

METHODS PAPER  

# Closing the domain gap: blended synthetic imagery for climate object detection

Caleb Kornfein<sup>1</sup> , Frank Willard<sup>1</sup>, Caroline Tang<sup>1</sup>, Yuxi Long<sup>1</sup>, Saksham Jain<sup>2</sup>, Jordan Malof<sup>3</sup>, Simiao Ren<sup>2</sup> and Kyle Bradbury<sup>2,4</sup> 

<sup>1</sup>Department of Computer Science, Duke University, Durham, NC, USA

<sup>2</sup>Department of Electrical & Computer Engineering, Duke University, Durham, NC, USA

<sup>3</sup>Department of Computer Science, University of Montana, Missoula, MT, USA

<sup>4</sup>Nicholas Institute for Energy, Environment & Sustainability, Duke University, Durham, NC, USA

**Corresponding author:** Kyle Bradbury; Email: [kyle.bradbury@duke.edu](mailto:kyle.bradbury@duke.edu).

**Received:** 01 March 2023; **Revised:** 23 July 2023; **Accepted:** 21 September 2023



**Keywords:** climate change; domain adaptation; energy infrastructure; object detection; remote sensing

## Abstract

Accurate geospatial information about the causes and consequences of climate change, including energy systems infrastructure, is critical to planning climate change mitigation and adaptation strategies. When up-to-date spatial data on infrastructure is lacking, one approach to fill this gap is to learn from overhead imagery using deep-learning-based object detection algorithms. However, the performance of these algorithms can suffer when applied to diverse geographies, which is a common case. We propose a technique to generate realistic synthetic overhead images of an object (e.g., a generator) to enhance the ability of these techniques to transfer across diverse geographic domains. Our technique blends example objects into unlabeled images from the target domain using generative adversarial networks. This requires minimal labeled examples of the target object and is computationally efficient such that it can be used to generate a large corpus of synthetic imagery. We show that including these synthetic images in the training of an object detection model improves its ability to generalize to new domains (measured in terms of average precision) when compared to a baseline model and other relevant domain adaptation techniques.

## Impact Statement

Existing methods of gathering information about energy and climate-related infrastructure and their impacts rely on self-reported information from organizations and governments, large-scale surveys, or crowd-sourced information. Object detection using overhead imagery offers a fast and low-cost way of identifying climate-related objects and other characteristics visible from above. These techniques, if scaled up, could democratize access to climate and energy infrastructure information cheaply and globally. Yet, as we show, the performance of object detection models can suffer when applied to geographically different regions, hindering the wider application of these models. To aid the broader applicability of detecting important climate infrastructure, we propose a domain adaptation technique that uses easy-to-generate synthetic overhead images across a range of diverse geographies. We show that this technique outperforms many alternative domain adaptation techniques, is simple to implement, and can be widely applied to other object detection problems.

  This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

## 1. Introduction

From power plants to wildfires, many of the causes and consequences of climate change are visible from overhead imagery. Accurate geospatial information about the causes of climate change, including energy infrastructure systems, is critical to planning climate change mitigation and adaptation strategies. However, spatial data on current energy infrastructure is often lacking. The data may not be publicly available, may be incomplete, or may not be of a sufficiently high resolution (Stowell et al., 2020). Recent research has demonstrated the potential of using satellite imagery to fill the data gaps by monitoring energy systems at unprecedented frequencies and scale (Donti and Kolter, 2021; Ren et al., 2022). Two remaining challenges to detecting climate objects at scale include (1) a lack of large datasets with labeled data for relevant applications, and (2) the difficulty of applying these techniques across diverse geographic domains.

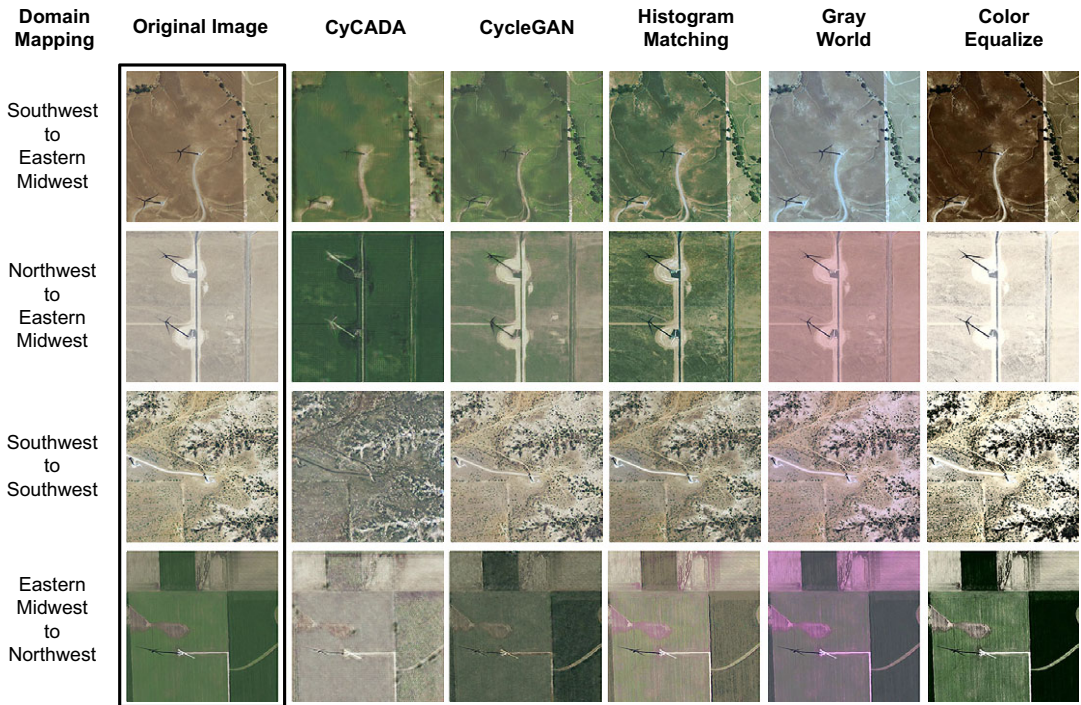
The first roadblock is the substantial number of labeled examples required to successfully train a machine learning model. Building a large dataset of overhead images can be challenging because of the labor-intensive nature of sifting through images and accurate data labeling (Tan et al., 2018). When the object of interest is rare and few instances may exist, the difficulty of detecting such an object is exacerbated. This is common for data related to energy and climate systems.

The second challenge is aligning the distribution of training images and testing images. It is frequently the case that the training and testing samples are not independently and identically distributed. The divide in the distribution of the images we use to train versus those we use to test our models is known as the domain gap. For example, we may only have training samples of an object in arid landscapes, but we might want to detect the object in mountainous or forestland landscapes as well. When the training samples come from different distributions than the testing samples, object detection performance deteriorates. Deep neural network models can be unreliable in cases where a domain gap exists (Tuia et al., 2016). Throughout this paper, we use the term geographic domain to refer to a landscape or region with an underlying distribution of visual properties that is regionally similar, but potentially distinct from other regions. In this case, when training on samples that originate from one geographic domain and testing on another, the potential for deteriorated object detection or segmentation performance, which we refer to as a domain gap, exists. Example images from diverse geographic domains can be found in Figure 2.

In their work, Tuia et al. (2016) lay out four general strategies for domain adaptation methods for remote-sensed classification: invariant feature selection, adapting the data distributions, adapting the model/classifier, and selective sampling. These strategies can be combined together or used separately for domain adaptation. Our contribution is a new technique that falls under the category of adapting data distributions proposed by Tuia et al. Changing the data distributions can mean expanding or transforming the training set in a way that enhances the ability of the algorithm to perform well on alternative domains (Shorten and Khoshgoftaar, 2019). In remote sensing applications, this enables the overhead imagery in a training dataset to become more representative of the data in the testing domains. This category of techniques also includes incorporating additional synthetic imagery to help close the domain gap by exposing a model to a broader variety of data (e.g., Synthinel-1, Kong et al., 2020; SIMPL, Hu et al., 2021; Xu et al., 2022).

Image transformation techniques are often color or pixel-based transformations, which leave image content unchanged but vary the color of pixels in an image, mapping pixels of each color to a new value. For example, histogram equalization adjusts image pixel intensity values such that they follow a uniform distribution to standardize image appearance regardless of domain. These pixel-wise techniques aim to better align the training and testing image sets. These techniques have the benefit of being able to transform a dataset in place without additional labels while alternative techniques generate unique synthetic images to add to the training set. Generative adversarial network (GAN)-based synthetic generation methods may modify both the pixel values and the information content of an image. These are neural models that learn a highly complex and expressive mapping between two domains. Examples of GAN-based models that have been used to generate synthetic training samples include CyCADA (Hoffman et al., 2018) and CycleGAN (Zhu et al., 2017).

In this work, we confine our comparisons to transformation and augmentation-based approaches. Within the subclass of color-based image transformation techniques, we compare to popular methods



**Figure 1.** Example images from selected domain adaptation methods. For each domain mapping, the original image is shown in the first column, while each image to the right shows that same image transformed by the technique to look like it came from the target domain. A detailed mapping of the domains can be found in [Figure 2](#).

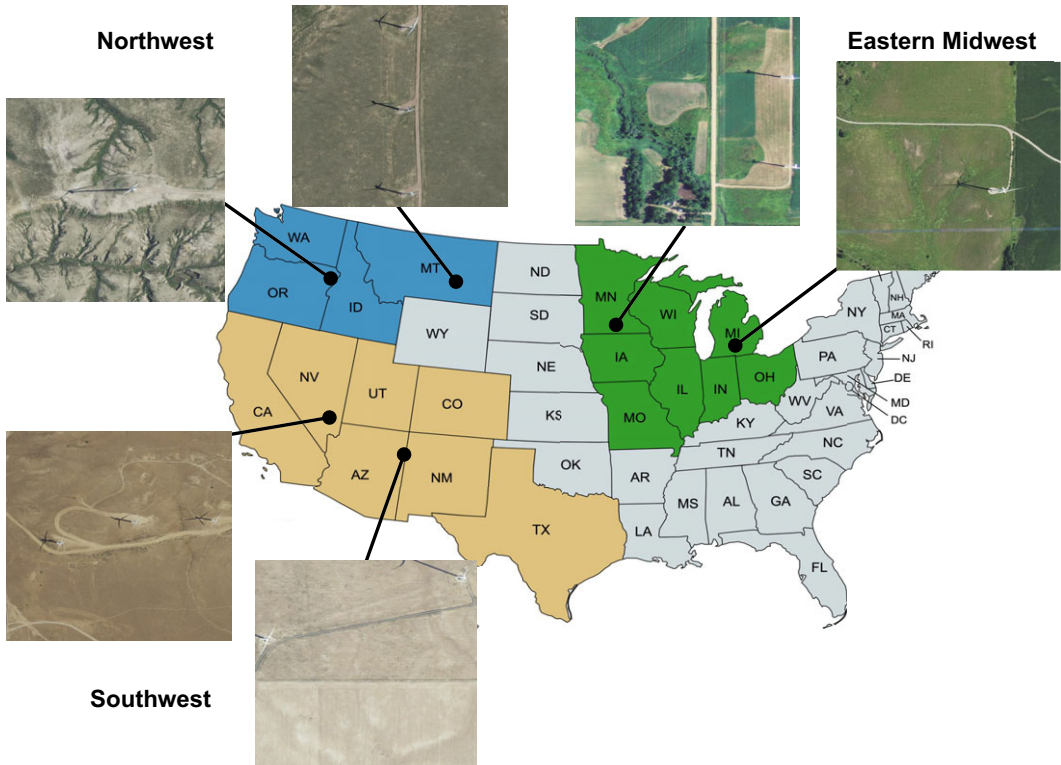
such as the gray world mapping (which assumes each color channel averages to gray) (Kanan and Cottrell, 2012), histogram matching (Abramov et al., 2020), and color equalization (Mustafa and Kader, 2018). We also compare the GAN-based synthetic image generation techniques to CycleGAN and CyCADA. We chose CycleGAN and CyCADA because our experiments simulate the detection of a rare climate object with minimal access to labeled data from the source domain and no labeled data from the target domain. While many other GAN-based techniques require labeled examples from both domains (a supervised context), CycleGAN and CyCADA do not have this requirement. Example images generated using each technique we compare to can be found in [Figure 1](#).

We contribute to the field of domain adaptation in remotely sensed data by proposing a new technique to generate synthetic overhead images that require minimal labeled examples of the target object. Our technique takes advantage of unlabeled overhead images to reduce the number of labeled examples required. Unlabeled remote sensed images are often easily acquirable through public datasets such as the National Agriculture Imagery Program (NAIP). Using the detection of wind turbines as a case study, we run a series of experiments and benchmarks demonstrating the benefit of our technique in augmenting training datasets. Our experiments show the potential of our technique to improve downstream model performance as compared to other domain adaptation techniques in this space—especially in situations with limited training data or when applying a model to new geographies.

## 2. Methodology

### 2.1. Dataset creation

To train our object detection models and test our domain adaptation technique, we created a dataset of overhead images containing wind turbines. Wind turbines were selected as an example of climate



**Figure 2.** Sample images and corresponding locations from our chosen geographic domains: the Northwest, Southwest, and Eastern Midwest United States.

infrastructure for three reasons. First, they are relatively homogeneous in appearance which helps to minimize intra-class variance and aid object detection. Second, they are found in a diverse variety of geographies and contexts (mountains, fields, etc.). Finally, they are relatively rare in occurrence, as the official 2021 U.S. Wind Turbine Database (Hoen et al., 2018) contains only 68,714 turbines across the entire United States.

We collected geographically diverse images from the Northwest, Southwest, and Eastern Midwest regions of the United States. We selected these regions, shown in Figure 2, with the intention of creating visually distinct domains. Since visual distinctiveness does not guarantee the presence of a domain gap, we also verified the presence of a domain gap experimentally (see Figure 6).

In each domain, we collected a set of training and testing images containing wind turbines using coordinates from the U.S. Wind Turbine Database. These covered the three geographies mentioned earlier (Northwest, Southwest, and Eastern Midwest regions of the United States). Overhead images over the selected coordinates were collected from the National Agriculture Imagery Program dataset (NAIP, 1 meter per pixel resolution) using Google Earth Engine (Gorelick et al., 2017).

To ensure that the training and testing sets had similar levels of variation in each domain, we used stratified geographic sampling. In this way, the coordinates of the labeled wind turbines from each region were first clustered using DBSCAN (Ester et al., 1996), and subsequently, the training and testing coordinates were selected using stratified random sampling to ensure representative sampling within each region. This avoided, for example, having a Northeast domain with training data only from Massachusetts and testing data only from New York.

Since we were using the coordinates of wind turbines for capturing overhead images, we wanted to ensure that the turbine was not in the center of every image in the dataset. To avoid this issue, we shifted

the coordinate center of each image uniformly randomly up to 75 meters both horizontally and vertically. The dimensions of each of the final images are  $608 \times 608$  pixels.

Finally, each image was quality-checked manually to ensure that wind turbines were present in the image and that the training and test datasets had no overlap.

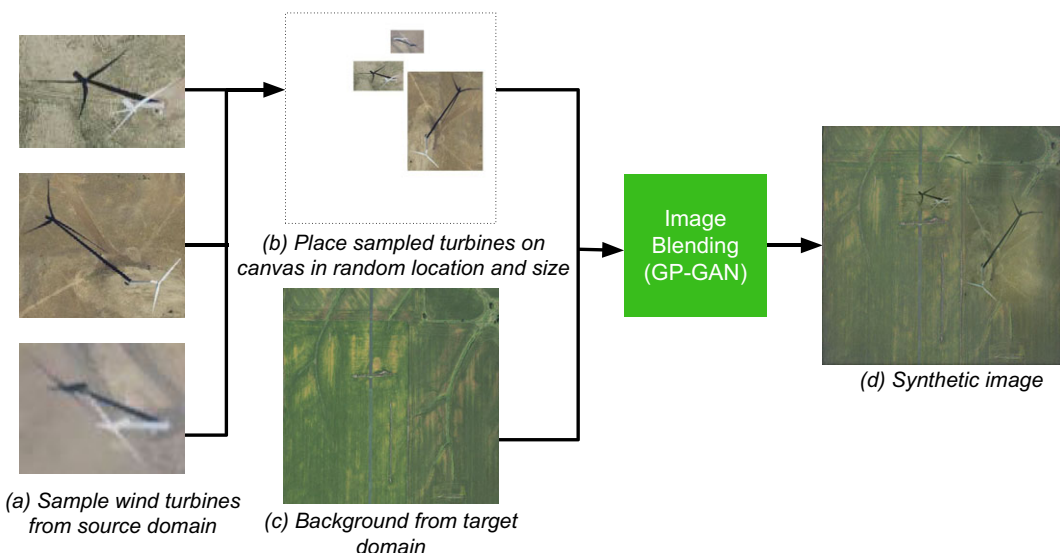
We also collected a supplementary set of images without turbines for each domain which we call “background” images. Our synthetic image generation technique uses these unlabeled images as a canvas on which to blend the object of interest. We captured images at a distance between four and six kilometers away from a known turbine location within the same domain, which were manually inspected to ensure that no wind turbines were present. The close distance was chosen to ensure visual similarity to the domain. The total distribution of images across domains and the number of labeled wind turbines contained within those images used for training, image generation, and validation can be referenced in Table E1 in Appendix E.

## 2.2. Generation of synthetic images

Our image generation process aims to produce synthetic images that are as similar as possible to real labeled data from any potential target domain. We use a pre-trained GP-GAN image blending model from Wu et al. (2019) to blend background target domain images together with source domain target objects. The GP-GAN model was pre-trained on webcam images depicting diverse seasonal, weather, and lighting settings from the Transient Attributes dataset by Laffont et al. (2014). Our process consists of four steps, as shown in Figure 3:

1. Sample a random background image from the *target* domain.
2. Sample objects from a set of *source* labeled domain objects (in our case wind turbines) from any domain.
3. Randomize the location, orientation, and size of the objects.
4. Blend the randomized objects into the target domain background image using GP-GAN.

One benefit of our technique is its ability to generate synthetic images while using only a few examples of the object. Namely, a small number of target object examples can be blended into as many contexts as desired. Second, the synthetic images can be stylized to look similar to any target domain given only



**Figure 3.** Diagram depicting our process to generate GP-GAN synthetic images. In brief, sampling target objects, randomizing their locations, selecting a background image, and blending via GP-GAN.

unlabeled images from that domain. Acquiring unlabeled datasets of overhead imagery is an easier task than manually labeling a dataset from the target domain. Although we manually inspected the background images to ensure that no turbines were present, for rare objects, it may be reasonable to assume that randomly captured snapshots from the target domain do not include the objects. In such cases, images from the target domain could be directly used as backgrounds without manual inspection. A final advantage is our technique's customizability along important dimensions, such as the number of object instances blended in. This stands in contrast to fixed color-mapping techniques, which simply transform an image in place and cannot control the number of object instances. Other controllable hyper-parameters include the spacing, size, and scale of the objects. Overall, our technique has unique advantages that could make it beneficial for domain adaptation contexts where the object is rare.

In our experiments, we used the background images from each domain as our GP-GAN canvases. We blended turbine examples from the source domain into target domain backgrounds. Our turbine examples included the shadows, as we thought this could give the object detection algorithm important contextual information. We customized the size and placement of the synthetically added turbines such that the wind turbines never overlapped one another, the size of the examples remained unchanged, and each synthetic image contained three turbines. Three turbines were chosen because we wanted the synthetic images to include ample examples to learn from and, for each source domain, three was the 90th percentile of the number of turbines in our training images.

Some of the generated synthetic images contain artifacts including artificially bright spots and blurred blending borders or turbines, as can be seen in [Figure 4](#). While not all of the data are visually perfect, this work evaluates whether or not these synthetic images are effective in overcoming performance differences between domains from an object detection perspective. Additional refinements to the image blending process would likely further enhance the performance improvements achieved through this work.

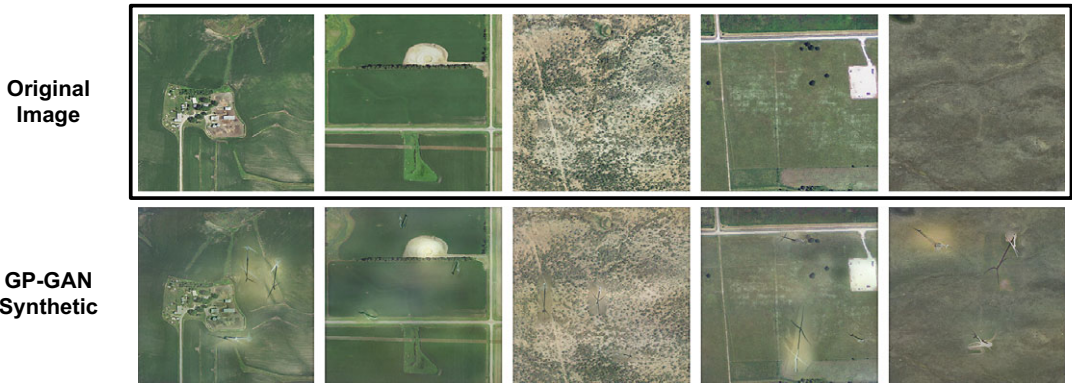
### 2.3. Experimental setup

Our experiments trained object detection models to detect wind turbines across a variety of geographic domains and to compare the performance of an object detection model trained using synthesized data versus other image transformation and augmentation approaches. Many types of infrastructure are quite rare, so this work focused on investigating applications where the objects were also rare. These could either be objects that are rare in a global sense, if the object is unlikely to be found in a randomly sampled point on Earth (a stricter requirement), or rare regionally within one domain (a likely situation to encounter when the availability of regional, high-resolution satellite imagery is expensive to collect).

While climate- and energy-relevant infrastructure and resources are often rare in the context of imagery for any given region, we made the assumption that we would be able to acquire more images from some domain, even if that is typically not the target domain, and selected that number to be 100, which is on the high end of past studies. For example, in [Martinson et al. \(2021\)](#) 10–50 images are used for training in the context of “rare” objects and [Wang et al. \(2019\)](#) used between 10 and 30 images in their investigations. Many few-shot techniques use as few as 1–10 instances such as in [Wang et al. \(2020\)](#) while the highest that we have seen was from [Xu et al. \(2022\)](#), which varied from 0 to 151.

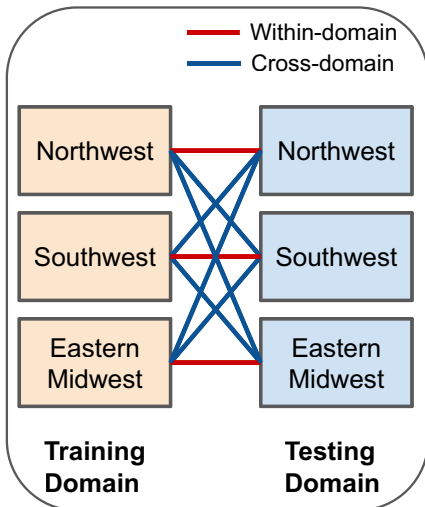
To simulate detecting a rare object, we assumed access to a small corpus of labeled wind turbine examples in the training domain and no additional wind turbine labels from the target domain. Each experiment tested all of the possible source/target domain pairings across the three domains: Northwest, Eastern Midwest, and Southwest, for nine pairings in total.<sup>1</sup> For each pair of domains, we trained and tested a YOLOv3 model five separate times to account for variation in training and to estimate model variance.

<sup>1</sup> With our three domains of Northwest (NW), Eastern Midwest (EM), and Southwest (SW), these pairings are: (1) train NW, test NW; (2) train NW, test EM; (3) train NW, test SW; (4) train EM, test NW; (5) train EM, test EM; (6) train EM, test SW; (7) train SW, test NW; (8) train SW, test EM; and (9) train SW, test SW.

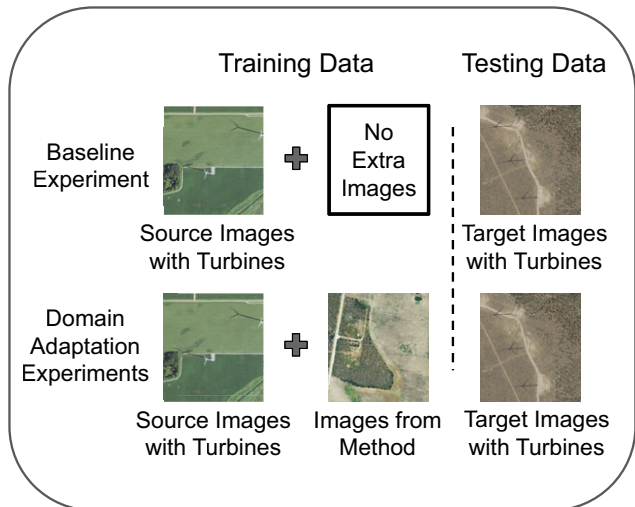


**Figure 4.** Example GP-GAN blended synthetic images. The upper original images without turbines have been transformed into the lower GP-GAN synthetic images by blending in turbine examples.

**1 - Select Domains**



**2 - Run Experiments**



**Figure 5.** The experimental setup. First, a pairing of a source and a target domain is selected. For each pairing, a baseline experiment is run as a benchmark by using labeled source domain images and labeled test images from the target domain. Then, a series of domain adaptation experiments are run that augment the baseline training set with additional supplemental images produced from a domain adaptation method.

For an experimental baseline, we evaluated model performance using the training data for each pair of domains without using any domain adaptation methods. The baseline experiments were trained on 100 labeled source domain images and tested on 100 labeled target domain images across all domain pairs.

Then we compared our approach and other domain adaptation techniques involving image transformation or augmentation to the baseline experiments. We evaluated whether the addition of these different types of supplementary imagery (including our synthetic images) could improve wind turbine detection performance across domains. These additional experiments were trained on 100 labeled source domain images, *supplemented* with an additional 100 images from the domain adaptation technique under evaluation, and were tested on 100 labeled target domain images. The experimental configurations are

shown in Figure 5. More information about the number of images and turbine labels used from each domain can be found in Appendix E.

We differentiated between *within-domain* experiment trials where the detection model was trained and tested on imagery from the *same* geographic domain, versus *cross-domain* experiment trials in which the model was trained on one domain and tested on a *different* domain. As an example, in an experiment, a within-domain trial may be trained on images of wind turbines from the Northwest U.S. and validated on other images from the Northwest. In contrast, a cross-domain trial may be trained on images from the Northwest and tested on images from the Southwest. We expected object detection performance to suffer in cross-domain contexts due to the presence of a domain gap.

For our wind turbine object detection model, we used a YOLOv3 model architecture with spatial pyramid pooling (Redmon and Farhadi, 2018). For each experimental trial, a YOLOv3 model was trained from scratch on the available training images.

Lastly, an important note on the experimental design that we want to highlight is that we used a fixed mixed batch ratio of real-to-synthetic data for all the experiments for this work with a mini-batch size of eight images. This allowed us to control exactly how many train and supplemental images were in each mini-batch and thus the relative influences of each image set. In a given mini-batch, seven of the images were from the baseline set while one image was from the supplementary set (the images generated through the domain adaptation method). One of the challenges we observed with this method is that synthetic data was not a perfect replacement for real data: if we used all synthetic data rather than real data, the cross-domain performance dropped substantially as shown in Figure B1 in Appendix B. We selected a real-to-synthetic mixed batch ratio of 7-to-1 a priori to ensure the presence of synthetic data in every minibatch. We varied the number of synthetic images in each mixed batch for GP-GAN training and, as seen in Figure B1 in Appendix B, we found that between 1 and 7 were effective ratios, while 0 or 8 (no synthetic, or all synthetic) performed markedly worse.

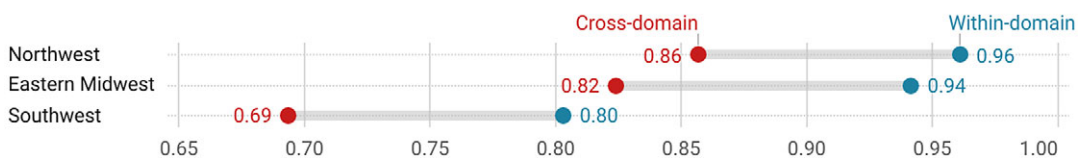
In the mixed batch setup, we fixed a number of images seen as an epoch. Since our typical experimental setup consisted of 100 training and 100 supplemental images, we defined an epoch as having passed 200 images or 25 mini-batches through the training process. Each model was trained over 300 epochs.

We evaluated model performance using average precision (AP). We also reported 95% confidence intervals (CI) constructed with *t*-distributions alongside AP results. Where applicable these intervals may represent the average of multiple confidence intervals—see Appendix C for further details on variability reporting.

### 3. Experimental results

We experimentally demonstrated the presence of a domain gap through the results of the baseline experiment which can be seen in Figure 6. On average, the cross-domain AP performance of our object detection model was 12% worse than the within-domain AP performance for each test domain.

Average precision by test domain



**Figure 6.** Baseline experimental results demonstrating evidence of a domain gap. In this plot, all domain pairings with the same test domain are grouped together and divided into the within- and cross-domain settings. The gap in performance between within-domain and cross-domain settings is shown in this figure as the distance between the red and blue points.



**Table 1.** Synthetic experiment results compared to the baseline experiment using average precision

Source domain	Target domain	Baseline CI	Synthetic CI
EM	EM	0.941 ± 0.018	<b>0.964 ± 0.015</b>
NW		0.905 ± 0.036	<b>0.930 ± 0.023</b>
SW		0.742 ± 0.042	<b>0.836 ± 0.030</b>
EM	NW	0.894 ± 0.015	<b>0.936 ± 0.014</b>
NW		0.960 ± 0.012	<b>0.963 ± 0.004</b>
SW		0.818 ± 0.017	<b>0.862 ± 0.026</b>
EM	SW	0.647 ± 0.034	<b>0.800 ± 0.016</b>
NW		0.739 ± 0.031	<b>0.772 ± 0.019</b>
SW		0.802 ± 0.027	<b>0.837 ± 0.018</b>
Within-domain average		0.901 ± 0.019	<b>0.921 ± 0.013</b>
Cross-domain average		0.791 ± 0.029	<b>0.856 ± 0.021</b>

Each experiment was run five times and the mean trial result along with 95% confidence interval widths are shown. Bold results indicate the higher result when comparing between the baseline and synthetic experiments.

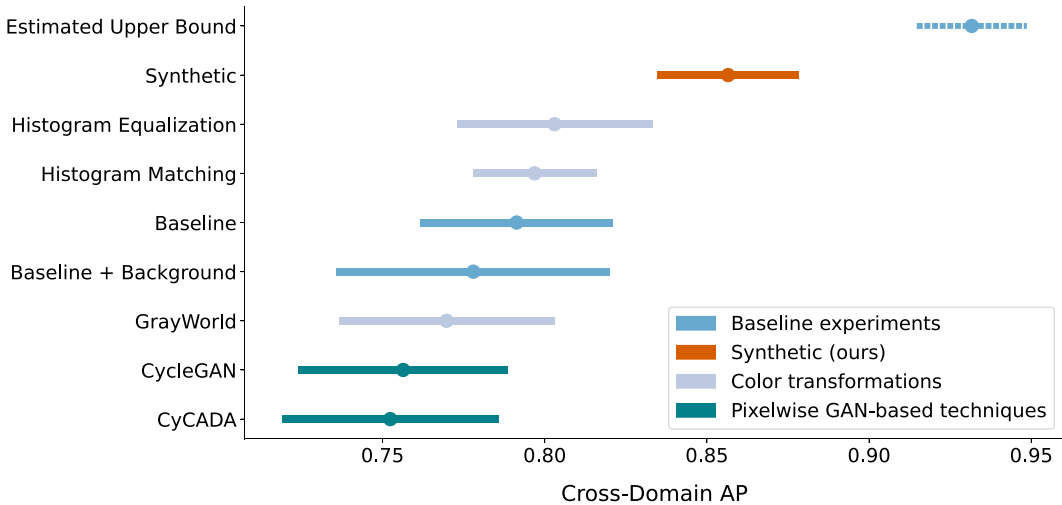
A full comparison of results between the baseline experiment and our synthetic images can be found in Table 1. Overall, baseline within-domain trials achieved an AP of 0.901 while baseline cross-domain trials achieved an average AP of 0.791. For each domain pair, the addition of synthetic images improved AP. This was especially true in a cross-domain context: on average, synthetic cross-domain trials achieved a 0.065 higher AP than baseline cross-domain trials, while synthetic within-domain trials achieved a 0.020 higher AP than baseline within-domain trials. Detecting turbines in the Southwest was especially challenging compared to other domains—we observed many small and clustered wind turbines in this domain and hypothesize that this could have contributed to weaker object detection performance.

We also compared our synthetic image blending technique to other domain adaptation techniques including histogram equalization, histogram matching, gray world, CyCADA, and CycleGAN.<sup>2</sup>

In addition to comparing each of the above techniques to each other and the baseline, we incorporated two further variations of the baseline experiment. First, to estimate an upper bound on the performance we might be able to achieve when training with 100 supplemental images, we supplemented the baseline training dataset with 100 additional *real* images from the *target* domain. Second, since we added 100 training samples of augmented or synthetic imagery in every experiment with a domain adaptation technique, we wanted to test if adding *unlabeled* imagery improved performance. If this were true then the domain adaptation techniques would not be enhancing performance. To investigate this, we supplemented the baseline experiment with 100 target domain background images (unlabeled target domain images without wind turbines present; these were the same images used for image blending in our synthetic case, but without having blended target objects).

The results of all these experiments are shown in Figure 7, with the estimated upper and lower-bound experiments shown in light gray. These experiments collectively tested whether adding each set of supplemental images improved cross-domain model performance when faced with highly limited training data availability (as is common with rare object detection, as is often the case with energy and climate applications). The results indicate that adding synthetically blended imagery is able to produce the greatest improvement in average precision of the techniques compared in this study. On average, our

<sup>2</sup> CycleGAN and CyCADA were trained from scratch for each domain mapping (e.g., from the Southwest to the Northwest domain). The models were trained using 100 background and 100 real images from the source domain and 100 background images from the target domain. CycleGAN and CyCADA learn a bidirectional mapping between a pair of domains. Thus, neither CycleGAN nor CyCADA are generic in their applicability more broadly than the specific domain pairings. Since these methods require the data for each new application, and such data would not generally be available in the context of rare objects, we limited the training data to be the same data as all of the rest of the algorithms were received to ensure a fair comparison.



**Figure 7.** Experimental results displaying the averaged 95% confidence intervals among cross-domain pairs. See Appendix C for more on how the intervals constructed.

synthetic image blending technique outperformed baseline trials by 8.2% in cross-domain pairings and 2.2% in within-domain pairings. A full table of results including the within-domain pairings can be found in Appendix A.<sup>3</sup>

There were many design choices that need to be made about these experiments including the number of synthetic images to include in the analysis, the number of objects per synthetic image, and the total number of labeled objects from the source domain available to the synthetic image blending tool. We investigated the sensitivity of the results to the parameters by varying each of them for one pair of domains: the Northwest and the Southwest. Those experimental results can be found in Appendix D. We found that as long as some amount of synthetic objects were included in the training process, we observed a performance improvement, but the results were not exceptionally sensitive to any of these parameters.

#### 4. Conclusions

The automated mapping of climate-relevant infrastructure from overhead imagery can help to fill critical information gaps, but global applications have been impeded by the challenges of geographic domain adaptation and a lack of training data. To combat this, we proposed a computationally inexpensive synthetic imagery generation technique designed to work with minimal labeled examples and help downstream models become more generalizable. Our approach uses GP-GAN to blend real images of an object onto unlabeled background images from the target domain. Our experiments provide evidence that supplementing training data with synthetically blended imagery can improve domain adaptation while requiring minimal time, no proprietary software, and few labeled object examples. This may aid in scaling up automated mapping to larger applications.

Future work could further refine our synthetic image generation methodology by adjusting the parameters of the image blending process and/or applying post-processing to generate even more realistic and well-blended synthetic images. Additionally, this approach could be applied to more energy and

<sup>3</sup> The “Baseline + Background” case, which adds additional unlabeled background images from the target domain to the source domain training data, resulted in a slightly lower mean cross-domain AP as compared to the baseline experiments. However, since on a domain-by-domain basis, the difference is within the confidence intervals of the two estimates, there was no statistically meaningful difference between the performance of those two experimental conditions.

climate objects for further evaluation. Lastly, this approach could be tested beyond one-to-one domain pairings to evaluate the technique in one-to-many, many-to-one, or many-to-many contexts.

**Acknowledgments.** We would like to acknowledge the support of the Duke University Data+ and Bass Connections programs as well as the Nicholas Institute of Energy, Environment & Sustainability, and the Rhodes Information Initiative. We would like to thank our collaborators Wei Pe, Madeleine Jones, Alena Zhang, Maddie Rubin, Alexander Kumar, Aya Lahlou, Boya (Jennie) Sun, and Katie Wu.

**Author contribution.** Conceptualization: K.B., J.M.; Data curation: C.K., F.W., Y.L., S.J.; Data visualization: C.K., F.W., C.T.; Methodology: C.K., F.W., S.R.; Writing—original draft: C.K., K.B.; Writing—review and editing: All authors. All authors approved the final submitted draft.

**Competing interest.** The authors declare none.

**Data availability statement.** Source code for this technique and all experiments is available at <https://github.com/energydatalab/closing-the-domain-gap>. A repository containing only the technique implementation for broader use can be found at <https://github.com/frankwillard/Blended-Synthetic-Imagery-for-Climate-Object-Detection/>. Our dataset of aerial images containing wind turbines can be downloaded at <https://zenodo.org/record/7385227#.Y4hf-zMKw5>.

**Ethics statement.** The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

**Funding statement.** This research was supported by the Duke University Bass Connections program and the Duke University Data + program.

## References

- Abramov A, Bayer C and Heller C (2020) *Keep it simple: Image statistics matching for domain adaptation*. *arXiv*.
- Bouthillier X, Delaunay P, Bronzi M, Trofimov A, Nichyporuk B, Szeto J, Mohammadi Sepahvand N, Raff E, Madan K, Voleti V, Ebrahimi Kahou S, Michalski V, Arbel T, Pal C, Varoquaux G and Vincent P (2021) Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems* 3, 747–769.
- Donti PL and Kolter JZ (2021) Machine learning for sustainable energy systems. *Annual Review of Environment and Resources* 46 (1), 719–747.
- Ester M, Krieger H-P, Sander J and Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD '96*. Washington, DC: AAAI Press, pp. 226–231.
- Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D and Moore R (2017) *Google earth engine: Planetary-scale geospatial analysis for everyone*. *Remote Sensing of Environment*.
- Hoen B, Diffendorfer J, Rand J, Kramer L, Garrity C and Hunt H (2018) United States wind turbine database (version 3.3, January 2021). *US Geological Survey, American Wind Energy Association, and Lawrence Berkeley National Laboratory Data Release*.
- Hoffman J, Tzeng E, Park T, Zhu J-Y, Isola P, Saenko K, Efros A and Darrell T (2018) Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*. London: PMLR, pp. 1989–1998.
- Hu W, Feldman T, Lin E, Moscoso JL, Ou YJ, Tarn N, Ye B, Zhang W, Malof J and Bradbury K (2021) Synthetic imagery aided geographic domain adaptation for rare energy infrastructure detection in remotely sensed imagery. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*.
- Kanan C and Cottrell G (2012) Color-to-grayscale: Does the method matter in image recognition? *PLoS One* 7, e29740.
- Kong F, Huang B, Bradbury K and Malof J (2020) The synthinel-1 dataset: A collection of high resolution synthetic overhead imagery for building segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. Snowmass, CO: IEEE, pp. 1814–1823.
- Laffont P-Y, Ren Z, Tao X, Qian C and Hays J (2014) Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics* 33(4), 1–11.
- Martinson E, Furlong B and Gillies A (2021) Training rare object detection in satellite imagery with synthetic gan images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, TN: IEEE, pp. 2763–2770.
- Mustafa WA and Kader MMA (2018) A review of histogram equalization techniques in image enhancement application. *Journal of Physics: Conference Series* 1019, 012026.
- Redmon J and Farhadi A (2018) *Yolov3: An incremental improvement*. *arXiv*.
- Ren S, Hu W, Bradbury K, Harrison-Atlas D, Malaguzzi Valeri L, Murray B and Malof JM (2022) Automated extraction of energy systems information from remotely sensed data: A review and analysis. *Applied Energy* 326, 119876.
- Shorten C and Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. *Journal of Big Data* 6(1), 60.

- Stowell D, Kelly J, Tanner D, Taylor J, Jones E, Geddes J and Chalstrey E** (2020) A harmonised, high-coverage, open dataset of solar photovoltaic installations in the UK. *Scientific Data* 7(1), 394.
- Tan C, Sun F, Kong T, Zhang W, Yang C and Liu C** (2018) A survey on deep transfer learning. In *International Conference on Artificial Neural Networks*. Cham: Springer.
- Tuia D, Persello C and Bruzzone L** (2016) Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine* 4(2), 41–57.
- Wang X, Huang TE, Darrell T, Gonzalez JE and Yu F** (2020) Frustratingly simple few-shot object detection.
- Wang Y-X, Ramanan D and Hebert M** (2019) Meta-learning to detect rare objects. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea: IEEE, pp. 9924–9933.
- Wu H, Zheng S, Zhang J and Huang K** (2019) GP-GAN: Towards realistic high-resolution image blending. *ACMMM*.
- Xu Y, Huang B, Luo X, Bradbury K and Malof JM** (2022) SIMPL: Generating synthetic overhead imagery to address custom zero-shot and few-shot detection problems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15, 4386–4396.
- Zhu J-Y, Park T, Isola P and Efros AA** (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. Venice, Italy: IEEE, pp. 2223–2232.

## Appendix A: Additional domain adaptation experiment results

Alongside AP, we measured experiment results using the percentage closure of domain gap (CDG) performance metric (Hu et al., 2021). CDG measures the percent closure of the gap between the within-domain and cross-domain trials that the current experiment recovers, assuming the within-domain represents the high end of performance for the given scenario. This is calculated as shown in equation A-1. Here, AP(E) represents the average precision resulting from experiment E.

$$\text{CDG} = \frac{\text{AP}(\text{cross-domain experiment}) - \text{AP}(\text{cross-domain baseline})}{\text{AP}(\text{within-domain baseline}) - \text{AP}(\text{cross-domain baseline})}. \quad (\text{A-1})$$

A full comparison of techniques using the CDG metric as well as both cross-domain and within-domain AP are shown in Table A1. Our synthetic imagery achieved the highest cross-domain and within-domain AP out of any experiment except the estimated upper-bound experiment. Additionally, our synthetic imagery achieved the highest closure of the domain gap out of any technique compared to in this study.

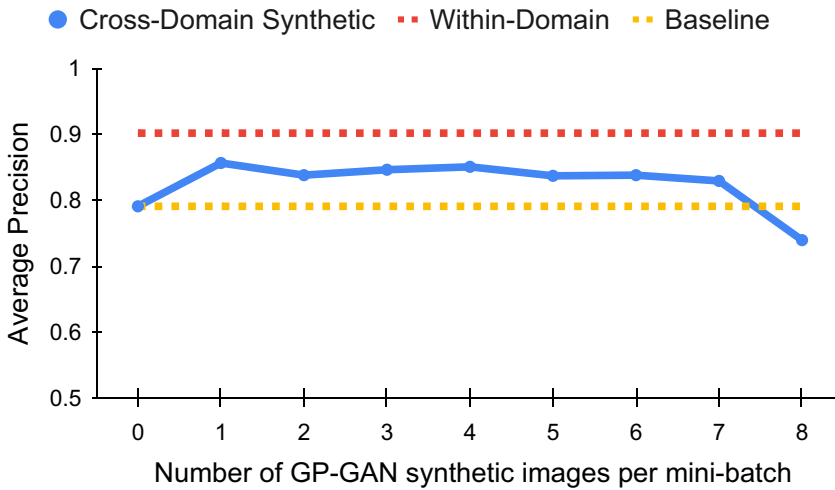
**Table A1.** Comparison of synthetic imagery to alternative domain adaptation techniques

Experiment	Cross-domain	Within-domain	CDG (%)
Synthetic	0.856	0.921	60.2
CyCADA	0.752	0.877	−39.2
CycleGAN	0.756	0.889	−30.7
Gray world	0.769	0.891	−15.1
Histogram matching	0.796	0.906	13.1
Histogram equalization	0.803	0.909	16.7
Estimated upper bound	0.931	0.938	139.4
Baseline	0.791	0.901	0.0
Baseline + Background	0.778	0.903	−16.3

Outcomes measured in average trial AP and split between the cross-domain and within-domain trials.

## Appendix B: Mixed batch ratio experiment

All domain adaptation experiments were trained using a mixed batch procedure with a mini-batch size of 8 and a mixing ratio of 7-to-1 real images to supplemental images. While this gave us precise control over the relative influences of the real and supplemental images across experiments, we conducted additional experiments with our GP-GAN synthetic images using all possible mixing ratios. The goal was to test how sensitive model performance was to the proportion of synthetic imagery used during training to see if the mixing ratio used during experiments was a reasonable choice. The results of these experiments can be found in Figure B1. In all realistic cases, the inclusion of our synthetic imagery improved performance. Only in the experiment where we trained using mini-batches with entirely synthetic data did performance degrade below baseline levels.



**Figure B1.** Average model performance (AP) over all domain pairs as a function of the number of GP-GAN synthetic images included in each mini-batch of 8. The case of 0 represents the baseline experiment as there are no synthetic images per mini-batch, while 8 represents training entirely using synthetic images. The red line represents the baseline experiment within-domain average, which the synthetic experiment hopes to approach. The blue line represents the baseline case of training entirely using non-synthetic imagery.

### Appendix C: Reporting results

To quantify the variability inherent in the training process and place confidence bounds around our reported results, we followed the example of Bouthillier et al. (2021) to report variability estimates of our model results. As each trial was repeated five times, we constructed 95% confidence intervals using a  $t$ -distribution with 4 degrees of freedom. In cases such as Table 1 where all the domain pairs are enumerated, these confidence intervals were exact. The mean of the trials along with the confidence interval bounds were also reported. In other instances, cross-domain or within-domain results were reported, which average information from multiple domain pairs and thus average multiple confidence intervals to give a sense of overall performance. These include Figure 7 and the cross-domain and within-domain reporting in Tables 1 and D1–D3. In this case, the mean of all the cross-domain trials was reported along with the mean confidence interval width.

### Appendix D: Sensitivity analysis

In this section, we present additional experiments to evaluate the sensitivity of object detection performance to various parameters of the synthetic data generation process. In particular, we varied (1) the number of labeled source domain turbines available to the image blender for generating the synthetic images, (2) the number of turbines present in each synthetic image, and (3) the number of synthetic images generated and included in each experiment.

We chose a single pair of domains to run these experiments on, namely the Northwest and Southwest domains. Each experiment here varied exactly one aspect of the experimental design. The default parameters included using the maximum number of turbine objects available from the source domain to generate synthetic images (see Table E1 for these numbers), blending in exactly three turbines per synthetic image, and generating 100 synthetic images per domain.

First, we varied the number of wind turbine objects available to the synthetic image generator. For each of the five trials testing  $N$  objects available to the generator, the  $N$  objects were randomly sampled from scratch each time.

Table D1 shows the cross-domain results. When few objects were present (0–5) the performance was lowest, and as long as there were about 10 or more objects available, a clear gain in performance was generally present. However, there was not a significant change as the number of objects available continued to increase.

Next, we varied the number of turbines per synthetic image. For all other results presented in this work, we used three per image. The results of this sensitivity analysis can be found in Table D2. The results show that for this domain pairing, object detection performance was not highly sensitive to the number of objects in the image as long as there was more than 1 turbine per image. The range of 2–5 resulted in approximately equally well-performing object detection performance.

The final experiment varied the number of synthetic images included in the supplemental image set. Our technique has no upper limit on the number of images you can generate given a set of object examples and background images. For the experiments

**Table D1.** Synthetic experiment results varying the number of labeled turbine instances accessible by the synthetic image generator

Objects available	Cross-domain CI	Within-domain CI
0	0.784 ± 0.026	0.887 ± 0.012
1	0.774 ± 0.039	0.884 ± 0.016
5	0.766 ± 0.040	0.895 ± 0.016
10	0.794 ± 0.022	0.883 ± 0.027
25	0.790 ± 0.032	0.878 ± 0.016
50	0.803 ± 0.028	0.882 ± 0.011
100	0.777 ± 0.020	0.857 ± 0.105
150	0.799 ± 0.029	0.887 ± 0.014

Zero objects available refers to the baseline plus background experiment where the supplemental background images contain no turbines. The within-domain results for 100 unique objects are dragged down by a single outlier trial which ran to completion successfully but had an AP value of 0.459.

**Table D2.** Density experiment results varying the number of turbines contained in each synthetic image

Turbines per Image	Cross-domain CI	Within-domain CI
0	0.784 ± 0.026	0.887 ± 0.012
1	0.782 ± 0.020	0.877 ± 0.014
2	0.788 ± 0.017	0.895 ± 0.019
3	0.804 ± 0.017	0.877 ± 0.024
4	0.789 ± 0.026	0.893 ± 0.011
5	0.805 ± 0.017	0.888 ± 0.023

Zero objects available references the baseline plus background experiment where the supplemental background images contained no turbines.

**Table D3.** Synthetic image experiment varying the number of synthetic images included

Synthetic images	Cross-domain CI	Within-domain CI
0	0.779 ± 0.024	0.881 ± 0.019
50	0.791 ± 0.022	0.893 ± 0.021
100	0.803 ± 0.018	0.900 ± 0.011
200	0.810 ± 0.011	0.896 ± 0.008
400	0.798 ± 0.027	0.887 ± 0.019
800	0.800 ± 0.017	0.894 ± 0.016

presented previously in this work, we generated 100 synthetic images throughout. As shown in Table D3, as long as there were 50 or more synthetic images present, the performance improvements from the added synthetic data were present. The greatest performance was in the range of 100–800 images but there was no meaningful difference in performance between adding 100 or 800 synthetic images.

The choice to use mixed-batch training means that during training a fixed number of supplemental images are seen. With one supplemental image per mini-batch, 25 mini-batches in an epoch, and 300 overall epochs, a total of 7500 synthetic images are seen per trial. Please refer to the notes on mixed batch training at the end of Section 2.3 for additional discussion of the mixed-batch training procedure.

## Appendix E: Experimental data

Table E1 gives the complete description of the image data used for experiments by domain. In each domain we collected at least 300 images containing turbines, reserving 100 images for training, 100 for the estimated upper bound experiment, and 100 for testing.

**Table E1.** Total number of images and labels used for training, synthetic image generation, and validation by domain

Domain	Training		Synthetic generation		Validation	
	Images	Labels	Backgrounds	Available objects	Images	Labels
EM	100	123	100	96	100	113
NW	100	270	100	204	100	277
SW	100	194	100	150	100	190
Totals	300	587	300	450	300	580

Train and test sets were generated using a two-part procedure: (1) DBSCAN was utilized to cluster the labeled wind turbine coordinates and (2) train and test coordinates were sampled using stratified random sampling from the clusters. As an additional data quality measure, we checked the distance between all train and test image coordinates and manually verified that the images were nonoverlapping. Each set of images was kept consistent across all experiments. The additional upper bound image sets of 100 images per domain contained an additional 125 labels for EM, 256 labels for NW, and 193 labels for SW, as coordinates were available for wind turbines in those regions.

Table E1 also indicates the subset of the total available labeled turbines from the training set of 100 images that were made available for synthetic image generation. To make an object available, we applied a bounding box around both the turbine and the shadow of the turbine for image blending. If the shadow or wind turbine was only partially visible in an image, the example was not used for blending. As shown in Appendix D, the results were not shown to be highly sensitive to the number of objects available for blending as long as there were at least 10, which there are for all conditions in this work.