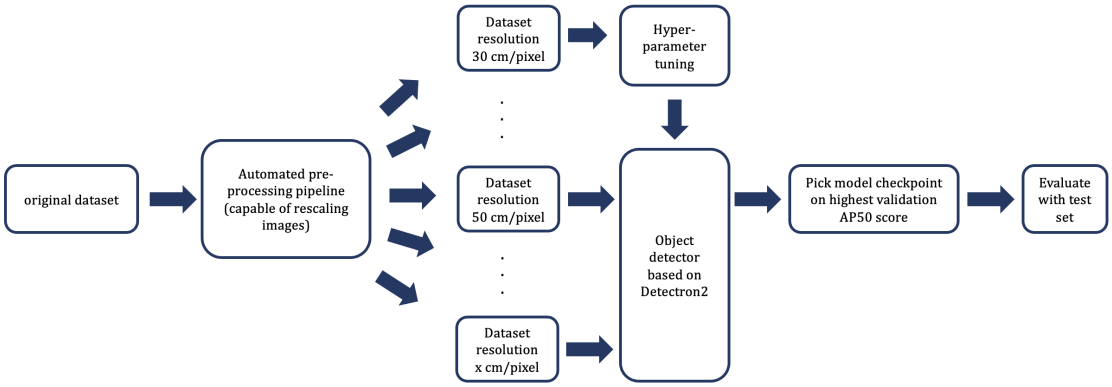


Figure 6: Workflow main experiment. We train a model for each resolution and evaluate the performance on the respective test set. Hyperparameter tuning is done on the data of resolution 30 cm/pixel.

pixels while distribution towers lie at merely 470 pixels. At the same time, only around 10% of the towers in our data are transmission towers. This class imbalance introduces further comparability constraints into the model evaluation. Further, differences in technical usage produce a change in the distribution of occurrences across the five locations. To mitigate basal noise, we remove transmission towers from our datasets for this experiment.

The standard sizes of the anchor boxes are 32^2, 64^2, 128^2, 256^2, and 512^2 pixels. After inspecting the distribution of anchor box sizes in the base data (Fig. 7), we reduced the anchor box sizes to 4^2, 8^2, 16^2, 32^2, 64^2.

We grid-search 48 different hyperparameter combinations and compare their maximum AP50 score on the validation set (Tab. 1). For comparability across batch sizes, we set the number of samples to be viewed to 120,000 and divide it by the batch size (24,000 iterations when batch size = 5 and 15,000 iterations when batch size = 8). The learning rate is reduced after 50%, 75% and 90% of the number of iterations. The models are evaluated every $\frac{1}{8}$ of the number of iterations. We train the full length of iterations and save the best model.The grid search produced no clear trend for a specific hyperparameter configuration outperforming across the board (see Appendix A.1 Fig. A.1). Thus, we use the top three performing configurations and train for the same number of epochs.

The final models are trained on artificially downsampled images with 30 cm/pixel, 35 cm/pixel, 40 cm/pixel, 45 cm/pixel, 50 cm/pixel, 70 cm/pixel, and 100 cm/pixel. We

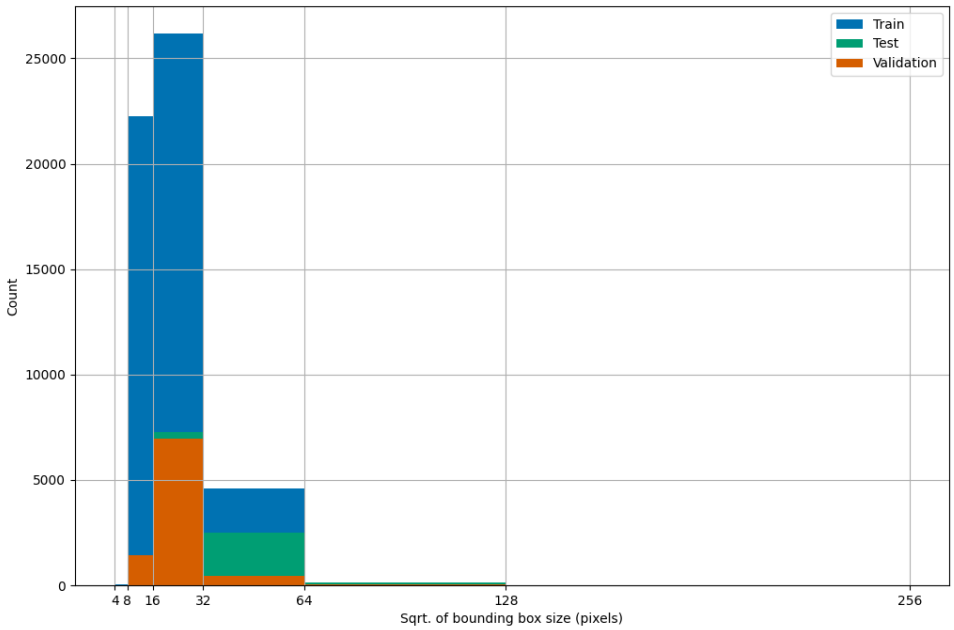


Figure 7: Size of ground truth boxes. Bounding box sizes (given as their square root) of the ground truth boxes for train, validation and test data. The bin sizes of the histogram represent the default box sizes of the model.

use the hyperparameters detailed above and three different configurations (see Tab. 1) providing greater robustness across the hyperparameter space. After training, the model checkpoint with the highest AP50 score on the validation set is evaluated on the test set with the corresponding resolution. We have removed the test images of Mexico due to irregularities discussed in 4.2, where location-specific results are discussed in greater depth.

4.1.2 Results

As expected, the performances of the models (for all three configurations), measured as the AP50 score on the test set, decrease with deteriorating resolutions (Fig 8).[6]

As expected, the best AP50 score of 39.4 is reached on the resolution of 30 cm/pixel. This AP score means that on average across all regions less than half of the towers detected are correct, which indicates that our model is performing rather poorly. In general,

[6]Results for the same models with the full test set are reported in App. A.4 Fig. A.2 and show the same behaviour but are systematically decreased by roughly 30%.

Hyperparameter	Description	Search Space	Configuration Rank		
			#1	#2	#3
Learning rate	Multiplication factor in weight change. Affects model's capability to reach a minimum of the loss function	[0.01, 0.001, 0.0001]	0.001	0.01	0.001
Batch size	No. of images per batch. Trades off learning time and validation accuracy	[5, 8]	5	8	8
Gradient clipping	Regularization and mitigation of exploding gradients	[True, False]	True	True	False
Weight decay	Regularization penalizing model complexity	[0.001, 0.0001]	0.001	0.0001	0.001
Freeze At	Sets the layer up to which the weights of the model are frozen during training	[1, 2]	1	2	2
AP50 Score	AP50 Score on the validation set during hyperparameter tuning with 30 cm/pixel		42.5	40.3	38.3

Table 1: Set-up for hyperparameter tuning. Overview of the chosen hyperparameter space and specifications for three selected configurations.

the different models' test performances are slightly worse than the performance on the validation set (configuration #1: AP50 39.4 $<$ 42.5, configuration #2: 37.2 $<$ 40.3, configuration #3: 35.3 $<$ 38.3). Focusing on the development of the curves for configurations #2 and #3, we can see that for the small steps between 30 cm/pixel and 40 cm/pixel model performance does not drop as stringently as expected. We even find a better performance of the models with configurations #2 & #3 on lower resolutions (e.g., configuration #3 40 cm/pixel $>$ 30 cm/pixel). For resolutions above 50 cm/pixel, the detection capability degrades faster and drops to an AP50 score around 16 at resolution 1.0 m/pixel and close to zero for resolutions above.[7]

In the following, model behaviour is visualised only for configuration #1 given that the model behaves consistently across configurations. In Fig. 9, the prediction capabil-

[7]We ran the same experiments and tests for resolution 3 m/pixel and 10 m/pixel for which model performance drops below 1. Our sample image in 5 is visualized in these resolutions in Appendix A.3 Fig. A.1

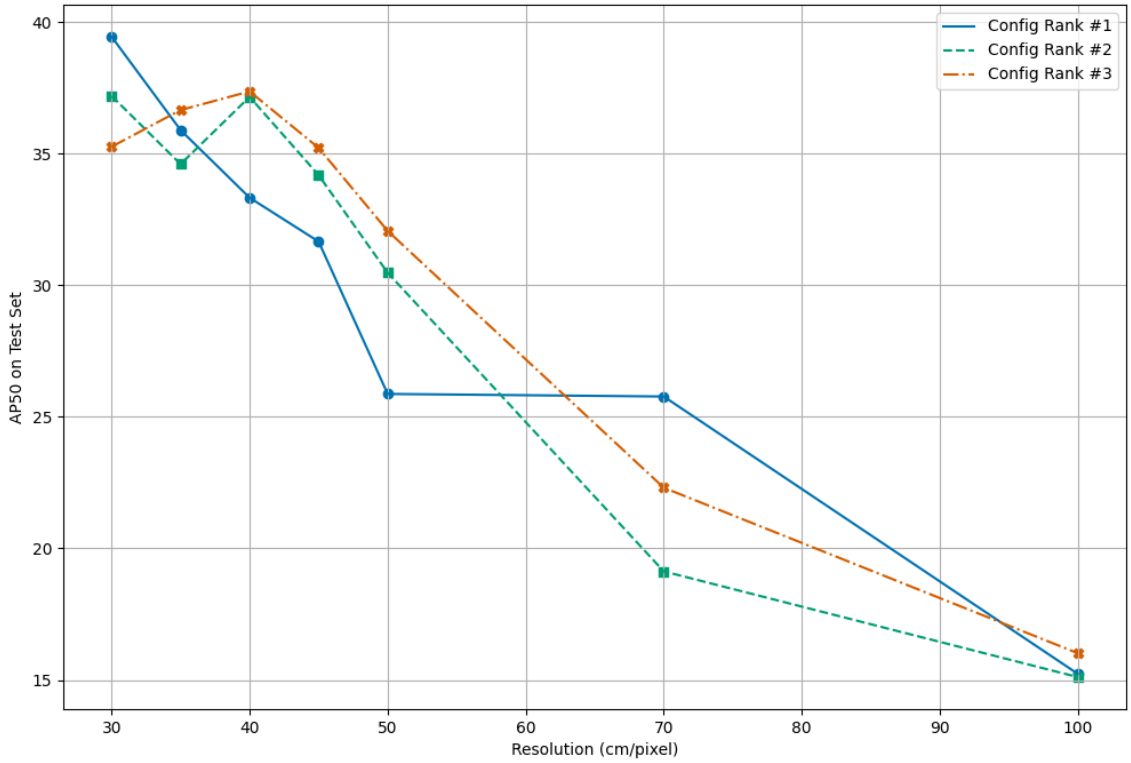


Figure 8: Effect of resolution on detection performance. AP50 scores for all three configurations on the test set for decreasing resolutions. The Mexican part of the test set is removed.

ity of our model is visualized for three exemplary images at different resolutions. The model predicts more of the ground truth boxes with higher confidence for the higher resolution imagery. In line with Fig. 8, performance appears very similar for resolutions 50 cm/pixel and 70 cm/pixel respectively. Decreasing model performance is also visible through decreasing confidence scores in the prediction of the tower class.

After an initial increase both classification and box regression loss decrease, resulting in a similar behaviour of the total loss (Fig. 10, explanation in 4.1.3). The total loss (Fig. 10c) is composed of the box regression loss (Fig. 10a) for the prediction of bounding boxes and the classification loss (Fig. 10b) for the prediction of the class *tower* (against the *background* class). Classification loss decreases more pronounced than box regression loss, but both curves seem to stabilize around the 14,000th iteration. An unexpected result is the development of the validation loss visualized in Fig. 10c together with the total loss. For almost all resolutions the validation loss steadily increases. There are only a few resolutions for which the loss stays constant at the beginning or even decreases,

Figure 9: Visualization of predicted bounding boxes. Predictions of our model (configuration #1) for resolutions 30 cm/pixel, 50 cm/pixel, 70 cm/pixel and 100 cm/pixel. The images in the left column show the ground truth labels.

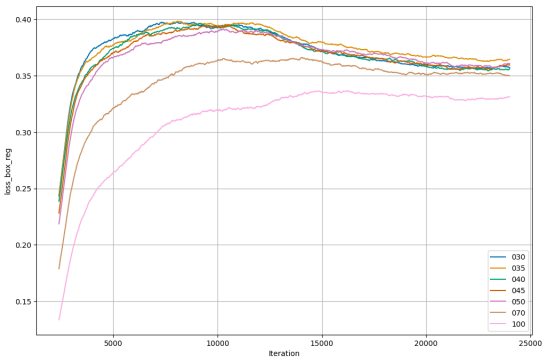
indicating that our model is overfitting.

Based on the performance on the respective validation set during training, the best model is selected for each resolution which is then finally used to test model performance on the respective test set (Fig. 11). Unsteady AP50 fluctuation with a slight easing towards the end of training was observed for all three configurations. This seemingly inconsistent behaviour is amplified by the fact that performance was only evaluated every 3,000 iterations.

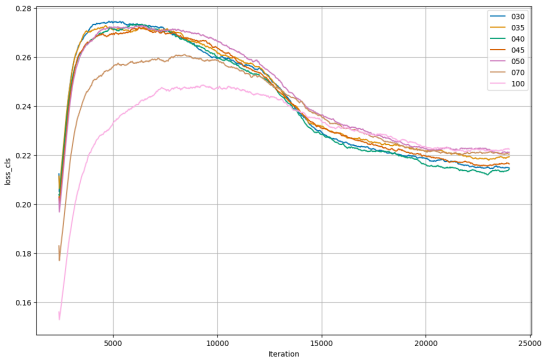
4.1.3 Discussion

Overall, the experiments yield strong evidence for a reduction in the detector's performance at lower levels of resolution. Nevertheless, the results are not conclusive enough to set a definitive threshold below which object detection of electricity infrastructure is no longer possible. The model performance is limited by a number of factors inhibiting the network's ability to achieve high performance, including the heterogeneity of the locations both in training and in testing data, and data quality. Further, capacity constraints introduced erratic metric behaviour during hyperparameter tuning.

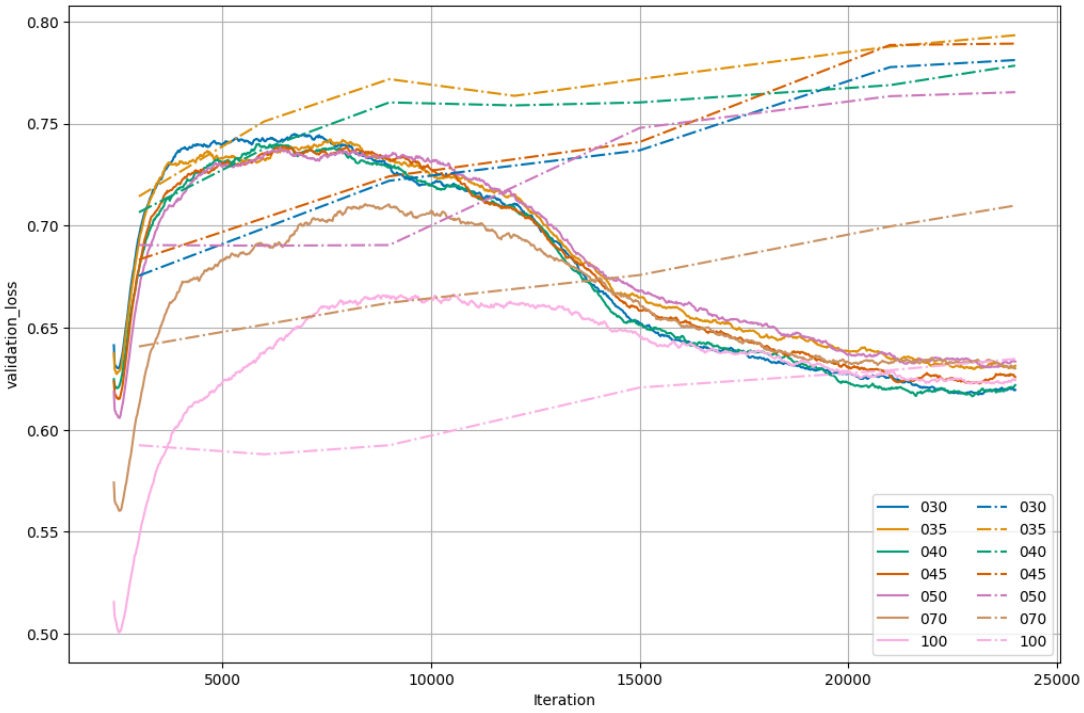
While our hyperparameter search allowed us to distinguish high performing config-



(a) Box Regression Loss.



(b) Classification Loss.



(c) Total (solid) and Validation loss (dashed).

Figure 10: Training loss curves. The box regression and classification losses decrease across resolutions (configuration #1). Consequently, the total loss also decreases. The validation loss (measured only every $\frac{1}{8}$ of iterations) increases almost throughout all resolutions and iterations. For better readability, the training's loss curves are averaged over a sliding window of size $\frac{1}{10}$ of the training iterations.

urations, its precision was limited by the long evaluation period. During hyperparameter tuning, the 48 runs were evaluated based on the highest AP50 metric. However, this evaluation is the most resource-intensive aspect of the workflow and thus could only be conducted every $\frac{1}{8}$th of the total run-time. Therefore, the best model checkpoint may have occurred between two evaluation checkpoints. The individual AP50

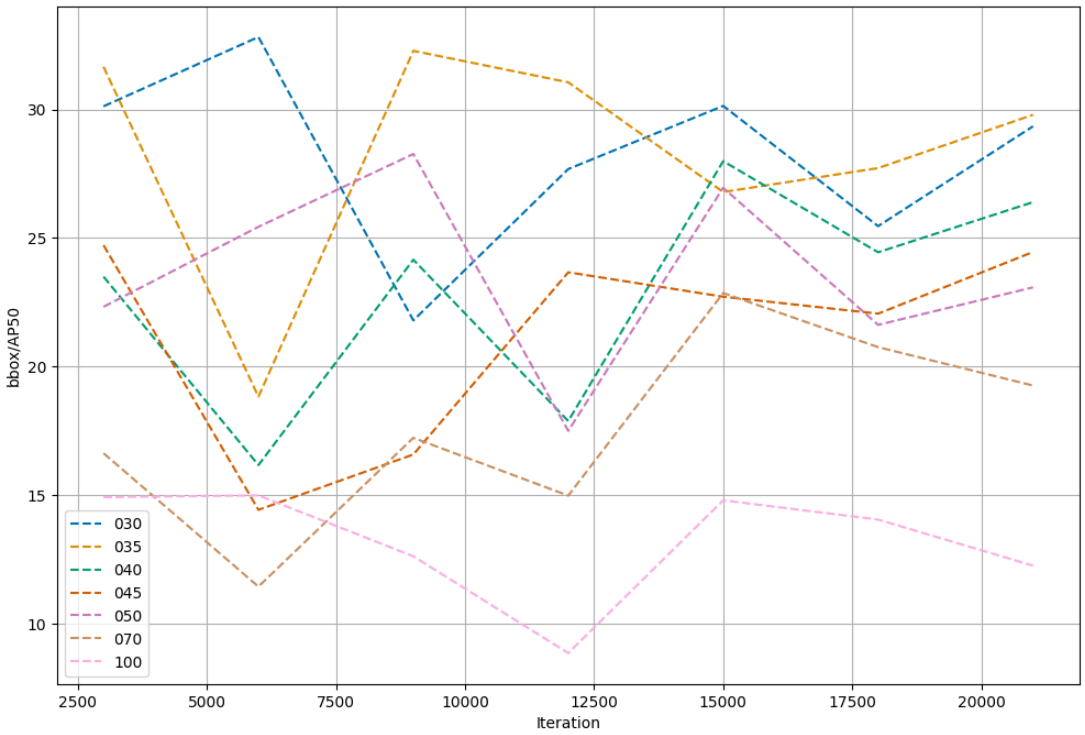


Figure 11: AP50 during validation. Configuration #1 AP50 scores during training for different iterations.

curves on the validation set for configuration #1 show strong fluctuation (Fig. 11). Thus, the chosen configuration of hyperparameters may be sub-optimal, and further, the best model for evaluating the test set may similarly be sub-optimal. This could lead to a situation in which the start of overfitting cannot be determined clearly. Nevertheless, each resolution-specific model is limited by this factor, thus making it a consistent noise across resolutions. Even with this additional variance, the expectation that performance should decrease with decreasing resolution is met (Fig. 8). Further, hyperparameter tuning was only conducted using data with a resolution of 30 cm/pixel. While we would expect an over-performance on the highest resolution, this cannot be observed in Fig. 8, e.g., for configuration #3. One explanation is the aforementioned variance in model selection both during hyperparameter tuning and model selection for the evaluation on the test set. The second explanation is that resolutions up to 50 cm/pixel all include sufficient information for the model to generally detect towers confidently.

As shown by Fig. 10, the model's total loss made up of the classification and box region losses is decreasing. The constituent parts are equally decreasing, thus showing

that the decrease in loss is not driven by one loss particularlyThe loss increases in the first periods, which is an expected behaviour given that in the loss function *background* predictions of the ROI heads are ignored (Girshick, 2015). In the beginning, the loss is very low, because the network proposes only a few boxes with a high enough confidence score to be included in the loss calculation. As the model "warms up", its predictions increase in confidence. At the same time, accuracy does not improve, leading to a higher loss. Finally, the model enters the phase in which it starts to learn the boxes correctly and the loss decreases.

Even though the model's training loss is decreasing, it overfits relatively quickly. The exact iterations of overfitting cannot be clearly determined, given the infrequent recordings of validation loss. Validation loss and the validation AP50 metric start deteriorating or peaking early in the modelling process. Thus, even though our model performance on the training set may be increased if the training time were to be prolonged, an improvement in the test and validation sets would not occur.

The performance is lower than in the Huang et al. (2021) performance (AP50 score of 52). However, they train their data on a more homogeneous set of locations, namely the US and New Zealand, while we explicitly wanted to understand the detection capabilities across locations. A 1:1 comparability cannot be assumed as we pre-process our images by cutting them into equal sizes with random offsets around the towers. Thus, our training data, while rooted in the same source, may differ from Huang et al. (2021).

Another limitation to increasing the model performance is the data quality. The images are from a variety of locations, some of which make it almost impossible to find the towers with a human eye (Figs. 2 and 9). For example, the images from Sudan show mostly urban areas, in dark tones and with small towers, making the prediction task particularly challenging. Urban areas can feature a variety of challenging objects that may be distribution towers at first sight (e.g., streetlights). The data annotators were in general agreement about tower locations. However, there were discrepancies in the sizes of bounding boxes drawn around the towers (Huang et al., 2021). One option to improve performance would have been to alter the loss function. For example, the Generalized IoU allows the calculation of different losses for non-overlapping bounding boxes. However, the improvement of Generalized IoU on both Mask-RCNN and Faster-RCNN is

negligible on common benchmarks such as MS COCO or PASCAL VOC (Rezatofighi et al., 2019).

An additional factor in model performance on the test set is that the training, validation, and testing data may differ. The train/validation/test split is stratified by base satellite image to avoid data leakage. Thus, even though drawn from the same locations, the images may feature differences in their underlying distributions. These differences, for example, further manifest in the shifting distributions of locations represented in train, validation, and test sets. These lead to the general model being evaluated on a slightly different distribution than it was trained on.

Overall, within the scope of our experiments, we could not determine a set threshold for tower detection due to rather poor baseline performances compared to human annotators (Huang et al., 2021). Further, any detection threshold would depend on the capacity to detect the entire electricity tower network beyond individual towers. We could nevertheless observe a decrease in performance with deteriorating resolutions, particularly with those lower than 50 cm/pixel.

Our LGRS models were confronted with the complexity of fitting to a diverse dataset in locations and object size. Therefore, we explore location-specific models in 4.2 and the impact of tower size in 4.3. A potential strategy that could be employed to enhance network detectability is post-processing using the information of grid structures which is elaborated upon further in 5.

4.2 Experiment 2 - Cross-Country Comparison

4.2.1 Set-Up

In this paper, we aim to make recommendations about the composition of potential training data to allow for predictions in a particular region of interest. One of the limitations in 4.1 is the model's limited ability to fit the heterogeneous data provided in the training set. One dimension of this heterogeneity consists of the different locations and the context they place the towers in. To understand the ability of the network to model a more homogeneous dataset and to then generalize to further datasets, we train a model for each location and test it on the other locations. Furthermore, we employ a Leave-One-

Out (LOO) strategy in which we train on all but one location and test on the other (Fig. 12).

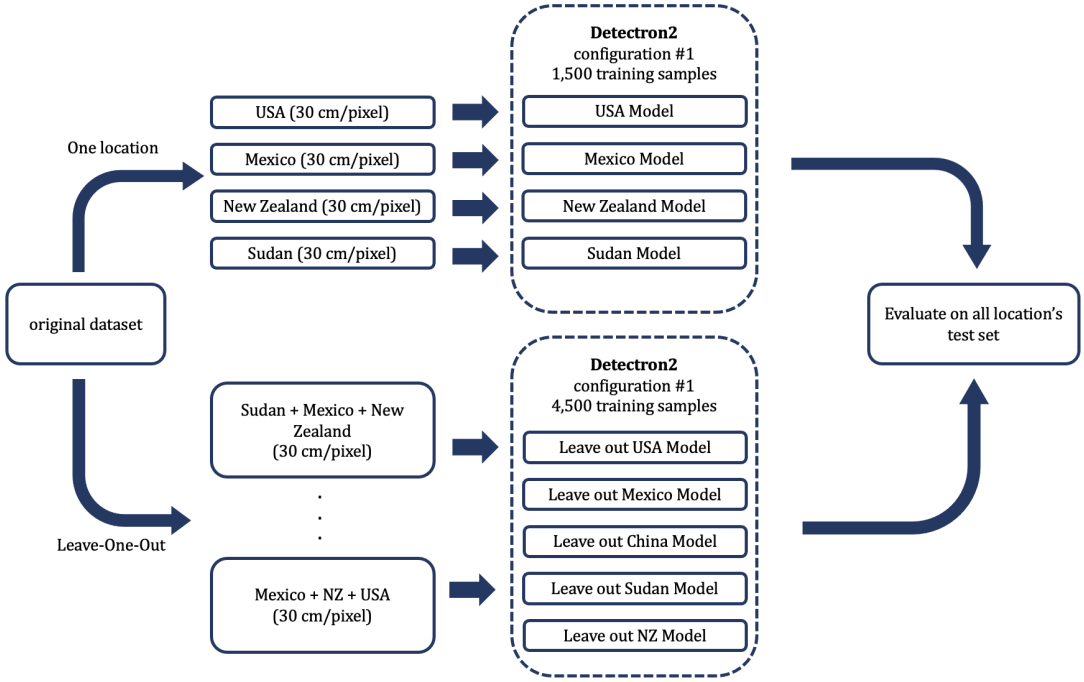


Figure 12: Workflow second experiment. In the second experiment, we compare countries. The upper part shows a one-location-per-model strategy the lower the leave-one-out approach.

To compare locations, a model will be trained based on the data of one of the locations. We choose to compare the locations on a national level giving us China, Mexico, USA, Sudan, and New Zealand. Because the locations have varying occurrences in the base dataset (Fig. 4a), the training sample size is restricted to a random selection of 1,500 images per country. Thus, in all combinations of LOO, we train with 4,500 samples.

China is left out of the model training, as only 750 images are available for this location. Nevertheless, we test on the Chinese test data with the individual-location models and specify an additional LOO model training with the four remaining locations, adjusting the ratio of images to 1,125 per location.

The original train/validation/test split for Mexico led to a disproportionally small validation set of only 60 images. To allow for an adequate validation set size, we reallocate images from the training to the validation set. The changed split leads to 2079 images in the train and 600 images in the validation set. The test set remains unchanged.

Given the decreased sample size compared to the LGRS models, overfitting is expected to occur earlier than during the first experiment. Hence, training time is reduced to 10,000 iterations and validation is performed every 500 iterations for the one-location model. The LOO has more data and is, therefore, trained for 15,000 iterations and evaluated every 750 iterations. The shorter evaluation intervals lead to a more precise detection of overfitting compared to in 4.1. Validation is only conducted within the respective samples, that is, on images from the same country/countries. Given that the underlying tower-detection task remains the same, the models are all trained with configuration #1 (Tab. 1) and 30 cm/pixel resolution imagery.

4.2.2 Results

Individual Models. The models on average perform similarly to the general model in 4.1 when tested in-sample but do so with great differences (minimum AP50 of 1.7 and maximum AP50 of 47) (Fig. 13). Further, they are unable to perform well on out-of-sample locations.

The models for each country do not perform consistently in-sample. Sudan (AP50: 47) outperforms the original model, scoring almost around 25% higher than the general model. The New Zealand (AP50: 32) and US model (AP50: 24) perform poorer than the model in 4.1 and the Mexican model performs the lowest (AP50: 1.7).

The models' out-sample scores are all below an AP50 score of 1. The only exceptions to this trend are models that are either trained or tested with the US data. The Mexican model performs better on US data than on its own. Likewise, the US-data model outperforms the Mexican-trained model on Mexican test data. Another exception is the out-of-sample performance achieved by the Sudanese model, which generalises to the Chinese test set with AP50 of 12.

The models train similarly, across the different locations and show similar behaviour in their loss and AP50 validation set curves to the original model (Appendix B.2 Fig. B.1). However, the loss curves peak earlier and then decline as the model fits the data. The validation losses increase throughout. The Mexican model shows no signs of training markedly different from the other locations, thus the lower test results cannot be

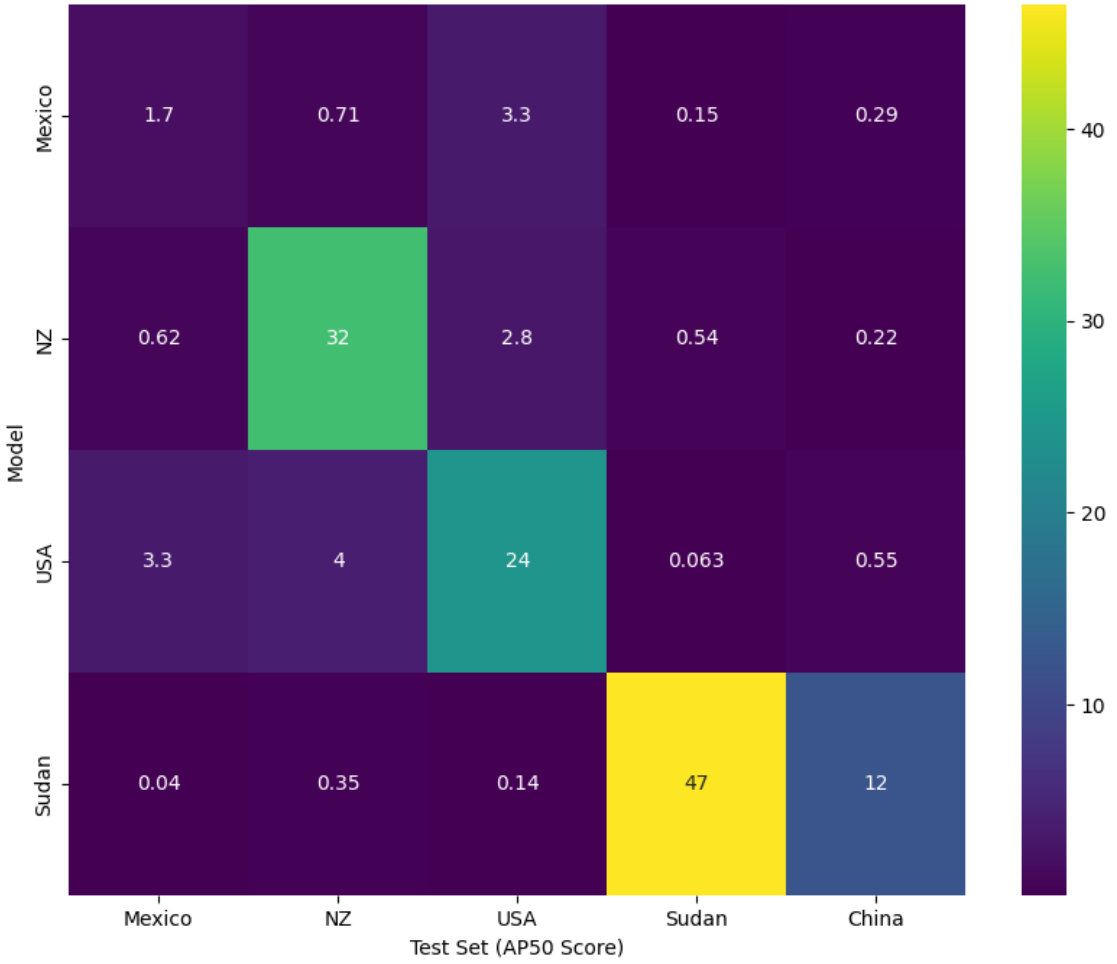


Figure 13: Model performance stratified by country. Heatmap of modelling and testing with different locations. The diagonal shows the performance on the in-sample test set and each row shows the performance of a model on the other countries.

attributed to a difference in training.

While the loss curves do not reveal a clear picture of overfitting, the AP50 validation scores (Fig. 14), paint a clearer picture than those of the LGRS models in 4.1 (Fig. 11). The models' performances on the respective validation sets increase until iteration 1,500/2,000 and drop afterwards. The behaviour confirms, that overfitting sets in early in the modelling process. The AP50 curves mirror the behaviour of the loss curves, as the stronger-performing models New Zealand and Sudan peak earlier than the weaker-performing US or Mexican models. The Sudanese model overfits the strongest, falling below the Mexican model. The Mexican validation scores reveal that the validation score performed similarly to the Sudanese and New Zealand model, scoring an AP50 score of 41 at its peak.

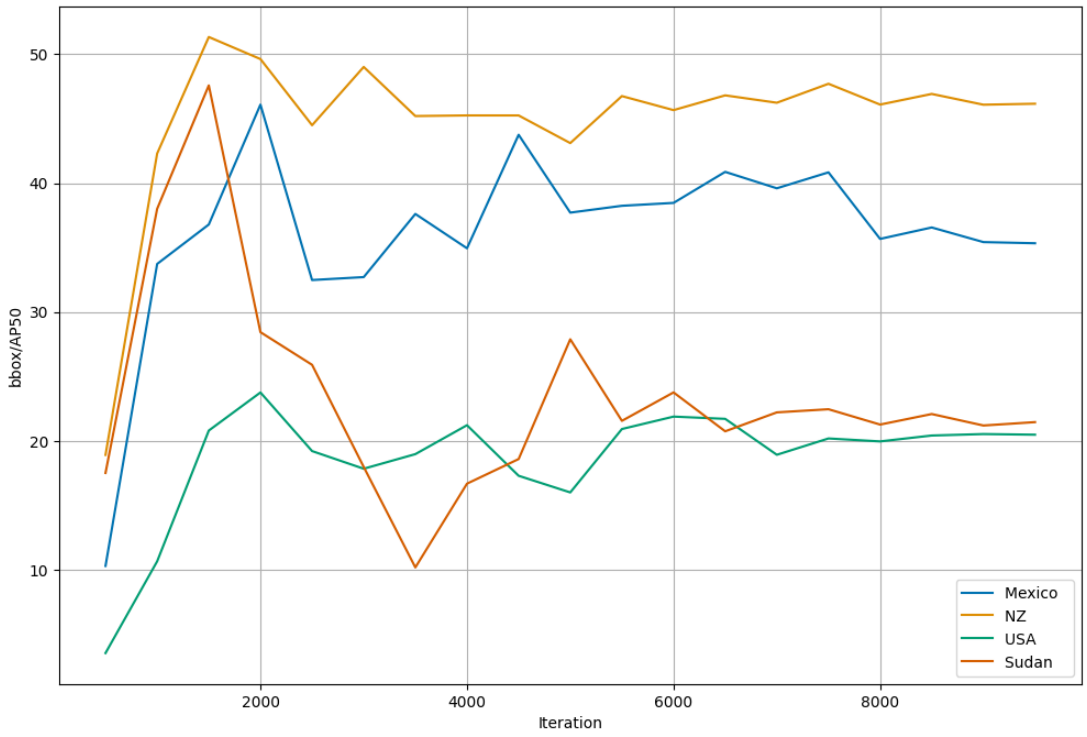


Figure 14: AP50 score during validation. AP50 score on in-sample validation set over the training iterations.

Leave-One-Out. Similar to the individual models, the LOO models perform very well on the in-sample test sets, except for the Mexican test sets. When a country is left out of the training data, then model performance is generally low when it is tested on this country's data (Fig. 15).

As observed already with the individual models, the performance of the Mexican test set is lower compared to the other locations. The Mexican model and test set perform the best when combined with the USA test set and model respectively.

Sudan performs the best out of all locations when included in the sample (>40) but the lowest when excluded (0.092). The only AP50 score lower than when Sudan is tested on a model where it is left out is China when tested on the same model. The inclusion of Sudan instead of New Zealand leads to a ten-fold increase in China's AP50 score (0.033 to 0.33). Switching in New Zealand for Mexico increases the score by another order of magnitude to 3.2. The score falls when the US is excluded and is lower again for the model with all the other locations.

The loss curves exhibit the same behaviour as previous sections (App. B.2 Fig. B.2).

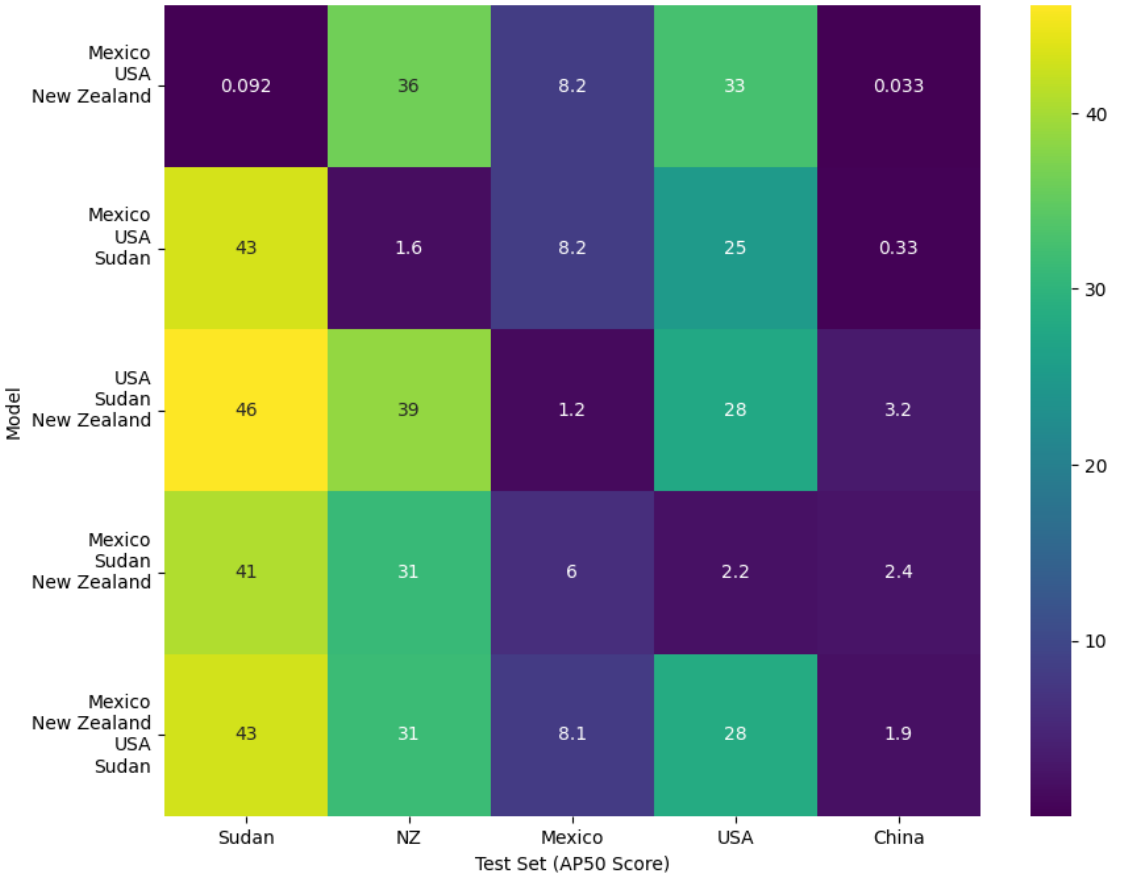


Figure 15: Model performance for the leave-one-out scheme. The models perform well in-sample (off-diagonal) and have low performance consistently on the country left out (diagonal). Mexico has a generally low performance.

4.2.3 Discussion

Individual Models. The models' performance is similar compared to the first model but drops on the out-of-sample test sets, in part due to the differences in training data heterogeneity in colour distributions.of the Mexican model are attributable to its test set.

In the Mexican case, the combination of regular training curves and the over 40-point AP50 score difference in validation and test score casts suspicion on the model selection. As in 4.1 the models are selected on the highest AP50 score of the validation set. These perform in line with the performance of the other models. Given the stratification in splitting the dataset, the test set is only taken from a single satellite image. This image differs significantly from the training and validation data, and thus the test performance is particularly poor.[8]

[8]To test this, we also switched the validation and test set and found that when selecting the model on

The models are validated every 500 iterations compared to only every 2,000 iterations in the models from 4.1. Thus, the AP50 scores on the validation set in Fig. 14 allow a closer look at the models' performance and a precise point at which overfitting sets in. The two models that perform higher on their respective in-sample sets (Sudan and New Zealand) also reach their peak earlier. The US model fits less tightly to the data but also overfits. This observation thus strengthens the analysis that the in-sample high-performers have more homogeneous underlying data.

The models in this second experiment have ceased to exhibit any usable performance on the out-sample sets. The only outlier from this trend is a Sudanese model tested on Chinese data. Given that the hyperparameter configuration has not changed and the other models test poorly on Sudanese and Chinese test data, this is related to the underlying data structure. Similarly, the inherent heterogeneity in US landscapes is partially represented in the dataset by including locations in Arizona (dry), Wyoming (green), Clyde/Texas (urban-green), and Kansas (urban-dry). This increases the sample space in which locations can be evaluated successfully and thus leads to New Zealand and Mexican models and test sets performing better in conjunction with the US data and models respectively.

There is thus evidence that each location's images form a multidimensional convex hull in which the respective testing images must fall to be performing well on. If the respective test images are within or overlapping this multidimensional space, the results improve. There are a variety of aspects that characterise the respective images, among them the build of distribution towers, the presence of further objects characterised by an urban or rural context, and the overall natural environment.

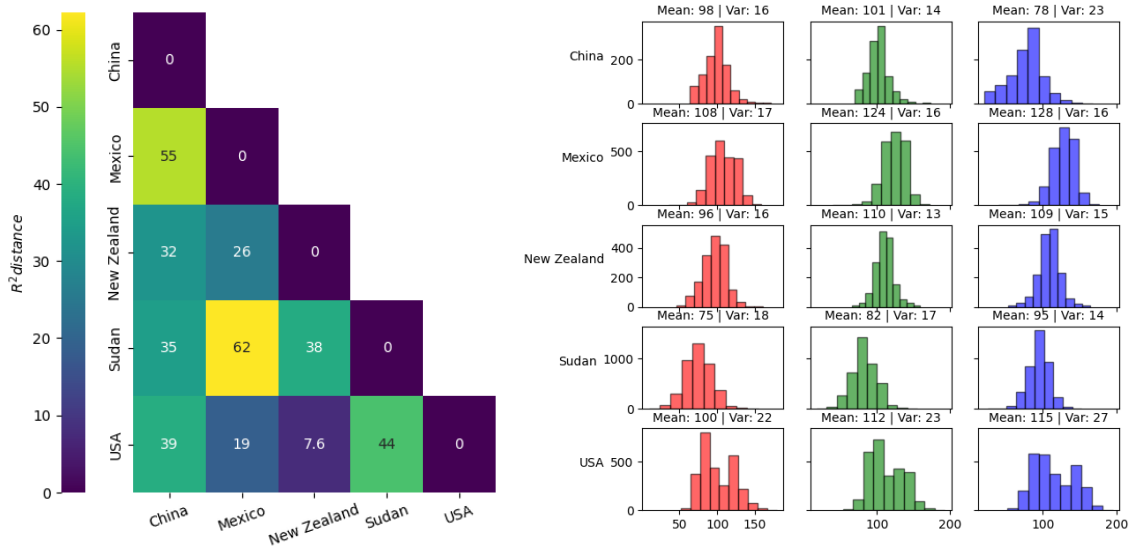
While the former two cannot be extracted directly from the data, the location of each location's hull on the dimension of natural environment can be approximated by the mean $R\vec{G}B$ vector of the location's image. The size of the hull can be characterised by it variance in $R\vec{G}B$ values:

$$R\vec{G}B_{loc}^{mean} = \frac{1}{n_{loc}} \sum^{n_{loc}} \frac{\sum^{512*512} (r, g, b)}{512 * 512} \tag{1}$$

the test set, the AP50 scores are consistently low. In contrast, the AP50 of the validation set - now the test set - are within the 40s.

$$R\vec{G}B_{loc}^{variance} = \frac{(\sum^{n_{loc}} R\vec{G}B_{loc}^{mean} - \frac{\sum^{512*512}(r,g,b)}{512*512})^2}{n_{loc}-1} \tag{2}$$

To understand the similarity of the images, Fig. 16 gives an overview of the R^2 distance between the average $R\vec{G}B_{loc}^{mean}$ of two locations (16a) and the respective variance of the colour channels (16b). The US stands out with the lowest distances from other locations and simultaneously the largest variances in the colour distributions. The variance in colours is able to explain why the US model performs better particularly on Mexico and New Zealand than vice versa. It is, however, also the reason why the US model fits its own data less as the space the model has to cover is larger. Sudan and New Zealand have lower relative variance and are hence able to fit their own data well.



(a) R^2 distances.

(b) Color distributions.

Figure 16: R^2 distance of the mean RGB vectors and by country RGB distributions. The USA (bottom-rows) has the widest distribution and smallest distances with the other countries.

The relationship between $R^2 distance$ in conjuncture with the training data variance is able to explain 45% of the variance in the $log(AP50)$ value. The R^2 distance is associated with the outcome value negatively and the average variance of the training data channels positively (Fig. 17 and Appendix B.1 Tab. B.1). The small sample size permits little generalizability of these results outside the inspected model space. Nevertheless, the data differ in measurable dimensions that negatively affect the out-of-sample performances. Hence, any - even if futile - attempt to use data from a country to estimate

the power infrastructure in another should aim to approximate the natural environment of their target. It should further allow the set to be diverse enough to overlap with the distributions of the test locations.

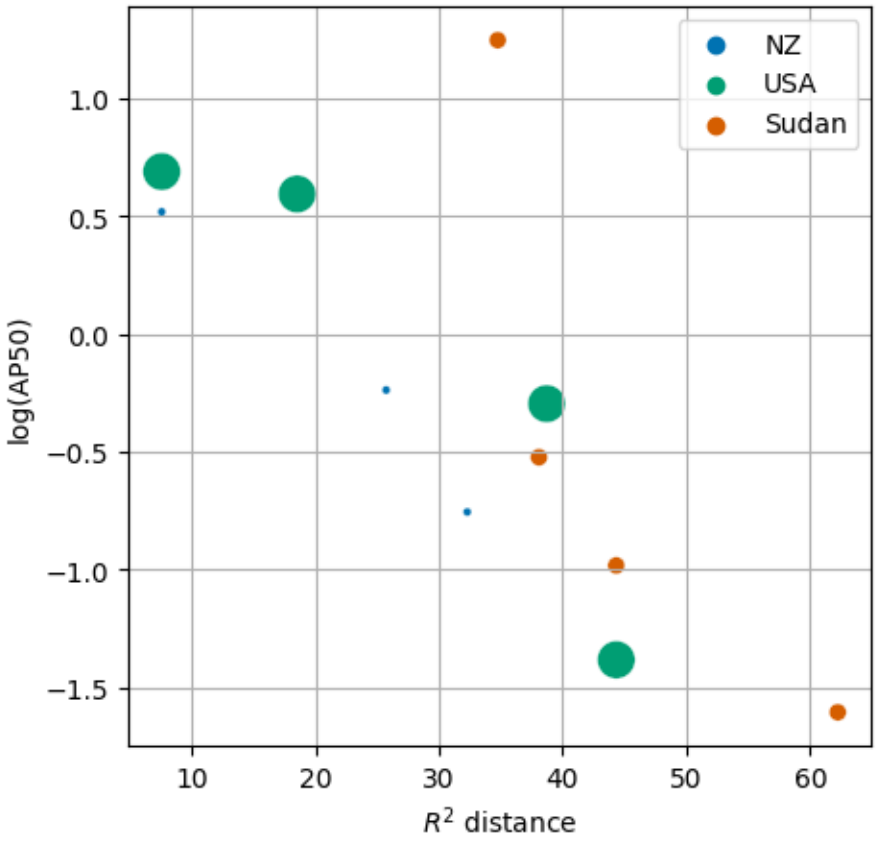


Figure 17: Correlation between R^2-distance between train-test $\vec{RGB}^{mean}$ and AP50 scores for out-of-sample testing. The respective marker size indicates the variance of the training data's $\vec{RGB}$. The plot does not show scores from the Mexican model.

Leave-One-Out. The LOO results are further evidence of the models' inabilities to generalise outside their training data. The effects of adding and removing individual countries from the combination are in line with the exploration of colour distributions above.

When left out of the training data, an unseen country's test set performance requires little elaboration. As found above, the models are unable to adequately generalise to out-of-sample data. The Mexican performance is still poor, but benefits from the longer training time and larger training data heterogeneity provided by the other locations.

Sudan and USA represent the two ends of the in-country variation. The marginal

improvements in their test scores given their interaction with the other locations, aids the understanding of how to best construct data sets when trying to optimise for a particular region of interest.

For example, Sudan has its highest performance together with USA and New Zealand. Adding Mexico into the mix worsens the results both for the three- and four-country model (right column Fig. 15). Fig. 16a highlights the stark average colour difference between Sudan and Mexico. Sudan's in-sample performance is the highest (43) and its out-of-sample performance (0.092) is the lowest, which can also be explained by the on average highest colour differences with the other countries and relatively low RGB variance. Thus, decision-makers should take the relative colour differences into account when composing a dataset, especially when the region of interest has a narrow distribution in biotopes.

The US on the other hand performs best with Mexico and NZ which have the lowest colour differences with it. Unlike the Sudanese model (47 vs. 46) it also outperforms the individual model (24 vs. 33). This can again be attributed to the fact that it has the highest variance. This also leads to the models that include the US data outperforming the instances when it is not included. Further, the US test set also has the highest AP50 score when being out-of-sample (2.2), which provides further evidence for the fact that the variance and colour similarities allow for the other datasets to be able to cover at least some parts of the test-set distribution.

Overall, if one were to compose a dataset out of multiple locations, one should consider the overall variance of the target location. If the location is characterised by one type of natural environment, a narrow approach of including location-specific data may be sufficient. With a greater variety in the location of interest, including more locations and a wider spread of areas may be beneficial.

4.3 Experiment 3 - The Effect of Tower Size

4.3.1 Set-Up

4.1 highlighted the challenges related to the heterogeneity in data on which to train models to detect tower infrastructure. In response, 4.2 explored a key source of variation:

geography. This section will deep dive into another source of data heterogeneity: tower size. This is an important exploration as the dataset used in 4.1 is unbalanced with regard to the tower size and, in addition, the bounding boxes of the towers still range between 36 and 17,880 pixels in size (Fig. 7).

Object size is one of the main factors influencing the complexity of the object detection task and therefore the detection capability of the model (Aguilar, Ortner, & Zerubia, 2022; Mansour, Hussein, & Said, 2019; Rabbi, Ray, Schubert, Chowdhury, & Chao, 2020). This correlation is intuitive as a smaller object size means that fewer pixels in the image are storing the information of the object. In this section, we first train and test multiple models for different tower sizes (Sub-Experiment 3.1 Fig. 18) and then compare the performance of these size-based models with the performance of the LGRS models ran in 4.1. Finally, we repeat the main experiment with datasets containing only large towers to specifically test the model performance for larger objects in lower resolution imagery (Sub-Experiment 3.2 Fig. 18).

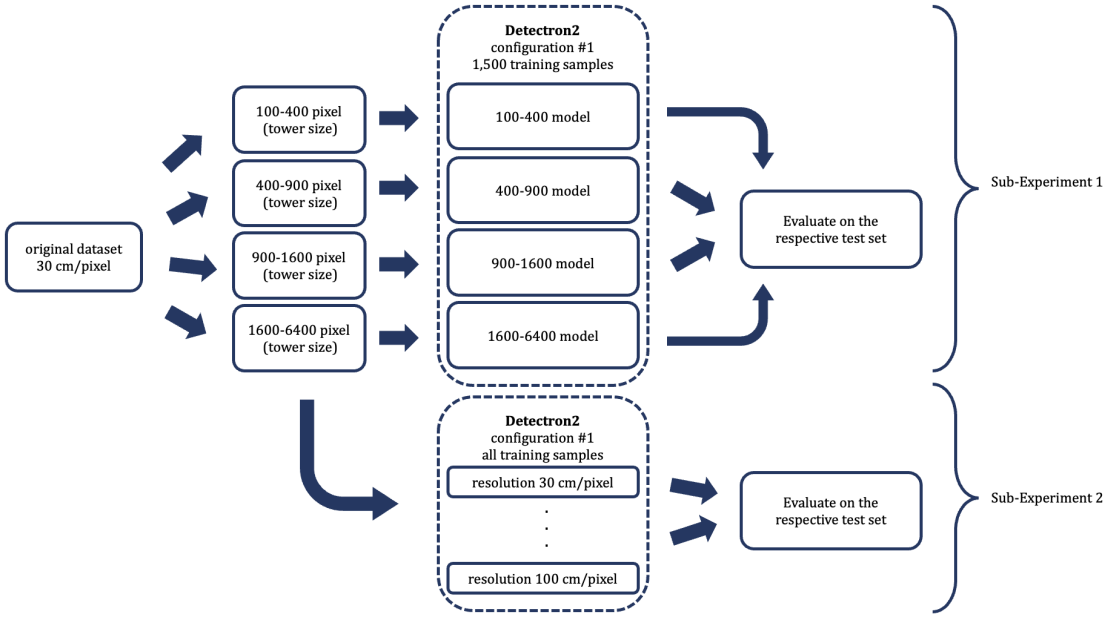


Figure 18: Workflow third experiment. Sub-experiment one compares model performance for different tower sizes, holding resolution constant. Sub-experiment two investigates the impact of resolution on detection capability for large towers.

To compare tower sizes, the dataset is stratified by the area of the ground truth boxes. We include all tower types (*distribution* and *transmission*) in the dataset and do not dis-

tinguish them in their labelling (both are labelled as *tower*). This is because, in contrast to the previous sections where the transmission towers presented a challenge given their larger average size, we are now interested in the detectability of differently sized towers and hence have good reason to include them in the dataset (Table 2).

Tower Type	Mean	Median	Mode	Minimum	Maximum	Std Dev
Distribution Tower	473	300	180	36	17,880	547
Transmission Tower	6,491	4,920	3,496	187	44,980	6,526

Table 2: Tower size distributions. Descriptive statistics of the distribution of ground truth bounding box sizes (in pixel) of distribution and transmission towers

To account for differently sized towers, we adjust the sizes of the anchor boxes in the model (same as in 4.1.1), which are characterised by their aspect ratio, corresponding to the square root of the size of the ground truth boxes. We specify four models with (1) tower size 100-400 pixels (extra small), (2) tower size 400-900 pixels (small), (3) tower size 900-1600 pixels (medium) and (4) tower size 1600-6400 pixels (large). [9][10] Due to the imbalance of train/validation/test split for some locations (Appendix C Fig. C.1) and to reduce variation caused by location (see RGB differences in 4.2.3 Fig. 16a), we limit the experiments to the locations in the US and New Zealand and therefore have to restrict each training sample to a random selection of 1000 images to ensure comparability.[11]

There are oftentimes multiple towers in a single image (Fig. 3). To adjust for potential variability of tower size in a single image, images are categorized based on the size of their primary tower the image is cut around (see 3.1). Labels of additional towers in the image are deleted if they do not belong to the same 'size category' as the main tower.

Similar to the set-up in 4.2 we reduce training iterations to 10,000 due to reduced sample size, increase the validation frequency and train the models with configuration #1 (Tab. 1) on data with a resolution of 30 cm/pixel.

[9] Due to the imbalance of the dataset with regard to tower size, the last bin has to include the wide range of tower sizes from 1600-6400 pixels (see Appendix C Fig. C.2).

[10] tower size 100-400 corresponds to an anchor box width/height of 10-20, tower size 400-900 corresponds to an anchor box width/height of 20-30, tower size 900-1600 corresponds to an anchor box width/height of 30-40 and tower size 1600-6400 corresponds to an anchor box width/height of 40-80

[11] the critical reader will find that we criticized Huang et al. (2021) before for the exact same choice of locations, however this choice was necessary due to given reasons.

4.3.2 Results

Tower Size Models. The experiment's results confirm the hypothesis formulated above as the model specified for the largest towers performs best (AP50: 34.0) and the model for the smallest towers performs worst (AP50: 17.5). Providing nuance to this observation, and also countering expectations, the model specified for towers of size 400-900 performs better (AP50: 24.6) than the model for towers of size 900-1600 (AP50: 29.2) (Fig. 19). Focusing on the LGRS model, we observe that it fits the skewed distribution of towers in its training data (Fig. 7), performing best for middle-sized towers. Yet it performs worse than the size specified models for both the small towers (AP50: 9.0 <17.5) and the large towers (AP50: 23.3 <34.0).[12]

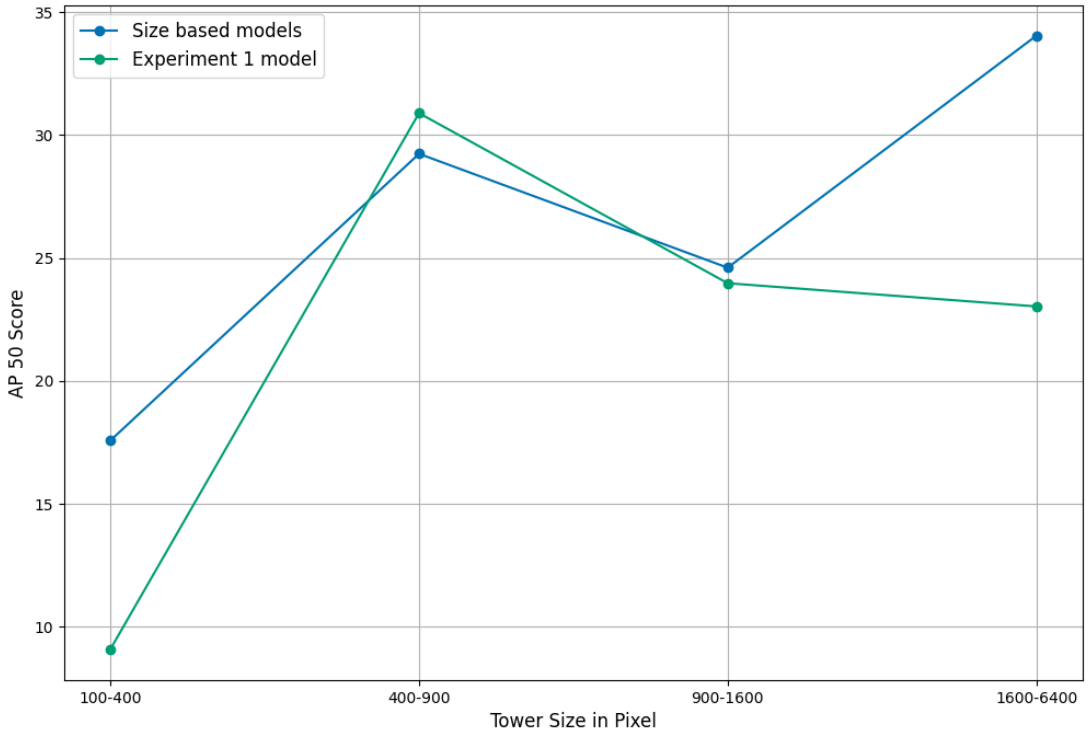


Figure 19: Model Performance on the test set for different tower sizes. Note: The size-based models are multiple models each specifically trained for detecting a specific tower size, whereas the Experiment 1 model is only a single model tested on the individual tower sizes.

[12] the performance of the general model is likely underestimated, as the model learned to detect all kinds of tower sizes, therefore detecting also the deleted tower labels which fall out of the respective size category, for which it gets penalized in this test setup.

The training behaviour of the models is different across tower sizes (Appendix C, Fig. C.3). The loss stays longer on the initial plateau for smaller tower sizes before it starts to decrease and then decreases less pronounced than for the models specified for larger towers. The loss curves for the models with tower sizes 100-400 and 400-900 pixels (extra small and small towers) indicate very similar training behaviour. Therefore the low test result for the model for extra small towers cannot be attributed to the training process. Besides these differences, learning seems to stagnate for all models around the 8,000th iteration. Similar to the prior experiments we do not see a decrease in the validation loss but a constant increase with temporarily somewhat constant losses around the time when the total loss starts to decrease. Again, the loss curves in combination do not show a clear sign of overfitting.

During training, the validation AP50 scores look similar to previous experiments (Appendix C, Fig. C.4). The AP50 scores peak between iteration 3,000-5,000, indicating that models overfit quite early in the training process. In contrast to the test results, and this time in line with expectations, the model for medium size towers performs better than the model for small tower sizes during training. Additionally, we find that the model for extra small towers outperforms its best validation score while testing (AP50 test: 17.5, AP50 validation: 13.5).

Resolution Models. In line with our expectation, we again observe a general trend of decreasing performance with decreasing resolution, but overall higher AP50 scores for the model specified to large towers (Fig. 20). These results are especially promising, as the variance of tower size is substantial (1600-6400 pixels). At a resolution of 30 cm/pixel, almost every second prediction is correct (AP50: 44.8). We find unstable results for the small resolution steps until 50 cm/pixel, an effect more pronounced than in 4.1. The performance drops to around 20 for both 70 cm/pixel and 100 cm/pixel resolutions and down to 5 for a resolution of 300 cm/pixel. Training and validation loss curves (Appendix C, Fig. C.6) indicate a very similar training behaviour for all resolutions up to 50 cm/pixel, with later peaks the lower the resolutions (70 cm/pixel and below). In contrast to the test results, during training, the AP50 scores show similar results for all resolutions with especially high scores (compared to their respective test results) for the low resolutions (Appendix C, Fig. C.6).

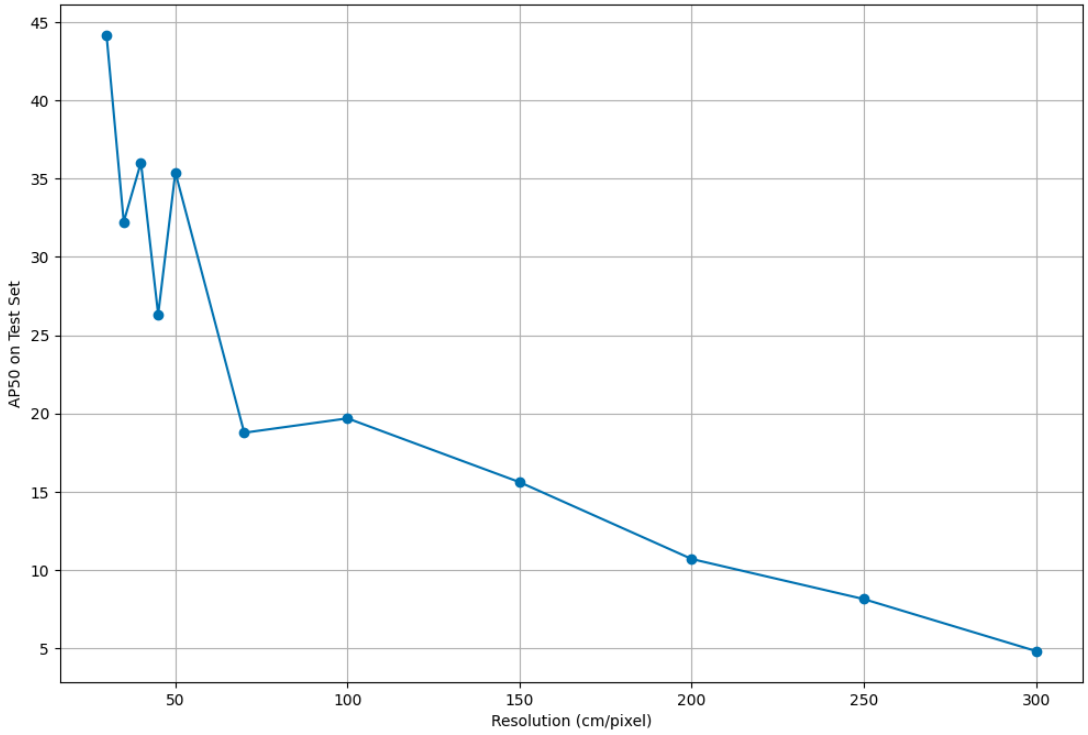


Figure 20: Model performance (model trained for tower size 1600-6400) on the test set for different resolutions.

4.3.3 Discussion

Tower Size Models. The experiment's results confirm the hypothesis that the model performs better for larger tower sizes compared to smaller towers, when resolution remains constant. Further, the results yield evidence that training models stratified by tower size can increase detection performance when the overall training dataset is skewed with regard to tower size.

The low performance of the LGRS model on the extra small towers is particularly striking, especially given that they are over-represented in the general training set. A potential explanation for this is that in the training data for the general model the small towers occur mostly in the data from Sudan (Appendix C, Fig. C.1b), thus fitting to this location, which is then not part of the stratified test set of this experiment.

The general difference in the development of the loss curves between smaller and bigger towers is most likely due to the fact that it is more difficult for the model to learn to recognise smaller towers than larger towers. This is especially evident in the training

curve for the 100-400 pixel model which remains rather flat, implying that the model is not picking up the signal of the towers. A lot of noise (other objects etc.) in the images could further complicate the task. Another source of poor training could be that images in the validation set are not representative, the model outperforming itself on the test set is an indication of this scenario.

Comparing the model for small towers (400-900 pixels) to the model for medium sizes towers (900-1600 pixels), the former performs better on the test set but slightly worse during training. This unexpected result might be attributable to the differences in validation set size. The model for tower sizes 400-900 has a validation set double the size than the model for sizes 900-1600 (612 images compared to 315) which potentially leads the model for sizes 400-900 to generalize better, and the model for sizes 900-1600 to overfit to the validation set. Overall, the assignment to distinct size categories could be blurred by annotators drawing bounding boxes around towers with varying degrees of tightness. In addition, the varying altitudes from which base images are taken determine the apparent size of the towers, so that the categorization may not correspond to the true size of the towers.

Resolution Model. A comparison of the results of the models stratified by tower size indicates that models generally perform better when specified for large towers compared to smaller ones. This effect is even more pronounced for low resolution imagery (<1 m/pixel). However, a direct comparison to the results of 4.1 should be taken with utmost caution, given that the datasets stratified by tower sizes are limited to more homogeneous locations and are therefore of significantly smaller size.

Even though the test results follow the same trend as in experiment 4.1, the decreasing trend is not as clear, especially for the small resolution steps between 35 and 50 cm/pixel and for resolutions 70 cm/pixel and 100 cm/pixel. A possible cause for these results could be that the hyperparameter tuning of the model was only done for a resolution of 30 cm/pixel and additionally for a model that was mainly trained for smaller tower sizes. Another limitation of model training is the small validation set (Appendix C, Fig. C.6). The small set size, containing only 88 images, contributes to the variance of the AP50 scores on the validation set, which could lead, in combination with limited val-

idation iterations, to a sub-optimal selection of the model used for the final performance testing. Overall, the loss curves and test results indicate that with further fine-tuning of the model for the task at hand, an even better and more robust performance could be reached.

Despite the limitations described, the experiment's results indicate, that the use of separate models for different tower sizes present a useful strategy to increase detection capacity. Further, if only low-resolution imagery is available, a focus on large towers could be beneficial. Future research should focus on model sensitivity to tower size and variance to determine whether the relationship between detection capability and tower size is linear or whether there exists some kind of threshold at which model performance drops significantly.

5 Conclusion

Extending existing research on electricity tower detection we demonstrate that even small energy infrastructure, which is more common in developing countries, can be automatically detected on high-resolution satellite imagery. With controlled experiments testing the effect of image resolution, we find that electricity tower detection on a large scale is only possible with commercial satellite imagery as the performance of our models has the steepest decline between 50 cm/pixel and 70 cm/pixel.

Tower detection based on satellite imagery remains a complex task. With additional experiments we demonstrate that the detection performance of the model is highly sensitive to tower size, locational characteristics such as ground colour, and the distribution of these features in the training data used. All of this strongly indicates that the task of tower detection cannot be reliably solved with a general model. The diversity of tower locations and sizes in practice suggests that out-of-sample training is not a promising approach in cases where training data from the location of interest is not available. However, we find that restricting the model to one tower size and to locations that are similar in their natural environment can lead to better detection performance.

Therefore, in order to create a dataset on which to train a model for a particular region, we suggest to first identify regions with similar characteristics, such as topogra-

phy, geographical environment, and energy infrastructure. Policy makers and modellers from similar regions could share high-resolution data and labelling to be utilised for the continuous development of more refined models while saving resources that can be invested in generally obtaining the highest-resolution images possible.

The question of a threshold for the minimum needed resolution of satellite imagery for tower detection cannot be definitively answered by our research. On the one hand, our results suggest that a resolution threshold will differ depending on the size of electricity towers and the level of noise in the respective context. Large towers on grasslands will have a lower threshold than small towers in similar environments and small towers in urban contexts will need even higher resolutions. On the other hand, our research on tower detection should be expanded with experiments on grid detection. By predicting grid lines using path detection models, the inherent nature of electricity grids can be utilised to filter out false positives and minimise the effect of non-detection. The ability of such an algorithm to draw a correct grid based on the detection predictions of the detection model could be a valid and practice-oriented metric to define a resolution threshold for the detection model.

Our findings offer a range of impulses for further research. Given the difficulty in obtaining affordable satellite imagery for electricity tower detection, further research could examine whether super-resolution methods could be leveraged to increase detection performance on lower-resolution imagery. Another crucial aspect of future work could be the development of the underlying dataset. With a more geographically and regionally diverse and larger dataset, the research on the impact of certain characteristics (like location and tower size) could be extended by including additional experimental variables and by providing an application-oriented benchmark for future algorithm development.

6 References

Aguilar, C., Ortner, M., & Zerubia, J. (2022). Small object detection and tracking in satellite videos with motion informed-CNN and GM-PHD filter. *Frontiers in Signal Processing*, *2*, 827160. Retrieved 2023-04-19, from https://www.frontiersin.org/articles/10.3389/frsip.2022.827160/full doi: 10.3389/frsip.2022.827160

Arderne, C., Zorn, C., Nicolas, C., & Koks, E. E. (2020, January). Predictive mapping of the global power system using open data. *Scientific Data*, *7*(1), 19. Retrieved 2022-12-12, from https://www.nature.com/articles/s41597-019-0347-4 doi: 10.1038/s41597-019-0347-4

Azimi, S. M., Vig, E., Bahmanyar, R., Körner, M., & Reinartz, P. (2018). Towards multi-class object detection in unconstrained remote sensing imagery. Retrieved 2023-04-26, from https://arxiv.org/abs/1807.02700 (Publisher: arXiv Version Number: 3) doi: 10.48550/ARXIV.1807.02700

Brown, J., Clark, C., Lomax, S., Rafique, K., & Sukkarieh, S. (2022, March). *Manipulating UAV Imagery for Satellite Model Training, Calibration and Testing.* arXiv. Retrieved 2023-03-10, from http://arxiv.org/abs/2203.11447 (Number: arXiv:2203.11447 arXiv:2203.11447 [cs, eess])

Brown, J., Qiao, Y., Clark, C., Lomax, S., Rafique, K., & Sukkarieh, S. (2022, February). Automated Aerial Animal Detection When Spatial Resolution Conditions Are Varied. *Computers and Electronics in Agriculture*, *193*, 106689. Retrieved 2023-03-10, from http://arxiv.org/abs/2110.01329 (arXiv:2110.01329 [cs, eess]) doi: 10.1016/j.compag.2022.106689

Castello, R., Roquette, S., Esguerra, M., Guerra, A., & Scartezzini, J.-L. (2019, November). Deep learning in the built environment: automatic detection of rooftop solar panels using Convolutional Neural Networks. *Journal of Physics: Conference Series*, *1343*(1), 012034. Retrieved 2023-04-20, from https://dx.doi.org/10.1088/1742-6596/1343/1/012034 (Publisher: IOP Publishing) doi: 10.1088/1742-6596/1343/1/012034

Chen, B., & Miao, X. (2020). Distribution line pole detection and counting based on

YOLO using UAV inspection line video. *Journal of Electrical Engineering & Technology*, *15*(1), 441–448. Retrieved 2023-04-28, from http://link.springer.com/10.1007/s42835-019-00230-w doi: 10.1007/s42835-019-00230-w

Ciocarlan, A., & Stoian, A. (2021, January). Ship Detection in Sentinel 2 Multi-Spectral Images with Self-Supervised Learning. *Remote Sensing*, *13*(21), 4255. Retrieved 2022-12-15, from https://www.mdpi.com/2072-4292/13/21/4255 (Number: 21 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/rs13214255

Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M. Y., ... Zhang, L. (2022, November). Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7778–7796. Retrieved 2022-12-15, from http://arxiv.org/abs/2102.12219 (arXiv:2102.12219 [cs]) doi: 10.1109/TPAMI.2021.3117983

Girshick, R. (2015, September). *Fast R-CNN*. arXiv. Retrieved 2023-04-14, from http://arxiv.org/abs/1504.08083 (arXiv:1504.08083 [cs])

Golovko, V., Bezobrazov, S., Kroshchanka, A., Sachenko, A., Komar, M., & Karachka, A. (2017). Convolutional neural network based solar photovoltaic panel detection in satellite photos. In *2017 9th IEEE international conference on intelligent data acquisition and advanced computing systems: Technology and applications (IDAACS)* (pp. 14–19). IEEE. Retrieved 2023-04-20, from http://ieeexplore.ieee.org/document/8094501/ doi: 10.1109/IDAACS.2017.8094501

Han, B., & Wang, X. (2017, April). Learning for Tower Detection of Power Line Inspection. *DEStech Transactions on Computer Science and Engineering*(iccae). Retrieved 2022-12-15, from http://dpi-journals.com/index.php/dtcse/article/view/7194 doi: 10.12783/dtcse/iccae2016/7194

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018, January). *Mask R-CNN*. arXiv. Retrieved 2023-04-08, from http://arxiv.org/abs/1703.06870 (arXiv:1703.06870 [cs] version: 3)

Hu, Z., He, T., Zeng, Y., Luo, X., Wang, J., Huang, S., ... Lin, B. (2018, December). Fast image recognition of transmission tower based on big data. *Protection and Control of Modern Power Systems*, *3*(1), 15. Retrieved 2022-12-15, from https://

pcmp.springeropen.com/articles/10.1186/s41601-018-0088-y doi: 10.1186/s41601-018-0088-y

Huang, B., Yang, J., Streltsov, A., Bradbury, K., Collins, L. M., & Malof, J. (2021). GridTracer: Automatic Mapping of Power Grids using Deep Learning and Overhead Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. Retrieved 2022-12-12, from https://arxiv.org/abs/2101.06390 doi: 10.48550/ARXIV.2101.06390

Kaack, L. H., Chen, G. H., & Morgan, M. G. (2019, July). Truck traffic monitoring with satellite images. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 155–164). Accra Ghana: ACM. Retrieved 2023-04-20, from https://dl.acm.org/doi/10.1145/3314344.3332480 doi: 10.1145/3314344.3332480

Kruitwagen, L., Story, K. T., Friedrich, J., Byers, L., Skillman, S., & Hepburn, C. (2021, October 28). A global inventory of photovoltaic solar energy generating units. *Nature*, *598*(7882), 604–610. Retrieved 2023-04-24, from https://www.nature.com/articles/s41586-021-03957-7 doi: 10.1038/s41586-021-03957-7

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., ... Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, *abs/1405.0312*. Retrieved from http://arxiv.org/abs/1405.0312

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017, April). *Feature Pyramid Networks for Object Detection*. arXiv. Retrieved 2023-04-08, from http://arxiv.org/abs/1612.03144 (arXiv:1612.03144 [cs])

Malof, J. M., Collins, L. M., Bradbury, K., & Newell, R. G. (2016). A deep convolutional neural network and a random forest classifier for solar photovoltaic array detection in aerial imagery. In *2016 IEEE international conference on renewable energy research and applications (ICRERA)* (pp. 650–654). IEEE. Retrieved 2023-04-20, from http://ieeexplore.ieee.org/document/7884415/ doi: 10.1109/ICRERA.2016.7884415

Malof, J. M., Rui Hou, Collins, L. M., Bradbury, K., & Newell, R. (2015). Automatic solar photovoltaic panel detection in satellite imagery. In *2015 international conference on renewable energy research and applications (ICRERA)* (pp. 1428–1431). IEEE. Retrieved

2023-04-20, from http://ieeexplore.ieee.org/document/7418643/ doi: 10.1109/ICRERA.2015.7418643

Mansour, A., Hussein, W. M., & Said, E. (2019). Small objects detection in satellite images using deep learning. In *2019 ninth international conference on intelligent computing and information systems (ICICIS)* (pp. 86–91). IEEE. Retrieved 2023-04-19, from https://ieeexplore.ieee.org/document/9014842/ doi: 10.1109/ICICIS46948.2019.9014842

Mapping the electric grid. (2018). Retrieved 2023-04-21, from https://devseed.com/ml-grid-docs/

Medjroubi, W., Müller, U. P., Scharf, M., Matke, C., & Kleinhans, D. (2017, November). Open Data in Power Grid Modelling: New Approaches Towards Transparent Grid Models. *Energy Reports*, *3*, 14–21. Retrieved 2023-04-20, from https://linkinghub.elsevier.com/retrieve/pii/S2352484716300877 doi: 10.1016/j.egyr.2016.12.001

Mitchell, D. P., & Netravali, A. N. (1988). Reconstruction filters in computer-graphics. *ACM SIGGRAPH Computer Graphics*, *22*(4), 221–228. Retrieved 2023-04-24, from https://dl.acm.org/doi/10.1145/378456.378514 doi: 10.1145/378456.378514

Padilla, R., Netto, S. L., & da Silva, E. A. B. (2020). A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (iwssip)* (p. 237-242). doi: 10.1109/IWSSIP48289.2020.9145130

Rabbi, J., Ray, N., Schubert, M., Chowdhury, S., & Chao, D. (2020). Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network. *Remote Sensing*, *12*(9), 1432. Retrieved 2023-04-19, from https://www.mdpi.com/2072-4292/12/9/1432 doi: 10.3390/rs12091432

Ren, S., He, K., Girshick, R., & Sun, J. (2016, January). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* arXiv. Retrieved 2023-04-06, from http://arxiv.org/abs/1506.01497 (arXiv:1506.01497 [cs])

Ren, S., Hu, W., Bradbury, K., Harrison-Atlas, D., Malaguzzi Valeri, L., Murray, B., & Malof, J. M. (2022, November). Automated Extraction of Energy Systems Information from Remotely Sensed Data: A Review and Analysis. *Applied Energy*, *326*,

119876. Retrieved 2022-12-12, from https://linkinghub.elsevier.com/retrieve/pii/S0306261922011424 doi: 10.1016/j.apenergy.2022.119876

Ren, S., Malof, J., Fetter, R., Beach, R., Rineer, J., & Bradbury, K. (2022, March). Utilizing Geospatial Data for Assessing Energy Security: Mapping Small Solar Home Systems Using Unmanned Aerial Vehicles and Deep Learning. *ISPRS International Journal of Geo-Information*, *11*(4), 222. Retrieved 2023-03-16, from https://www.mdpi.com/2220-9964/11/4/222 doi: 10.3390/ijgi11040222

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019, June). Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 658–666). Long Beach, CA, USA: IEEE. Retrieved 2023-04-15, from https://ieeexplore.ieee.org/document/8953982/ doi: 10.1109/CVPR.2019.00075

United Nations. (2018, June). *The Sustainable Development Goals Report 2018.* Retrieved 2023-04-27, from http://desapublications.un.org/publications/sustainable-development-goals-report-2018

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... Yu, T. (2014). scikit-image: image processing in python. *PeerJ*, *2*, e453.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). *Detectron2.* https://github.com/facebookresearch/detectron2.

Yang, X., Yang, J., Yan, J., Zhang, Y., Zhang, T., Guo, Z., ... Fu, K. (2018). SCRDet: Towards more robust detection for small, cluttered and rotated objects. Retrieved 2023-04-26, from https://arxiv.org/abs/1811.07126 (Publisher: arXiv Version Number: 4) doi: 10.48550/ARXIV.1811.07126

A Experiment 1

A.1 Hyperparameter Tuning

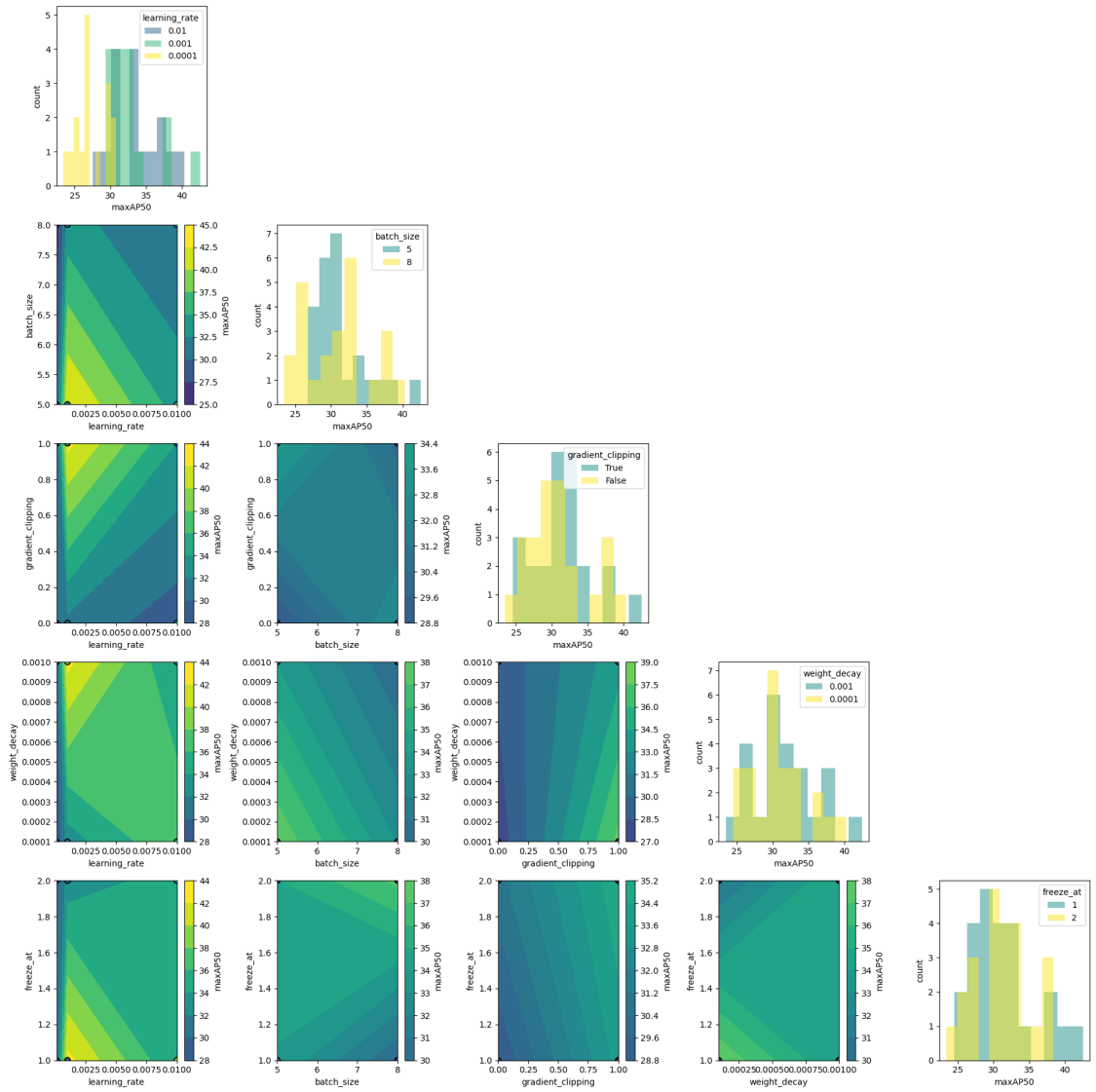


Figure A.1: Results hyperparameter tuning. Contour plots showing the change in AP50 given the change of two parameters. On the diagonal, the AP50 scores are grouped by values of hyperparameters.

A.2 GridTracer Replication

To replicate the results of Huang et al. (2021) we selected the same model type (Resnet101) the same subset of the data and the same hyper-parameters and model settings, as far as they were discussed in their paper, to achieve the most similar experiment set-up as possible. Huang et al. report an average AP50 score of 0.52 for their experiment. In our

replication experiment, in which we trained a model on all locations and retain an average AP50 score for all three locations, we only achieve an AP50 score of 0.42. There are several potential explanations for this performance gap: (1) unfortunately, due to different explanations of their data handling scheme it remains unclear whether Huang et al. trained three separate models for each region or one model for all three regions, which would influence the task complexity (2) the pre-processing of the data, including the way images are cut into sub-images and downsampling, could have let to significant differences, as the choices made for these steps are not described in the paper of Huang et al.. (3) we assume further differences in the experiment set up, as they mention a training duration of 50,000 iterations and we observed a peak in learning already around iteration 15,000.

A.3 Downsampling for 3 m/pixel and 10 m/pixel

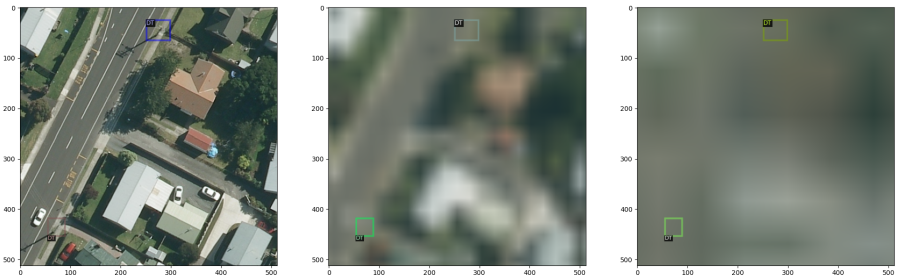


Figure A.1: Example images for resultions 3 m/pixel and 10 m/pixel. Downsampled images for resolutions 3 m/pixel and 10 m/pixel with the ground-truth bounding boxes. The test performance for either resolution are close to 0 AP50.

A.4 Results on Full Test Set

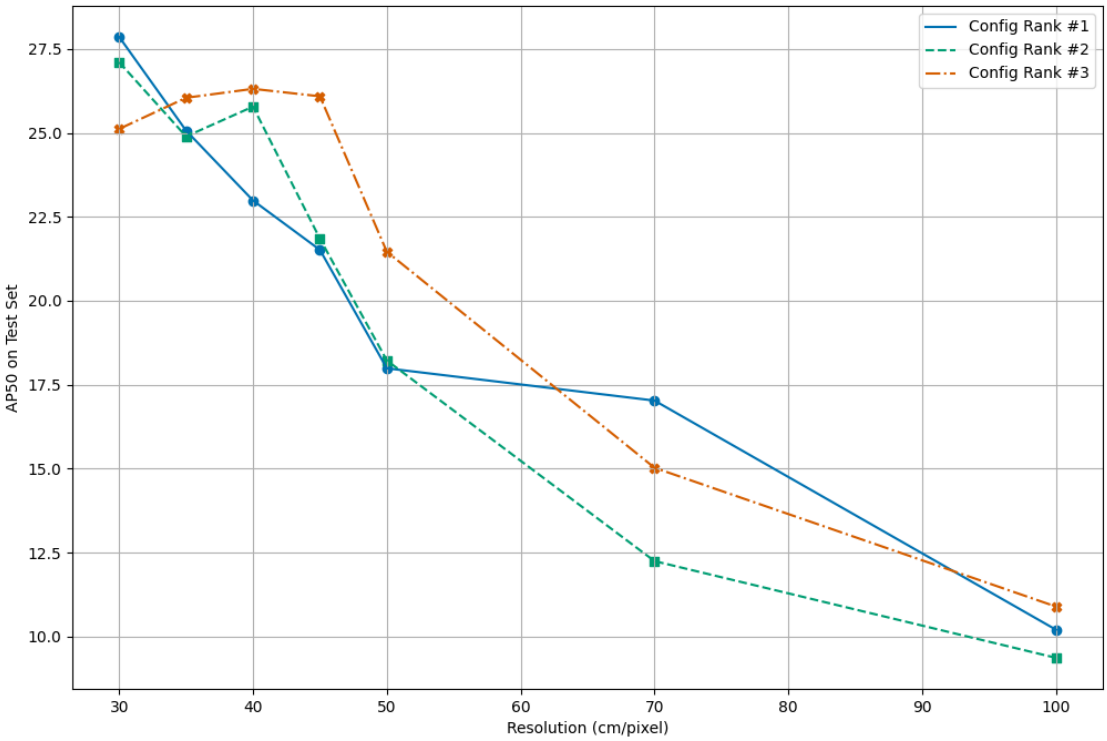


Figure A.2: AP50 during validation. AP50 Scores for all three configurations on the test set for decreasing resolutions for the full test set including Mexico.

B Experiment 2

B.1 Regression Outputs

Dep. Variable:	log_AP50	**R-squared:**	0.505
Model:	OLS	**Adj. R-squared:**	0.455
Method:	Least Squares	**F-statistic:**	10.19
Date:	Mon, 24 Apr 2023	**Prob (F-statistic):**	0.00962
Time:	10:43:16	**Log-Likelihood:**	-10.682
No. Observations:	12	**AIC:**	25.36
Df Residuals:	10	**BIC:**	26.33
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P$>$ |t|**	**[0.025**	**0.975]**
r2	-0.0339	0.010	-3.249	0.009	-0.057	-0.011
var	0.0448	0.020	2.220	0.051	-0.000	0.090

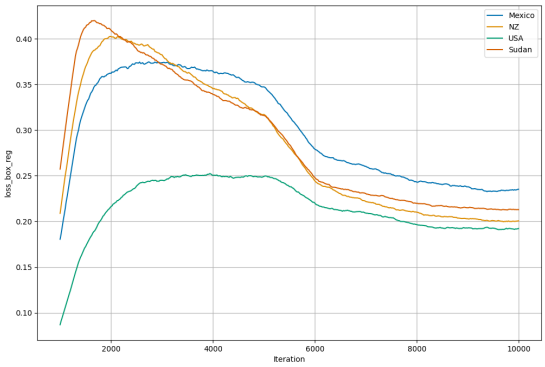
Omnibus:	13.886	**Durbin-Watson:**	1.972
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	8.983
Skew:	1.467	**Prob(JB):**	0.0112
Kurtosis:	6.058	**Cond. No.**	4.73

Table B.1: . OLS Regression Results. Regression of log(AP50) against the $R^2R\vec{G}B$ distance (**r2**) and the variance of the training country's data (**var**) Note: Standard Errors assume that the covariance matrix of the errors is correctly specified.

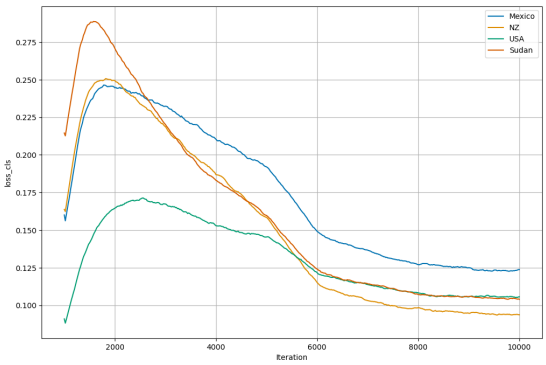
Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

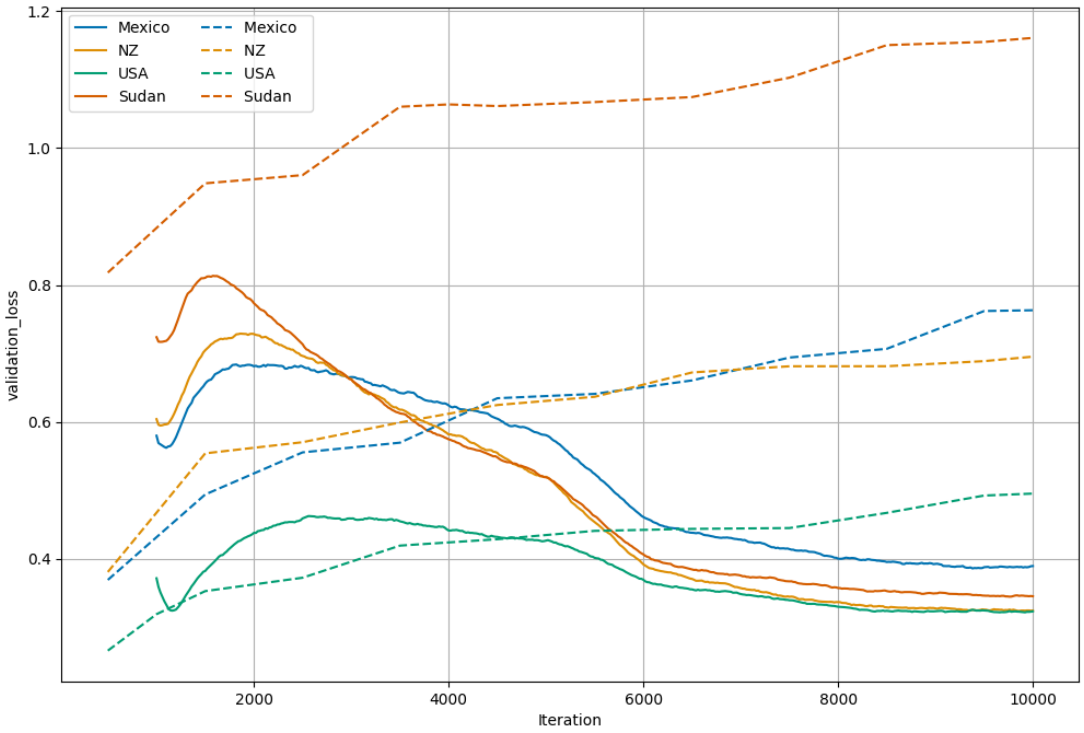
B.2 Figures for Experiment 2 - Impact of Biotopes



(a) Box Regression Loss.

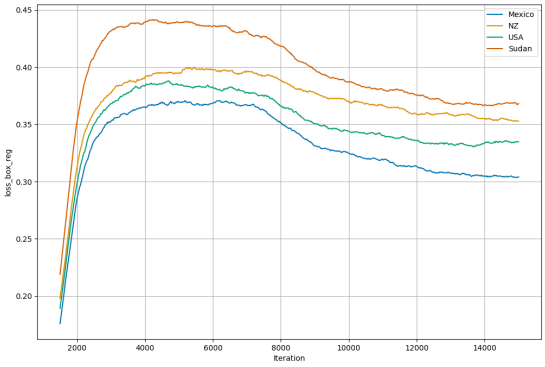


(b) Classification Loss.

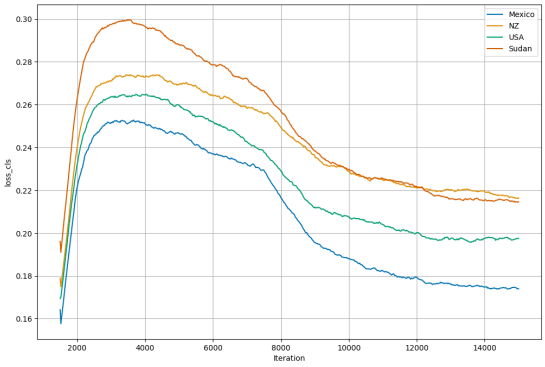


(c) Total (solid) and Validation loss (dashed).

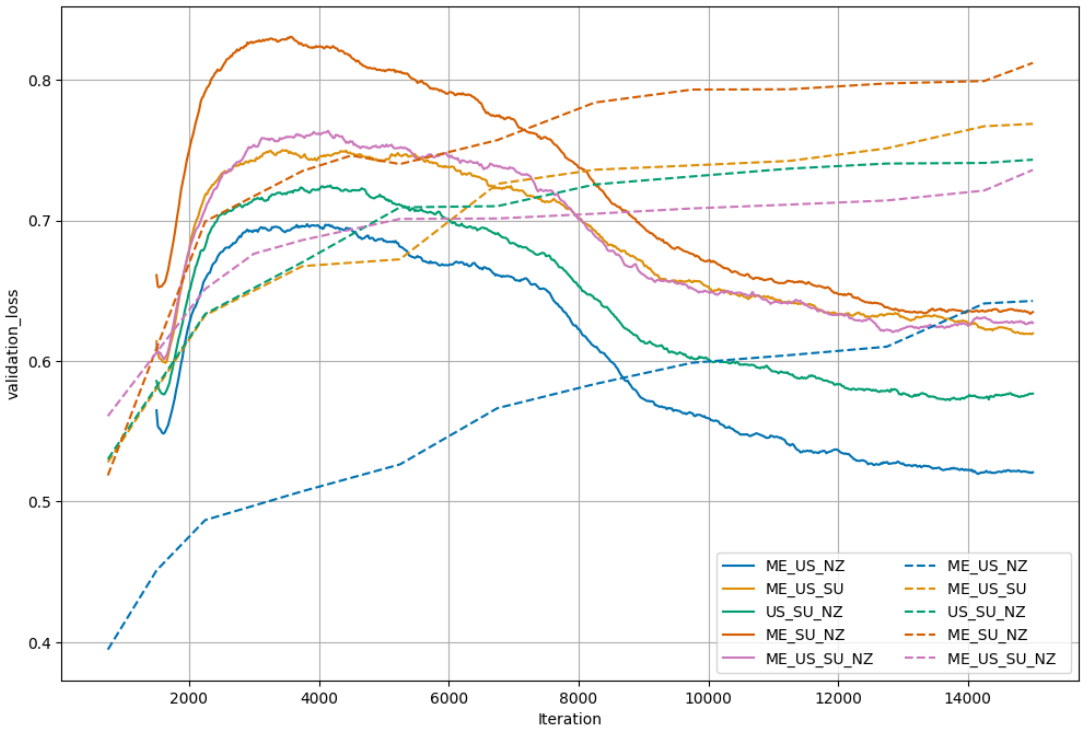
Figure B.1: Training loss curves. The box regression and classification losses decrease across resolutions (configuration #1). Consequently, the total loss also decreases. The validation loss (measured every 500 iterations) increases throughout. For better readability, the training's loss curves are averaged over a sliding window of size $\frac{1}{10}$ of the training iterations.



(a) Box Regression Loss.



(b) Classification Loss.



(c) Total (solid) and Validation loss (dashed).

Figure B.2: Training loss curves. The box regression and classification losses decrease *LOO* configurations (configuration #1). Consequently, the total loss also decreases. The validation loss (measured every 750 iterations) increases throughout. For better readability, the training's loss curves are averaged over a sliding window of size $\frac{1}{10}$ of the training iterations.

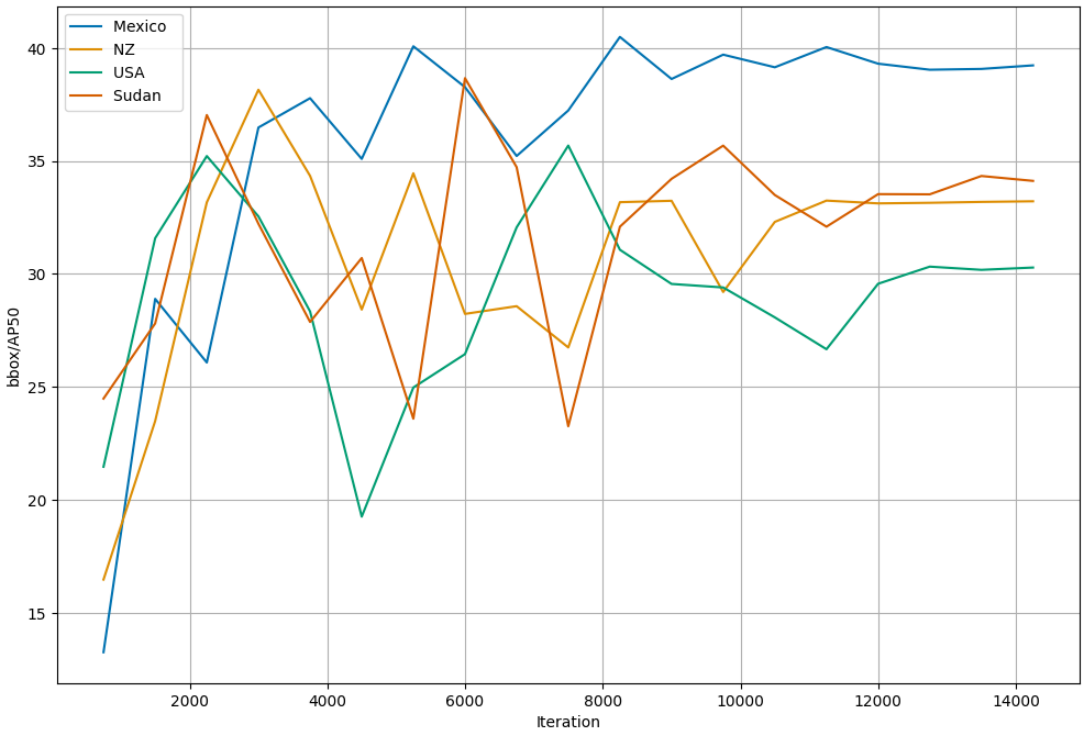
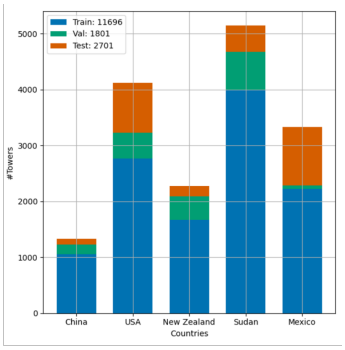
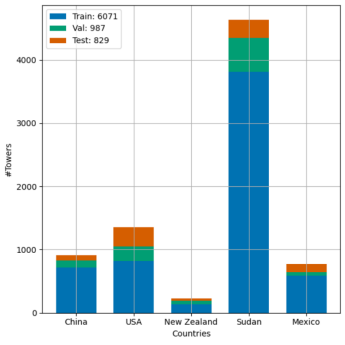


Figure B.3: AP50 during validation. AP50 score on validation set over the training iterations for models specified on LOO datasets.

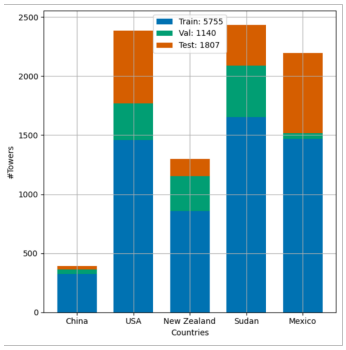
C Experiment 3



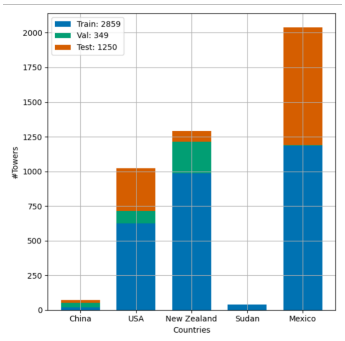
(a) All Tower Sizes.



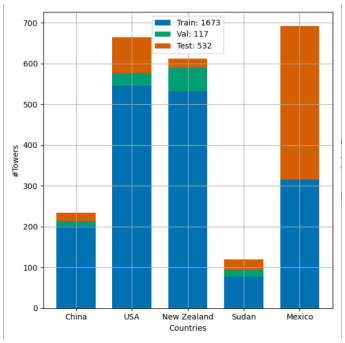
(b) Tower Size 100-400.



(c) Tower Size 400-900.



(d) Tower Size 900-1600.



(e) Tower Size 1600-6400.

Figure C.1: Train/val/test split by tower size and location. We only have representative data over all tower sizes for locations in the USA and New Zealand.

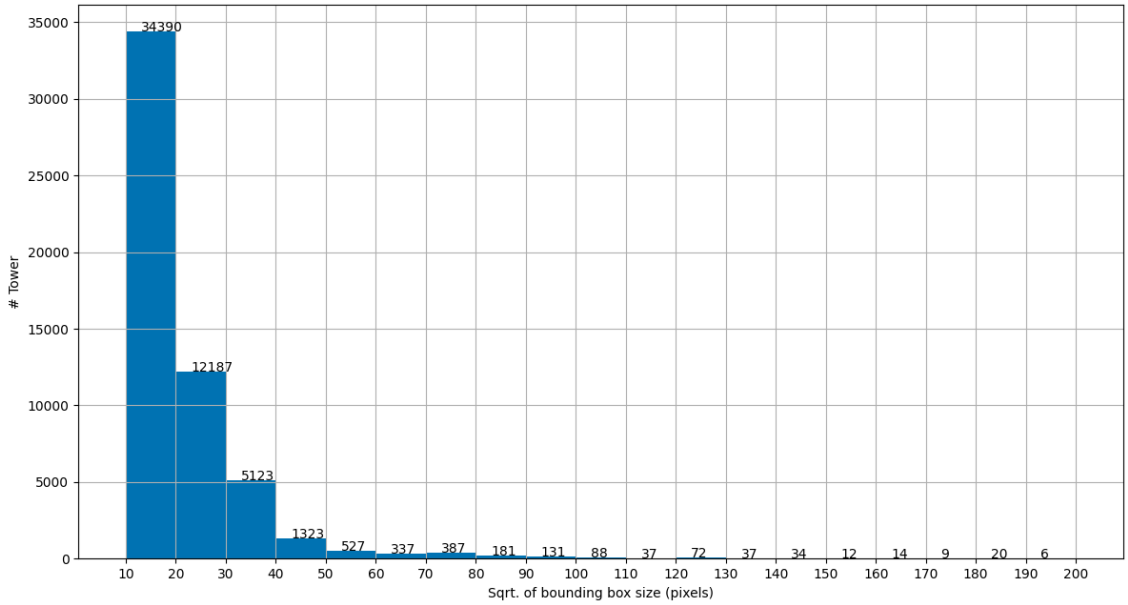
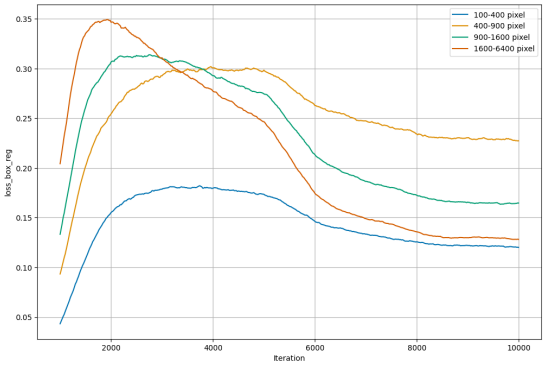
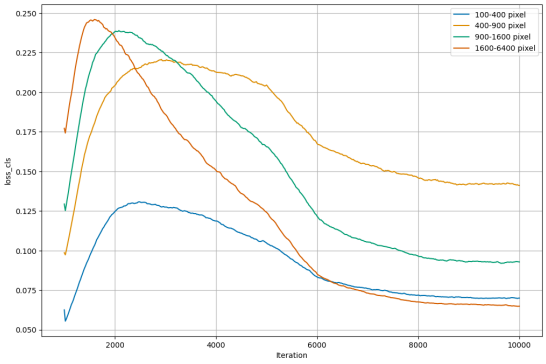


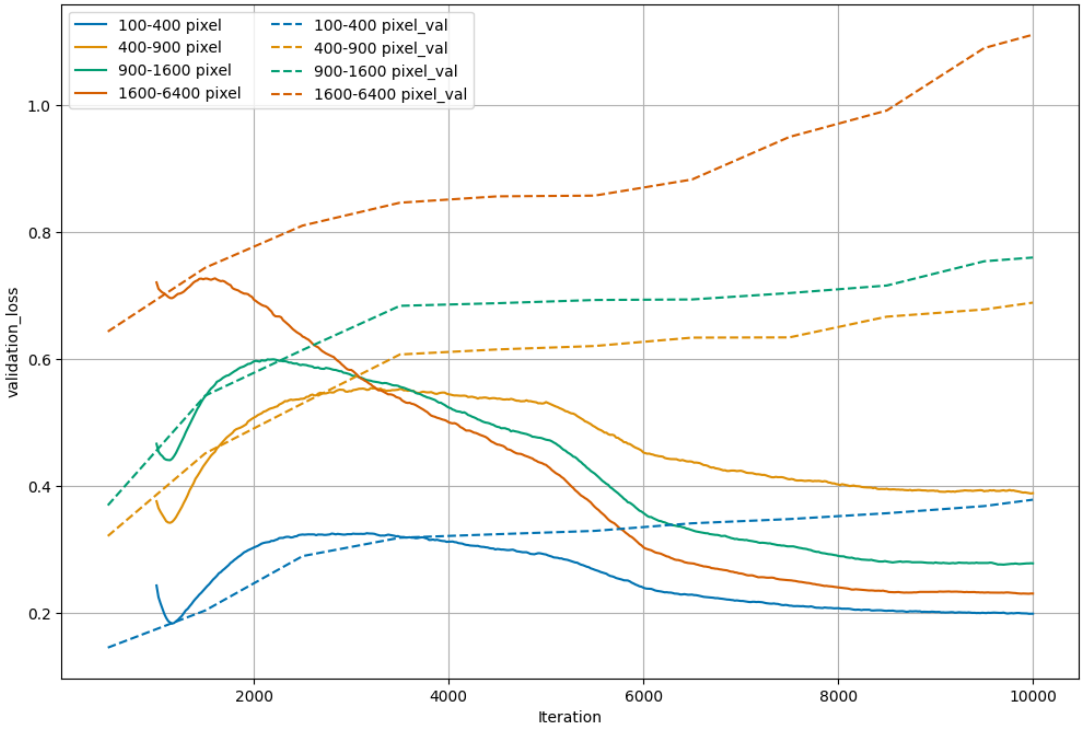
Figure C.2: Distribution of tower sizes in training set. Tower size is given in the square root of bounding box size, corresponding to the anchor box size of the model.



(a) Box Regression Loss.



(b) Classification Loss.



(c) Total (solid) and Validation loss (dashed).

Figure C.3: Training loss curves. For the model specified to varying tower sizes at resolution 30 cm/pixel (configuration #1). The box regression and classification losses decrease across resolutions. Consequently, the total loss also decreases. The validation loss (measured every $\frac{1}{20}$ of iterations) increases almost throughout all tower sizes and iterations. For better readability, the training's loss curves are averaged over a sliding window of size $\frac{1}{10}$ of the training iterations.

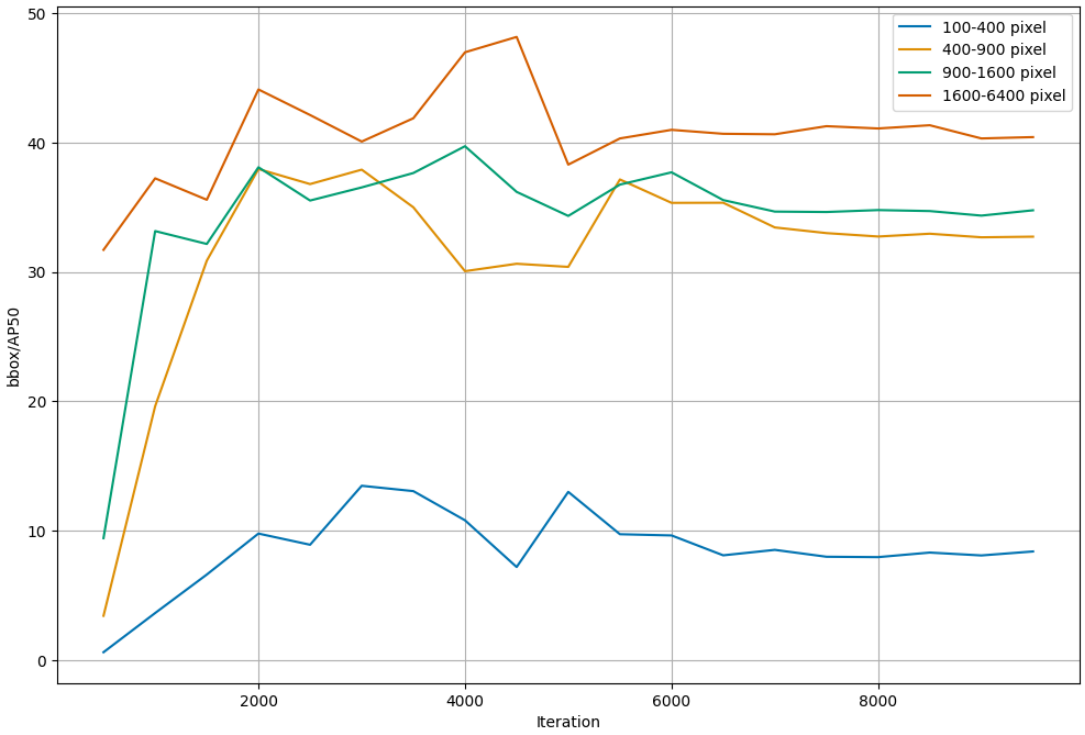
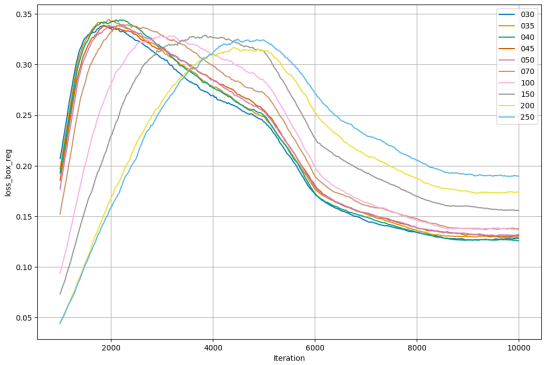
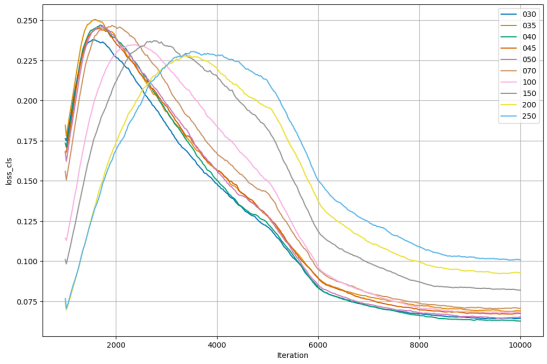


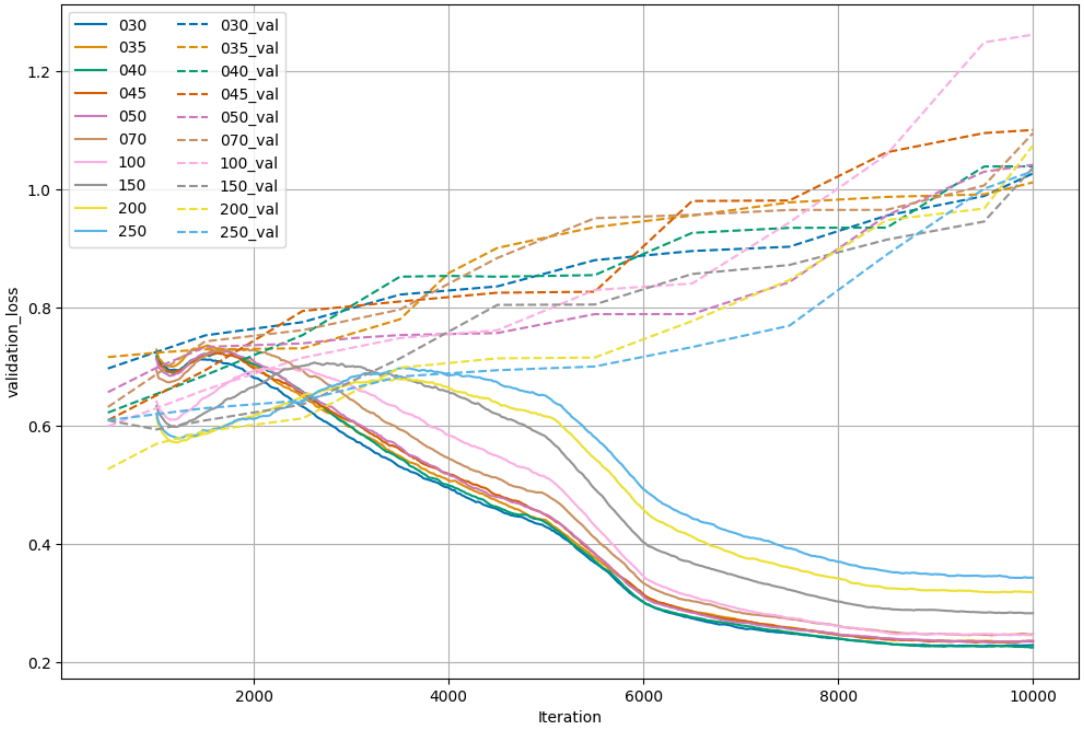
Figure C.4: AP50 during validation. AP50 score on validation set over the training iterations for models specified to different tower sizes.



(a) Box Regression Loss.



(b) Classification Loss.



(c) Total (solid) and Validation loss (dashed).

Figure C.5: Training loss curves. For the model specified to tower sizes 1600-6400 pixel for decreasing resolutions (configuration #1). The box regression and classification losses decrease across resolutions. Consequently, the total loss also decreases. The validation loss (measured every $\frac{1}{20}$ of iterations) increases almost throughout all resolutions and iterations. For better readability, the training's loss curves are averaged over a sliding window of size $\frac{1}{10}$ of the training iterations.

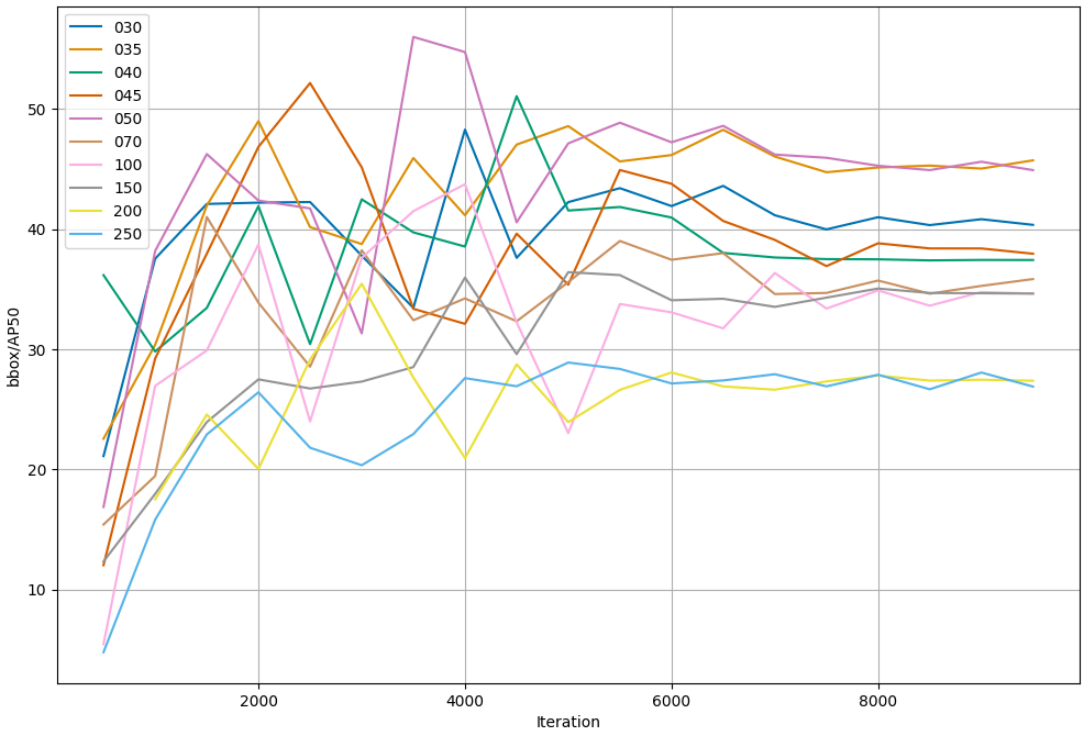


Figure C.6: AP50 during validation. AP50 score on validation set over the training iterations for the model specified to tower sizes 1600-6400 for different resolutions.

D Technical Hardware

All models were trained on the Hertie School GPU Cluster and one GPU at a time:

- GPU: 4 x NVIDIA A100 40GB HBM2
- GPU memory: 16GB
- Available RAM: 512 GB (8 x 64GB) ECC DDR4 3200 Mhz
- Disk Space: 3,8 TB
- Idle time until kernel cut-off: no limit
- Maximum continuous training: no limit.