Mapping Industrial Poultry Operations at Scale With Deep Learning and Aerial Imagery

Caleb Robinson, Ben Chugg, Brandon Anderson, Juan M. Lavista Ferres, and Daniel E. Ho

***Abstract*—Concentrated animal feeding operations (CAFOs) pose serious risks to air, water, and public health, but have proven to be challenging to regulate. The U.S. Government Accountability Office notes that a basic challenge is the lack of comprehensive location information on CAFOs. We use the U.S. Department of Agriculture's National Agricultural Imagery Program 1 m/pixel aerial imagery to detect poultry CAFOs across the continental USA. We train convolutional neural network models to identify individual poultry barns and apply the best-performing model to over 42 TB of imagery to create the first national open-source dataset of poultry CAFOs We validate the model predictions against held-out validation set on poultry CAFO facility locations from ten hand-labeled counties in California and demonstrate that this approach has significant potential to fill gaps in environmental monitoring.**

***Index Terms*—Concentrated animal feeding operations (CAFOs), convolutional neural networks (CNNs), deep learning, National Agricultural Imagery Program (NAIP), poultry barns, semantic segmentation.**

I. Introduction

ONE of the most substantial environmental challenges stems from modern agriculture. An area of specific concern has been concentrated animal feeding operations (CAFOs), which are intensive large-scale industrial farms that can pose serious environmental and health threats. In 2008, CAFOs produced more than 50% of the total livestock in the U.S. [1]. The number of CAFOs has continued to increase [2]; yet, obtaining reliable information on the precise locations and prevalence of CAFOs is difficult, due in part to litigation, limited regulatory capacity, and permit evasion [3]. Indeed, the U.S. Government Accountability Office noted that the Environmental Protection Agency (EPA) "does not have comprehensive, accurate information on the number of permitted CAFOs nationwide. As a result, EPA does not have the information it needs to effectively regulate these CAFOs" [4].

CAFOs present significant public health and environmental risks. A single CAFO can produce more manure than a large city of over 1 million people [4]. CAFOs produce more than 13 times the amount of human waste annually [6]. Such waste is often inadequately handled, leading to the contamination of nearby lands and waterways with pathogens [7], pharmaceuticals [8], heavy metals [9], and hormones [10]. This causes nitrogen and phosphorous runoffs, which severely affect water quality [11] and result in algal blooms [12]. CAFOs are also associated with air pollution. Living or attending school in proximity to CAFOs, even at a distance of some miles, is associated with reduced lung function and asthma [13], [14], [15]. Additional public health impacts of CAFOs include the propagation and incubation of disease and aggravating climate change [2]. Former Secretary of Energy Stephen Chu noted that agriculture and meat production may be more consequential to climate change than power generation [16], but such claims are hard to validate without a systematic enumeration of facilities. These challenges have given rise to recent work on automating CAFO detection and monitoring [17], [18], [19], [20].

Here, we augment this existing work by taking a deep learning approach to provide the first public, comprehensive, national map of poultry CAFOs. Our approach is fully automated, open source, and does not rely on the Census of Agriculture (CoA), a limitation of some of the existing methods (see Section II). Specifically, we employ a two-step modeling process. First, we use high-resolution (1 m/pixel) four-band (RGB and near-infrared) aerial imagery and poultry barn labels with an open dataset from the Delmarva Peninsula to develop a convolutional neural network (CNN) model. Second, we develop a rule-based filtering methodology to remove false-positive predictions made by our model. For instance, predicted poultry barns should follow the distribution of shapes of the labeled barns observed in the Delmarva Peninsula. Similarly, one of the main sources of concept drift is that poultry barns are oriented toward the wind (due to ventilation), and we use rotation augmentation to overcome significant orientation differences across states. The advantage of this approach is that it enables us to leverage the open and high-fidelity Delmarva data when ground truth segmented data can be expensive to acquire in new domains. We validate our model by comparing its predictions to a hand-labeled dataset of poultry CAFOs from ten counties across California, which were not used in the model development. We show that our model is able to obtain the recall of 87% and the precision of up to 83%.

Our contributions are fourfold:

1) a method for combining image segmentation with data augmentation and object based filtering to scale CAFO detection using a small set of ground truth labels;

Manuscript received 31 December 2021; revised 14 May 2022; accepted 9 July 2022. Date of publication 18 July 2022; date of current version 12 September 2022. *(Corresponding author: Caleb Robinson.)*

Caleb Robinson and Juan M. Lavista Ferres are with the Microsoft AI for Good Research Lab, Redmond, WA 98052 USA (e-mail: calebrob6@gmail.com; jlavista@microsoft.com).

Ben Chugg, Brandon Anderson, and Daniel E. Ho are with the Stanford RegLab, Stanford, CA 94305 USA (e-mail: benchugg@law.stanford.edu; branders@law.stanford.edu; dho@law.stanford.edu).

Digital Object Identifier 10.1109/JSTARS.2022.3191544

2) the first national open-source poultry CAFO dataset with facility locations and size estimates based on building footprints;
3) out-of-sample validation based on large-scale independent labeling by a trained team;
4) results that demonstrate the potential for this approach to detect CAFOs and uncover significant blind spots, for instance, in the CoA.

We make all the detected facilities, the model, and replication code available at https://github.com/microsoft/poultry-cafos/.

II. Related Work

Our work touches on several strands of prior research. First, there has been much success in applying CNNs to analyze high-resolution satellite imagery, for both object detection (see, e.g., [21] and [22]) tasks and semantic segmentation (see, e.g., [23] and [24]). Applications stemming from this progress have included ship detection [25], footprint extraction [26], and wildfire detection [27], among others.

As noted above, there has also been work specifically aimed at applying computer vision methods to CAFOs. Handan-Nader and Ho [20] employed a CNN with hand-labeled imagery collected by environmental interest groups over the course of years to automate the mapping of poultry and swine CAFOs across North Carolina. Their approach uses transfer learning with a pretrained network (Inception V3) to classify image tiles. Our work builds on this effort by moving from an image classification to a semantic segmentation framework and focusing on model generalization in order to scale our analysis to the entire U.S. In contrast with a classification approach, the semantic segmentation approach allows us to identify individual barns and, thus, calculate barn-level features (such as area). Our hypothesis is that an additional filtering step based on barn-level features will improve the performance and generalization ability of our approach (compared to an image classification approach) because industrial poultry production is relatively homogeneous regardless of geography.[1]

Maroney et al. [29] detect poultry operations in 35 counties across seven southeastern states, using the ArcGIS Feature Analyst [30], [31] on National Agricultural Imagery Program (NAIP) 1 m/pixel aerial imagery. Feature Analyst proceeds by: 1) having the user to manually select a training set; 2) using ensembled supervised learning (with a range of models, including neural networks, decision trees, and k-nearest neighbors); and 3) retraining after the user removes false positives. Patyk et al. [32] augment this ArcGIS-based detection with probabilistic estimates based on the CoA [33] to detect poultry operations in almost 600 U.S. counties.

These efforts are promising, and our work builds and improves on these efforts in several key ways. First, the modeling approach by Patyk et al. relies on data from the CoA. The CoA, however, is taken only once every five years, masks results in small counties, and is based principally on a survey that is affected by nonresponse and unknown farm locations [34]. We demonstrate that our approach identifies significant blind spots in the CoA, identifying counties with zero reported CAFOs that, in fact, have a large number of CAFOs. Second, the proprietary ArcGIS Feature Analyst system relies on manual inputs and provides little detail on the exact model deployed for a task; this black box nature makes replication, model verification, and improvements difficult. Third, while Patyk et al. estimate locations for nearly 600 counties, the resulting data are not publicly available. We, hence, provide CAFO locations for all 3000+ counties in the U.S.

III. Problem Formulation

We would like to perform an instance segmentation of remotely sensed imagery to identify individual poultry CAFO barns over the continental U.S. We break this problem in two steps:

1) a binary semantic segmentation of the remotely sensed imagery into a "barn" and "background" class;
2) a supervised grouping of contiguous pixels classified as "barn" into barn objects.

Formally, we are given a dataset of N labeled multispectral images and corresponding label masks, $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$. $X_i \in \mathbb{R}^{h \times w \times c}$ denotes the imagery, where h and w are the pixel height and width, respectively, and c is the number of spectral bands. $Y_i \in \{0,1\}^{h \times w}$ denotes the label, with 1 denoting a barn and 0 background.

Step 1 is a standard supervised learning problem, in which we aim to learn the parameter, θ, of a semantic segmentation model $f(X_i; \theta) = \hat{Y}_i$ that produces a probabilistic estimate of whether each pixel in an input image is part of the foreground "barn" class, $\hat{Y}_i \in [0,1]^{h \times w}$. Step 2 is less standard in machine learning and involves aggregating contiguous groups of predicted foreground pixels into *objects*, O, and then classifying each of these objects as a "barn" or not in terms of their *object-level features*. We engineer or extract d features per object (e.g., the object's area), $O_j \in \mathbb{R}^d$, which allows us to incorporate accessory data, besides the multispectral imagery, at the object level. The classification step can be done through hand-crafted rules, or, for example, by gathering *object-level labels*, Y_j^{object}, and training a classifier, $g(O_j; \phi) = \hat{Y}_j^{\text{object}}$.

While it is in principle possible to build an end-to-end model, we believe there are strong practical reasons to keep object-level detection as a separate step. First, manual rules can leverage expertise on CAFOs (e.g., dimensions, features, and geography). Such rules are transparent, easily scrutinized, and are readily modifiable by future research terms if needed. Second, encoding *object-level* features (e.g., predicted CAFO area and aspect ratios) would be less straightforward as these features depend on the output of the model. Third, the computational burden is extensive for end-to-end approaches. Keeping recall high at step 1 and inspection false positives at step 2 enables a rapid discovery of filters. This is particularly the case when our training data

[1]Poultry production in the U.S. operates under the paradigm of "vertical integration," meaning that there are a handful of large companies—*integrators*—across the country responsible for contracting out the majority of production to smaller farm operations. Integrators contractually require that production facilities adhere to specific standards. This results in structurally similar poultry barns across the country. See [28] for an overview.

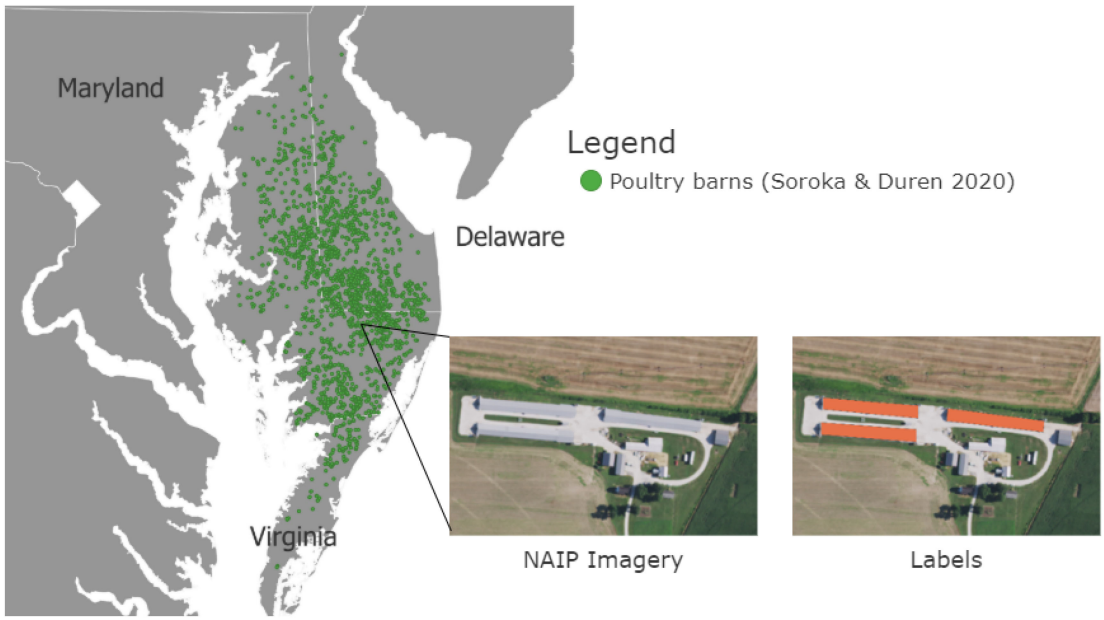


Fig. 1. Locations of poultry barns from the Delmarva dataset [5].

may not cover all the areas that reveal systematic false positives. That said, future work investigating an end-to-end approach would be valuable. Such approaches might 1) add an additional image channel for each additional feature, similarly to CNNs for structured data [35] or 2) add a separate network for the metadata, connecting its output with that of the CNN with additional layers (see, e.g., [36]).

In the following sections, we describe the datasets we use, the different modeling approaches for the two steps, and experiments to validate our modeling choices and final dataset.

IV. Data

In this article, we use aerial imagery from the U.S. Department of Agriculture's NAIP imagery, the Soroka and Duren dataset of poultry barn polygons from the Delmarva Peninsula [5], a dataset of CAFO facility polygons from an internal validation process, and road network data from OpenStreetMap (OSM).

A. NAIP Imagery

The NAIP imagery has a 0.6–1 m/pixel spatial resolution, RGB and near-infrared channels, and is collected on a state-by-state basis once every three years in the U.S. For consistency, we use the most recent 1 m imagery per state, as not all the states have transitioned to 0.6 m/pixel imagery. The resulting imagery dataset is substantial in size (>42 TB) and has the advantage of providing comprehensive coverage of the U.S. at high spatial resolution.

B. Delmarva Data

The Delmarva data were produced by the U.S. Geological Survey Chesapeake Bay Studies program to study disease transmission from migratory birds to farms. It consists of 6013 poultry barn polygons from NAIP 2016/2017 aerial imagery over the Delmarva Peninsula (containing portions of Virginia, Maryland, and Delaware) [5]. We augment these data using the methodology described by Robinson et al. [18] to create estimated construction dates back to 2010/2011 for each poultry barn. Fig. 1 depicts the region and instances of the Delmarva data.

C. Validation Data

In a separate project examining environmental and public health dimensions of CAFOs in California, we developed an active-learning-based approach to find CAFO facilities *of any type* or confirm where CAFO facilities exist. We scanned ten large agricultural counties in California, yielding a validation dataset of 8869 polygons. Each polygon is labeled as "empty" or with the type of CAFO as determined by expert annotators. In contrast to the Delmarva data, the polygon contains the entire facility, including areas that are not the barn. Appendix A provides more detail on the data labeling process. Because there are substantial changes in the landscape, surroundings, and adjacent facilities in different regions, we use these validation data as a tough test for our approach to generalize to an agricultural setting that is quite distinct from Delmarva.

D. OSM Data

Street data are particularly useful for our object-based filtering for two reasons. First, CAFOs require access to major roads for the distribution of livestock or livestock products. Second, roadways can themselves be visually similar to CAFO barns (see Appendix C), and road networks are, hence, valuable to filter out false positives. We download all road network location data in the U.S. from OSM using the `osmnx` library [37] and incorporate these data as we describe in the following.

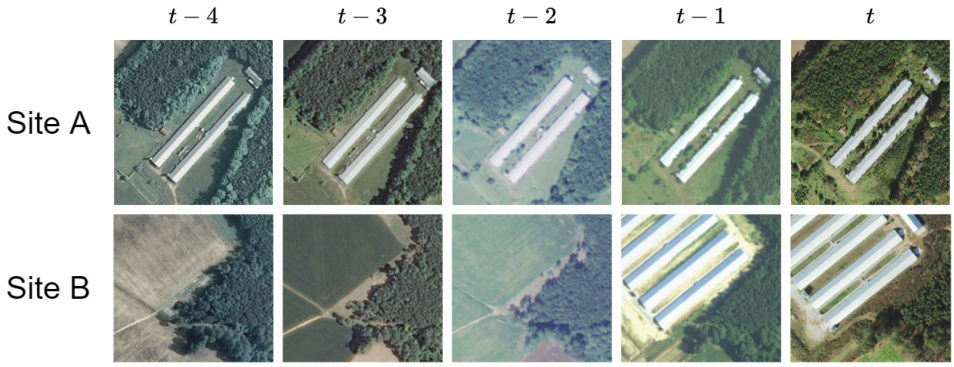


Fig. 2. Example of our *temporal augmentation* method. We have labeled data (i.e., pixel masks) and aerial imagery for poultry barns at time t; however, we also have historical aerial imagery and know many barns will have existed before t. With our *single* training strategy, we use imagery and masks from just t, with *all* we use imagery from all time points and assume that the same masks apply, and with *augmented*, we use an unsupervised method to determine the previous points in time that the masks are valid for. At "Site A," both *all* and *augmented* will allow us to train with four additional labeled samples in different imaging conditions. However, at "Site B," the *all* strategy will introduce noisy labels, while the *augmented* strategy will correctly identify that only the imagery at $t-1$ contains a valid mask for training.

V. Methods

A. Supervised Model Training

We implement Step 1 from the problem formulation (see Section III) by training U-Net models [38], to segment NAIP imagery into "barn" and "background" classes. Despite the success of some image classification approaches on aerial imagery [39], [40], [41], we chose an image segmentation approach because it enables easier rule-based filtering (see Step 2). We do not pursue few-shot segmentation (see, e.g., [42]) because we have sufficient training data. We use a ResNet-18 encoder as implemented in the `segmentation-models-pytorch` package [43]. We split the Delmarva Peninsula spatially into a northern *training area* and a southern *testing area* (roughly 75% and 25% of the total area). We further use a 7×6 km^2 area from the training area as a validation set. We do not use a fixed dataset of pre-extracted patches during training, but instead sample patches of size 256×256 pixels from imagery covering the training area. As only 0.1% of the training area contains positive "barn" masks,[2] we weight the sampling procedure by randomly discarding a patch with probability α if it does not contain an instance of a positive label. Intuitively, lower values of α result in training on more background pixels, while higher values of α result in training on a more balanced dataset. A value of $\alpha \approx 0.999$ would correspond to a class-balanced dataset.

Given that we only have labeled data from the Delmarva Peninsula—but would like to apply a trained model over the entire U.S.—our main concern is that our models will not be exposed to sufficient variation in the data during training and will, thus, fail to generalize when applied to imagery throughout the country. While poultry barns are relatively standard in appearance regardless of location, their relationship with their surroundings and their orientations can and do vary with location. Furthermore, the NAIP imagery itself varies considerably across states due differences in the time of year, day, and environmental conditions of image capture. For example, Fig. 2 shows NAIP imagery for two locations at five points in time—there are considerable nonsemantic differences between the images due to the different imaging conditions.

Considering this, we train the models with random rotation augmentation (and random horizontal/vertical flips) to account for shifts in the distribution of poultry barns' orientations. More generally, training with rotation augmentation, or using rotation equivariant networks [44], [45], is a necessary component of modeling pipelines that use remotely sensed imagery as objects in remotely sensed imagery can be observed in any orientation.

1) Temporal Augmentation: We also train with *temporal augmentation* to account for differences in NAIP imagery. This method involves pairing multiple years of NAIP imagery over the training area with a single layer of labels. For example, the Soroka and Duren dataset was created with NAIP imagery from 2016 and 2017 (depending on the state), so without temporal augmentation, we would train on a dataset consisting simply of pairs (X^t, Y^t), where t represents the 2016/2017 layer. However, we also have NAIP imagery of the Delmarva Peninsula at the same resolution dating back to 2010. With temporal augmentation, we can use the fact that many labels will not change over time and augment our training set with samples of the form (X^{t-k}, Y^t) for offsets, k, that point to valid years of NAIP imagery. While this is a strong assumption, it allows us to expose the model to more variance in the NAIP imagery at the cost of introducing label noise during training. We further use the unsupervised method described in [18] to estimate the construction dates of each poultry barn to avoid augmenting with imagery samples before the construction of the labeled barn. Specifically, we create three versions of our training data, corresponding to different levels of temporal augmentation:

1) *single*, which only uses the 2016/2017 Delmarva labels paired with the corresponding NAIP 2016/2017 imagery;
2) *all*, which pairs the Delmarva labels with all valid NAIP imagery back to 2010;

[2]The training area covers ~ 14.091 km^2, while the area of poultry barns in the training area is 13.190 km^2.

TABLE I
SIZE OF THE DIFFERENT TRAINING SPLITS (DESCRIBED IN SECTION V-A), THE TESTING SPLIT, AND THE FULL U.S. IMAGERY (USED IN SECTION VI-C)

Split	Pixels	Tiles	Labeled Barns
single	∼18 billion	397	5 280
augment	∼124 billion	1 983	24 867
all	∼124 billion	1 983	26 400
test	∼5 billion	114	733
US-wide imagery	∼13 trillion	212,354	–

While the contiguous U.S. is ~ 8 trillion m^2, the size of the 1 m NAIP imagery over the U.S. is ~ 13 trillion pixels (at a 1 m/pixel resolution) due to overlap between adjacent tiles of Imagery.

3) *augmented*, which uses estimated construction dates from an unsupervised model to pair the Delmarva labels with appropriate imagery (i.e., in which the labeled poultry barns existed in) back to 2010.

See Fig. 2 for an example of the samples selected by these three approaches. We apply each of these methods to create three training datasets, but keep the test labels fixed to their 2016/2017 imagery. See Table I for the size of each set.

We train all the models with a pixelwise cross entropy loss, an initial learning rate of 0.01 using the AdamW optimizer [46], and learning rate decay that drops the learning rate by a factor of 10 on training loss plateaus. We compare models trained: with and without filtering (described in the following section); with values of $\alpha \in \{0.05, 0.1, 0.5\}$; with the three different measures of temporal augmentation (single, all, augmented); and with and without rotation augmentation. In all the cases, we perform model selection based on validation set performance.

B. Object-Based Filtering

The second step of our approach aims to incorporate data, such as a vector road map, that is not possible to directly use in the semantic segmentation model from the first step. Here, we group the per-pixel outputs of the semantic segmentation model to create *objects* and then classify these objects as "barns" or not based on object-level features. Specifically, given the probabilistic output of a semantic segmentation model over a large area, $\hat{Y}$, we first threshold the output to create a binary "barn" prediction per pixel. The threshold value used can be tuned over a held-out validation set to achieve a desired point on the precision/recall curve. Next, we group contiguous sets of pixels predicted as the positive "barn" class into objects (or polygons in geographic space) using a polygonize operation[3] with a 4-pixel neighborhood. We then compute the following set of features for each object/polygon.

1) *Area:* We compute the minimum rotated rectangle [48] that fits the group of pixels and record the area of this rectangle as the area of the predicted "barn." A minimum rotated rectangle for a given polygon is the smallest rectangle with at least one edge coincident to an edge from the polygon. Such a rectangle can be determined in $O(n)$ time, where n is the number of vertices in the given polygon [49].
2) *Aspect ratio:* We again compute the minimum rotated rectangle and record the ratio of the length of longer side of the rectangle to the length of the shorter side.
3) *Road distance:* We compute the distance from each object to the nearest road in the OSM database of public and private roads over the U.S.[4]

We use these features to create a rule-based classifier using the *range* of the same features computed over the Soroka and Duren polygons (see Fig. 3). Any object that has an area outside of the range $[525, 8106]$ m^2, an aspect ratio outside of the range $[3.4, 20.49]$, or is immediately intersecting with a road in the OSM data is classified as background. In contrast, any object that fits these criteria is classified as a poultry barn. Another way of interpreting this method is a filtering step that aims to discard groups of false-positive predictions made by the segmentation model. The segmentation model naively operates on image-based data with pixel-level supervision without a direct way to incorporate higher order labels such as the distribution of aspect ratios of known poultry barns or road network vector data. These features are useful in modeling, and while we used a rule-based classifier with three features in this article, the classifier could be learned using labeled data over a more in-depth object-based feature representation.

VI. EXPERIMENTS AND RESULTS

A. Evaluation

To evaluate a model, we run the model over the entire test area and then group contiguous sets of predicted poultry barn pixels into objects. We consider an object as a true positive prediction if it has greater than a 50% intersection over union (IoU) with a labeled barn. Similarly, an object is a false positive if it does not have a 50% IoU with a labeled barn, and a labeled barn is a false negative if no predicted object is counted as a true positive with it.

We evaluate performance using precision, recall, and the F_2 score. We choose the F_2 metric to more heavily weight recall in the model evaluation as it is possible to reduce the number of false positives in postprocessing; however, it is not possible to reduce the number of false negatives (i.e., find new objects). This point broadly applies to object detection models being run over large amounts of satellite imagery—there is a much higher cost of missing a potential object of interest than that of making a false-positive prediction.

B. Results

Table II summarizes our main results for model variants trained with rotation augmentation. The table presents performance measures based on the model's raw predictions (unfiltered columns) and predictions after applying our proposed object-based filtering method (filtered columns), varying α, and temporal augmentation in rows. Results are sorted by the F_2 score of filtered models.

[3]A search is performed over predicted pixels to group each contiguous set of pixels into a polygon using the `rasterio` package [47].

[4]For more details about performing this computation at a national scale, see Appendix B.

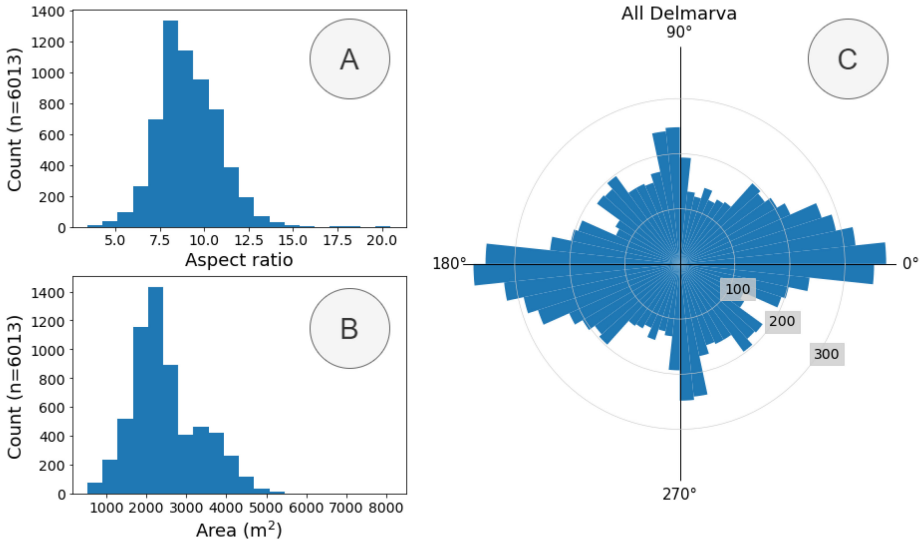


Fig. 3. Distributions of properties of the poultry barns in the Delmarva dataset. (a) Distribution of the aspect ratio of poultry barns (ratio of the length of the long side of the barn to the length of the short side of the barn). (b) Distribution of the area of poultry barns. (c) Distribution of the orientations of poultry barns.

TABLE II
TEST RESULTS OF MODELS TRAINED WITH ROTATION AUGMENTATION

		Unfiltered			Filtered		
Training set	α	Precision	Recall	F_2	Precision	Recall	F_2
Augmented	0.05	27.35%	97.25%	64.36%	87.05%	94.68%	93.05%
All	0.1	45.64%	96.78%	79.07%	89.94%	93.73%	92.94%
Augmented	0.5	48.90%	96.04%	80.51%	92.01%	93.13%	92.91%
Augmented	0.1	37.74%	96.70%	73.68%	88.51%	93.64%	92.57%
All	0.05	32.56%	96.32%	69.21%	85.57%	93.73%	91.98%
All	0.5	36.06%	96.61%	72.32%	87.57%	92.45%	91.43%
Single	0.1	15.95%	96.88%	48.09%	80.60%	93.44%	90.55%
Single	0.05	8.64%	98.08%	31.93%	68.54%	93.71%	87.30%
Single	0.5	43.14%	87.71%	72.69%	86.62%	84.30%	84.75%

Results are displayed by: 1) whether or not object-level filtering is applied (left and right panels); 2) the level of temporal augmentation (*single, all, and augmented*); and 3) α weight in the training data. Rows are sorted by the filtered F_2 score. Augmentation and α values affect only the training set, and evaluation is conducted on the natural distribution of the test set.

1) Impact of Filtering: We find that the filtering step is highly effective based on rules derived from the training set statistics. On average (over the rows of Table II), recall decreases by 3.29, while precision increases by 52.27. This underscores the substantial utility of expert-based rules in complementing CNN-type approaches.

2) Impact of Temporal Augmentation: We also observe that the U-Net trained with temporal augmentation and an α value of 0.05 performs the best in terms of the filtered F_2 score. Overall, "augmented" data with lower values of α, corresponding to including a higher proportion of background imagery, result in better performance. We also observed in our training process that methods trained with *some* type of temporal augmentation appear to generalize better than those without temporal augmentation—despite potentially introducing noise into the training process.

3) Impact of Rotation Augmentation: We observe a distinct rotation bias in our labeled dataset, where most barns have an East–West orientation (panel C from Fig. 3). The reason for this rotation bias lies in ventilation. Among growers, one piece of received wisdom is that "longest side of your pen, that is, the breadth, should face the prevailing direction of the wind."[5] Of course, wind direction can differ dramatically across regions, so the ability to generalize will be affected by this rotation. Because our training and testing data are from the same area, we do not observe any significant differences between models trained with and without rotation augmentation on our fixed test set. To demonstrate the positive effect of rotation augmentation, we conduct a further experiment with test-time augmentation, where we crop a patch of imagery around each poultry barn in the test set, and then test each model on all possible 45° rotations of these images. The average difference in recall between models trained with and without rotation augmentation

[5][Online]. Available: https://www.justagric.com/poultry-house-construction-guidelines/

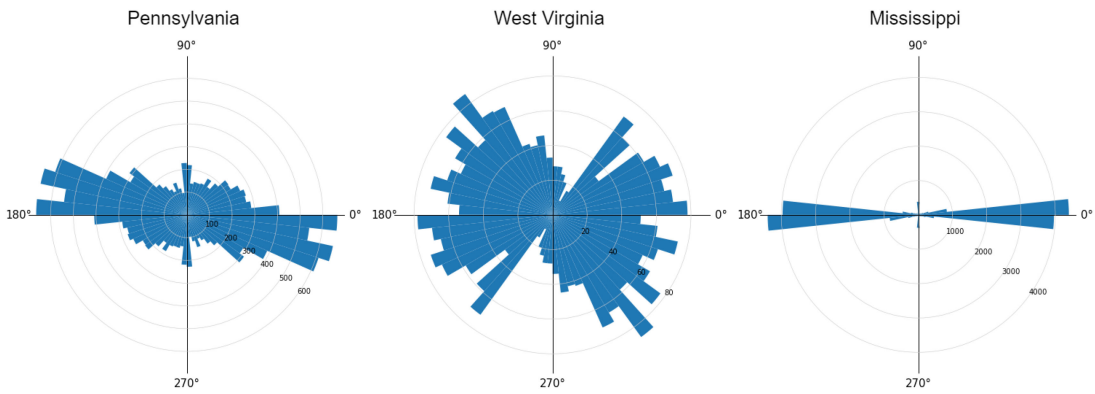


Fig. 4. Distributions of orientations of *predicted* poultry barns by state. (Left) Pennsylvania. (Middle) West Virginia. (Right) Mississippi.

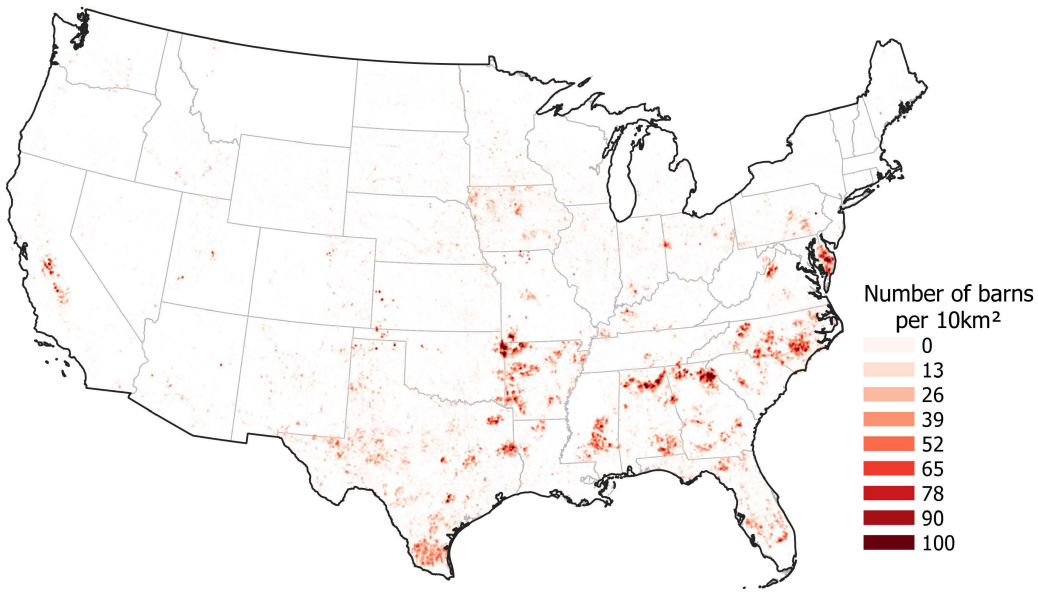


Fig. 5. Density rendering of our filtered version of predicted poultry barn locations over the U.S. colors range from 0 barns in a 10-km^2 radius (transparent) to 100 barns in a 10-km^2 radius (dark red).

is 22.32, while the best model with rotation augmentation shows a 27.47 improvement in recall over its nonrotation augmented counterpart. In other words, if poultry barns are oriented randomly over a landscape, the models with rotation augmentation will identify 22.32% more poultry barn pixels than the models without rotation augmentation. As expected, we find that other states in the U.S. (outside of the Delmarva Peninsula) have different distributions of orientations. For example, Fig. 4 shows the distribution of orientations of predicted poultry barns in Pennsylvania, West Virginia, and Mississippi.

C. Out-of-Sample Inference and Validation

1) Inference: We create a U.S.-wide map of predicted poultry CAFO locations by applying the best-performing model over 42 TB of the most recent 1 m/pixel NAIP imagery from each state in the continental U.S. This computation was run on Microsoft Azure, specifically with an NC24v3 virtual machine[6] that was located in the same cloud region as the NAIP imagery. The NAIP imagery is broken up into 212 354 *tiles*, each of which is $\sim 8500 \times 7000$ pixels, for a total of ~ 13 trillion pixels. We create a grid of 256×256 patches with 64 pixels of overlap between neighboring patches from each tile of NAIP imagery, run the model to compute the per-pixel class probability estimates for each patch of imagery, and then stitch together the resulting predictions while averaging the class predictions in overlapping areas. The process of averaging predictions along the edges of tiles serves to reduce any border artifacts caused by zero padding within the model [50]. We performed this process $4\times$ in parallel (one process for each GPU on the virtual machine) for a total run time of approximately 2.5 days. This results in 7 108 719 predicted barn polygons before filtering and 360 857 predicted barn polygons after the filtering step. A density rendering of the predictions is shown in Fig. 5.

2) Out-of-Sample Validation: We compare our final filtered set of poultry CAFO barn predictions to the out-of-sample

[6]The NC24v3 virtual machine type contains four NVIDIA V100 GPUs.

TABLE III
ABLATION RESULTS

Model	Unfiltered			Filtered		
	Precision	Recall	F_2	Precision	Recall	F_2
U-Net	27.35%	97.25%	64.36%	87.05%	94.68%	93.05%
U-Net++	did not converge					
DeepLabV3+	44.49%	97.72%	78.85%	86.79%	94.79%	93.07%
MA-Net	50.64%	97.36%	82.19%	91.13%	94.17%	93.54%

Test results of different models trained with rotation augmentation, the *augmented* training set, and $\alpha = 0.05$. Results are displayed in a similar manner to Table II.

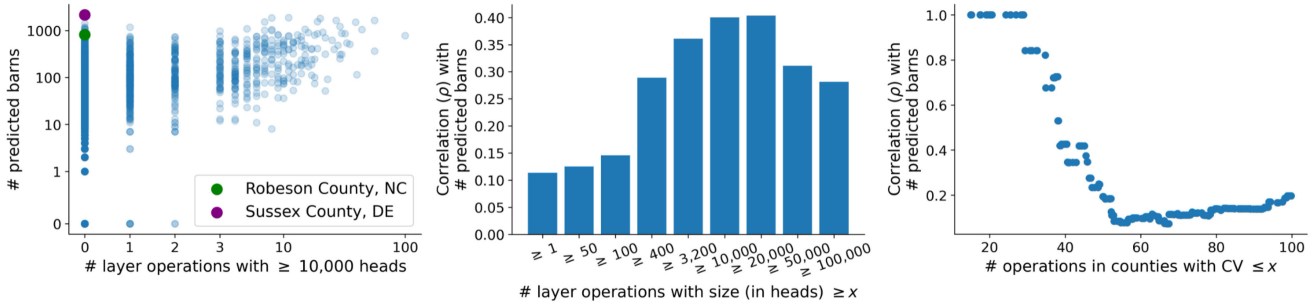


Fig. 6. Comparison of our U.S.-wide model predictions with the CoA data at a county level. (Left) Scatter plot showing the number of poultry layer operations with $\geq$ 10 000 heads reported by the CoA to the number of predicted barns per county. We highlight Sussex and Robeson counties as extreme examples, where the CoA is missing data (see Fig. 7). (Middle) Spearman's correlation coefficient between the number of poultry layer operations over different size categories and our number of predicted barns per county. The correlation is highest for the subset of operations with $\geq$ 10 000 heads (i.e., the plot shown in the left panel) indicating that our model is picking up on larger operations. (Right) Spearman's correlation coefficient between the number of poultry operations from subsets of counties with increasing coefficients of variation (CV) as reported by the CoA and our number of predicted barns per county. Our predicted number of barns are highly correlated with the CoA data over the subsets of counties that the CoA data have the most certainty in.

validation dataset described in Section IV and Appendix A. This dataset covers ten counties in California and consists of facility-level annotations (not barn polygons) for *all* types of CAFOs, as well as "empty" annotations for polygons that have been confirmed to not have any CAFOs in them.

We count true positives as predictions that are within 100 m of a validated poultry facility and false positives as predictions that are not within 100 m of any validated poultry facility. For computing recall, we evaluate predictions within the validated area, where false negatives are validated poultry facilities that do not have a prediction within 100 m. We apply this methodology over all the ten counties and achieve a recall of 86.90% and a precision of 83.02%.

D. Comparison With Other Methods

We run an ablation study on model architecture, comparing the original U-Net architecture we use to more modern U-Net++ [38], DeepLabV3+ [51], and MA-Net [52] architectures. In this experiment, we use identical hyperparameters to the best-performing model from Table II—training with the augmented training set, setting $\alpha = 0.05$, the same decaying learning rate schedule, etc. From Table III, we observe that the DeepLabV3+ architecture results in very similar performance to the U-Net, while the MA-Net architecture has significantly improved precision and slightly improved recall—even after filtering. This suggests that the MA-Net approach is able to eliminate some classes of false positives that are *not* captured by the filtering step and represents a promising direction for improving performance on the task of detecting poultry barns. The U-Net++ approach does not converge under the same training parameters.

We also compare to the state-of-the-art (open-source) results in detecting CAFOs from Handan-Nader and Ho [20], who use a tile classification approach to detect CAFOs. We ran their pretrained model in the Delmarva Peninsula on images centered on existing poultry barns and obtained a recall of 93.72% and a precision of 71.16%. Note that this is a drop of 16 percentage points in precision compared to our model. Moreover, their model's performance drops significantly when images are not centered around the facility (but instead tiled over the entire study area) with recall dropping to 54%, almost 40 percentage points lower than our method. Fig. 6 also compares our method with the status quo, i.e., the CoA. There we find several counties whose numbers in the CoA differ drastically with our predictions.

VII. IMPLICATIONS

Our study has substantial implications for environmental research and governance. First, we have generated and released the first open-source national map of poultry CAFOs, which can be used for environmental research, monitoring, and enforcement.

Fig. 7. Example model predictions (highlighted in blue) from two counties that the CoA reports zero poultry operations in. (Left) Twenty-four predicted poultry barns with NAIP 2017 imagery over a 2.6-km^2 area from Sussex County, Delaware. (Right) Ten predicted poultry barns with NAIP 2016 imagery over a 3.1-km^2 area from Robeson County, North Carolina.

Fig. 5 shows a heatmap of the filtered version of the predicted poultry barn locations over the U.S. This dataset includes a polygon for each predicted poultry barn with attributes including the properties used in the filtering step, the average modeled probability over the pixels in the polygon, and the timestamp of the imagery that the predictions were generated from. We have released both the sets of filtered and unfiltered predictions for future work to build on.

Such a map has a myriad of potential use cases. Epidemiological researchers can examine the impact of livestock agriculture of disease spread [5]. Environmental interest groups can use such information to disseminate and monitor specific facilities [53]. In 2014, Earthjustice, for instance, filed a petition with EPA about disparate impact of CAFO permitting in North Carolina [54]. Government agencies can use such information to prioritize inspections and understand permitting failures [3], [55]. For example, the EPA estimated that nearly 60% of CAFOs do not hold permits [56]. The national map also enables risk scoring facilities based on the proximity to waterways, vulnerable communities, and other potential environmental impacts.

Our approach can also significantly reduce the time currently spent on manually scanning for such facilities by providing a set of likely poultry barns with high recall. The U.S. contains over 4 million km^2 of agricultural and pastoral land [57]. While constructing the out-of-sample validation set, we found the task of CAFO identification to be nontrivial even to a human validator, requiring 90 s on average for specialists at a third-party company to label a square kilometer image. At this rate, we can estimate that it would require approximately 100 000 h of human effort to cover the U.S. For states and environmental groups, humans scanning satellite imagery took three or more years to complete a single state [20].

Second, our results illustrate major limitations in the CoA, due to coverage gaps, high nonresponse rates, and, even, internally contradictory data [34], [58]. The left panel of Fig. 6 plots the detected number of barns by our model per county against the number of operations with over 10 000 heads of layer chickens reported the CoA. While the two are indeed correlated (Spearman's rank-order correlation coefficient is 0.38), there are many deviations, particularly in counties where the CoA reports no poultry operations at all. For instance, the CoA lists both Robeson County, North Carolina and Sussex County, Delaware, as having no poultry operations with over 400 heads. In both the counties, however, our model detected thousands of poultry barns, which we have validated through the visual inspection of the NAIP imagery (see Fig. 7 for examples from both the counties). Both the counties also have been previously reported as containing large CAFOs [59], [60], [61].

The middle panel of Fig. 6 plots the Spearman's rank-order correlation coefficient between the number of barns predicted by our model per county against counts from the CoA based on different poultry operation size thresholds. This shows that the two approaches are most highly correlated with mid- to large-sized CAFOs. Finally, the right panel compares the number of barns predicted by our model to the CoA count of total operations for subsets of counties based on different uncertainty thresholds reported by the CoA. The CoA reports a coefficient of variation (CV) for each point estimate at the county and state levels, where smaller values represent less uncertainty, with attempts to adjust for the undercount of facilities [62]. We collect the subset of counties corresponding to increasing CV value cutoffs and find that our predictions are highly correlated with the CoA operation counts for counties that the CoA has less uncertainty.

These results show that prior methods to map CAFOs that rely on the CoA may inherit the same coverage gaps. Most promisingly, our approach provides an independent enumeration of CAFOs, without relying on self-reported information by producers. This study also illustrates how advances in computer vision can supplement areas where administrative data have fallen short [63], [64].

Third, our results illustrate advantages to a hybrid approach that leverages the segmentation of regular objects (barns) and expert-based heuristics. Earlier approaches that classify image tiles grapple with centering the CAFO facility in the image and may be sensitive to changes in the background, which could change significantly across U.S. regions (e.g., arid versus humid landscapes). For instance, the image classification model by Handan-Nader and Ho results in recall of 93.72% and 71.16% precision in Delmarva. This is a drop in 16 percentage points in precision. But that model's performance drops significantly when images are not centered around the facility, with recall dropping to 54%. This suggests that image segmentation is a far more efficient by requiring inference on fewer overlapping image tiles.

In short, focusing on the key object of the distinct poultry barn and combining expert-based heuristics (e.g., aspect ratio of barns, proximity to road for distribution, and wind orientation of barns) can enable efficient learning with much lower costs in labeling and inference.

VIII. Limitations

While this study has made a major advance toward an open-source national poultry CAFO map, we spell out several limitations and opportunities for further work.

First, the precision of our results varies by geography. From a qualitative error analysis, the most common class of false positive are predictions on roadways that are not covered by OSM data and predictions on other linear features that are not included in our current filtering step (e.g., railways). See Appendix C for examples. A future avenue of improvement could involve obtaining more comprehensive road data to increase the efficiency of our proximity filter. A tradeoff with this approach, however, is that more comprehensive datasets (e.g., datasets from ESRI or Google Maps) are not open source. Another alternative would be for researchers to tailor the filtering step to improve the precision based on local knowledge. Future work might also treat the filtering step as an independent learning problem, in which we attempt to learn the distribution of false positives and correct for it.

Second, while our out-of-sample validation yields strong results, it is not possible to calculate the precision where California data were not labeled. Since those labels resulted from an active learning model to ensure 80% recall of facilities—i.e., human labelers did not visually inspect 100% of each county—it is possible that our precision is biased. If a predicted polygon is outside of the validated area, we cannot tell whether it is a false-positive or a true-positive prediction, as it could be a poultry facility that was missed in the validation data. If we assume that *all* predictions outside of the validated areas are false positives, this yields an extreme lower bound on precision of 18.84%. The large discrepancy between precision based on overlapping imagery and this lower bound suggests that either our model has a high false-positive rate and/or there are missing facilities in the validation data. Qualitatively, we find a mix of both the effects, with the model making false-positive predictions in desert areas around San Diego, but also identifying likely poultry facilities missed in the validation data. These results suggest that combining sources information will be most promising for a full enumeration of CAFOs.

Third, our approach has focused specifically on poultry CAFOs and has not examined cattle or hog CAFOs. The reason is that poultry CAFOs are visually distinguishable by the barn, whereas distinguishing hog facilities requires the identification of the manure storage system, and cattle CAFOs can rely extensively on outdoor feedlots. A next natural step would be to extend our approach to hog facilities that typically have, for instance, different aspect ratios.

Fourth, while we demonstrate substantial gains from object-based filtering, there may be other object-level features that can help improve performance. The main challenge here lies in whether such data are comprehensively available across the continental U.S. and validating filters based on such data. For instance, in an earlier iteration, our team attempted to cropland layer data, but, surprisingly, found that such data did not significantly help in classifying CAFOs. Advances in weak supervision may help to improve the development of such heuristics [65].

Finally, we note that like all machine learning models, we are constrained by our ground truth data. Consequently, it is unclear how robust our method will be to changes in the imagery over time or to changes in CAFO construction standards. While we have applied temporal augmentation to our training data (see Section V) to combat some of these temporal effects in past, we are not immune from such changes in the future. In particular, it may be harder to directly apply our model to time series of NAIP imagery, as the resolution itself changes, and applicability to more real-time (higher cadence) imagery will require adaptation to lower spatial resolution (see, e.g., [19]).

Notwithstanding these limitations, our findings show that this hybrid approach makes a substantial advancement in the large-scale detection of CAFOs in a transparent open-source fashion.

IX. Conclusion

In this article, we provided the first freely available dataset and corresponding open-source model aimed at locating poultry CAFOs across the continental U.S. We hope that this work provides an initial step toward improving the regulatory capacity of federal and state environmental agencies—filling blind spots in the CoA and permit records from state authorities—as well as the research capacity of academics and the ability of environmental interest groups and the public to monitor these consequential facilities.

Appendix A
Validation Data

Using a YOLOv3 model [66] built for a separate project, the inference was run over California to detect CAFOs of any type (swine, dairy, and poultry). The predictions were then validated using a team of trained annotators (Stanford undergraduate students), employing an active learning approach based on upper confidence bound (UCB) sampling [67] to reduce the necessary number of images to check.

Each detected facility is assigned a score in $[1, \infty)$, representing the model's confidence in the prediction. We partition the space $[1, \infty)$ into K discrete buckets $B_1, \ldots, B_K$, each bucket B_i associated with an interval $[a_i, b_i)$ where $\cup_i [a_i, b_i) = [1, \infty)$. Bucket B_i contains those images containing a facility with a score in the interval $[a_i, b_i)$. If an image contains multiple detected facilities, it is placed in the bucket associated with its highest scoring facility. We add an additional bucket B_0 containing those images with no detected facilities.

For each bucket B_i, let μ_i denote the fraction of examined images in B_i containing a facility, and let n_i indicate the number of times we have validated an image from this bucket. The UCB score associated with bucket B_i is then

$$S_i = \mu_i + \alpha \left(\frac{\log \sum_j n_j}{n_i} \right)^{1/2}$$

where $\alpha \in \mathbb{R}$ is an "exploration parameter." The UCB score reflects the past success of the bucket at containing positive images, but by favoring those buckets that are visited less frequently, the second term ensures that we do not focus exclusively on buckets with the largest percentage of past successes. As $\alpha \to \infty$, we rely less on past success and more on a uniform selection of the buckets. To decide which images to sample, we form a probability distribution π over the buckets in the natural way

$$\pi_i = \frac{S_i}{\sum_j S_j}.$$

We sample from π with replacement m times, where m is the number of images validated in that round, and then update the scores for the subsequent rounds. For the first round, we take π as the uniform distribution.

This process is repeated in each country until it is estimated that we have found 80% of the facilities. The estimate is constructed by extrapolating success rates in each bucket over all the images. That is, the total number of estimated CAFOs is

$$N_C = \sum_{i=0}^{K} |B_i| \pi_i.$$

Once we have identified at least $0.8 \cdot N_C$, we stop the process and move onto another county. The validation data used in this article employed this process in ten counties across California. Note that the process does not result in every image being examined.

Appendix B
Computing Road Distance Features at Scale

Across the U.S., the OSM road network contains ~ 37.8 million edges,[7] and our model initially predicted 7.1 million poultry barn polygons. Calculating the distance between each polygon and the closest edge in the road network is, thus, nontrivial. First, we break the problem up such that it can be solved in parallel by splitting the road network and predictions across the 212 354 NAIP tiles and only considering nearest matches within a tile (we download the road network aligned to these tile definitions using the `osmnx` library [37], and our model predictions are saved at the tile level). Now, we need to compute the shortest distance between N polygons and M lines per tile; however, in urban areas, both M and N will be prohibitively large (as our model's false positives are often correlated with roads and white buildings) for a naive approach. Our approach is as follows.

1) Split each road edge into pieces of at most length d. If an edge of length D contained two nodes previously, after this splitting step, it would contain $2 + \lfloor D/d \rfloor$ nodes.
2) Add all the nodes from the now split road network into a K-dimensional tree data structure. This data structure offers logarithmic time lookup of the r nearest neighbors of a query point.
3) For each predicted polygon, query the K-dimensional tree with the polygon's centroid for all the neighbors up to a distance of $2d$ away. The set of points returned are guaranteed to include the nodes from the closest line.
4) Compute the polygon-line distance between each predicted polygon and lines corresponding to the nearest points returned in the previous step. This distance will be the minimum distance.

The road splitting step in this algorithm is necessary to calculate the nearest road line for each given polygon. For example, if we try to find the nearest line for a given polygon based on the endpoints of the line, then we can easily miss long straight roads that pass near to the given polygon. This approach is implemented in the accompanying repository and can process road network data and predictions corresponding to a given tile in seconds.

Appendix C
Examples of False Positives

Fig. 8 shows different types of false-positive predictions and a true-negative prediction from across the entire U.S. Road false positives are the most common and occur where OSM data are incomplete (e.g., in rural areas or on informal dirt roads), which breaks the filtering step, or where the OSM roads and the NAIP imagery we use are not aligned. In nonroad cases, we observe that the arrangement of the false positives usually mimics a plausible arrangement of barns in a poultry facility (e.g., the pair of false positives shown in the right panel of Fig. 8). These cases often happen in nonurban contexts, where poultry facilities will be found. We find that the model consistently does *not* make false-positive predictions in more urban settings (e.g., with airplane hangars shown in the bottom right subfigure of Fig. 8), suggesting that the model takes the spatial context of an area into account while making predictions.

Acknowledgment

The authors would like to thank Microsoft Azure for support in cloud computing, and Schmidt Futures, Stanford Impact Labs, the Chicago Community Trust, and Sarena Snider (Snider Foundation) for research support.

[7] The number of "highway" tags reported by https://taginfo.geofabrik.de/ web service.

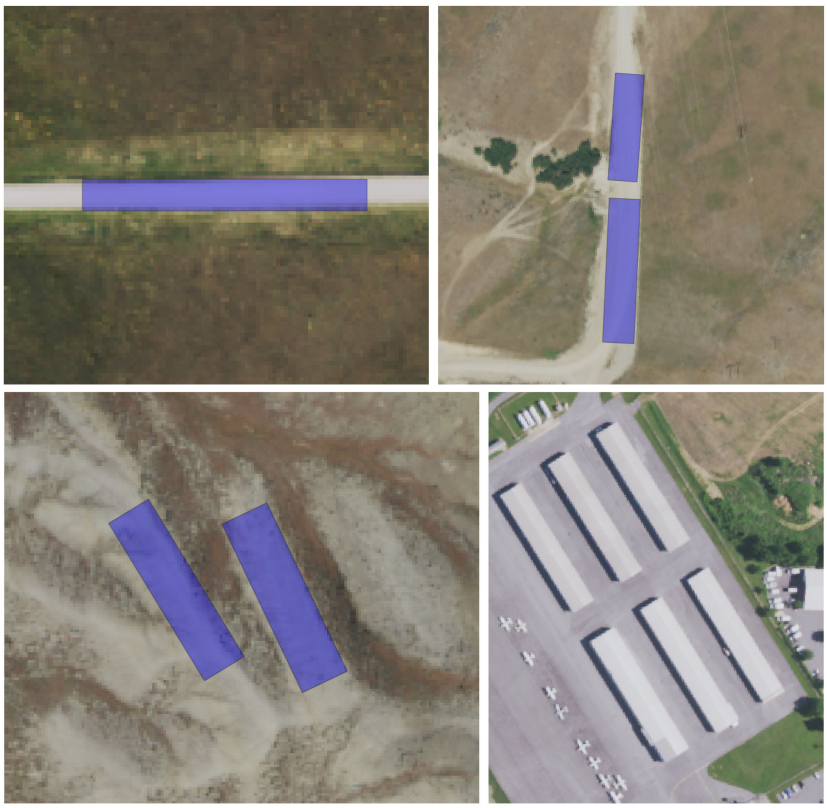


Fig. 8. Examples of false positives and true negatives. (Top left) False-positive prediction along a section of white, straight road in rural South Dakota. (Top right) False-positive predictions on dirt roads in Utah. (Bottom left) False-positive predictions on barren ground in Wyoming. (Bottom right) True-negative predictions for airplane hangars in Baltimore, Maryland.

REFERENCES

[1] D. Gurian-Sherman, *CAFOs Uncovered: The Untold Costs of Confined Animal Feeding Operations*. Cambridge, MA, USA: Union of Concerned Scientists, 2008.

[2] C. Hribar, "Understanding concentrated animal feeding operations and their impact on communities," Nat. Assoc. Local Boards Health, Washington, DC, USA, Tech. Rep. 59792, 2010.

[3] C. Handan-Nader, D. E. Ho, and L. Y. Liu, "Deep learning with satellite imagery to enhance environmental enforcement," in *Data Science Applied to Sustainability Analysis*. Amsterdam, The Netherlands: Elsevier, 2021, pp. 205–228.

[4] *Concentrated Animal Feeding Operations: EPA Needs More Information and a Clearly Defined Strategy to Protect Air and Water Quality From Pollutants of Concern*, US Government Accountability Office, Washington, DC, USA, 2008.

[5] A. Soroka and Z. Duren, "Poultry feeding operations on the Delaware, Maryland, and Virginia Peninsula from 2016 to 2017: U.S. Geological Survey data release," 2020. [Online]. Available: https://doi.org/10.5066/P9MO25Z7

[6] J. Burkholder et al., "Impacts of waste from concentrated animal feeding operations on water quality," *Environ. Health Perspectives*, vol. 115, no. 2, pp. 308–312, 2007.

[7] C. P. Gerba and J. E. Smith, "Sources of pathogenic microorganisms and their fate during land application of wastes," *J. Environ. Qual.*, vol. 34, no. 1, pp. 42–48, 2005.

[8] E. R. Campagnolo et al., "Antimicrobial residues in animal waste and water resources proximal to large-scale swine and poultry feeding operations," *Sci. Total Environ.*, vol. 299, no. 1–3, pp. 89–95, 2002.

[9] J. C. Barker and J. P. Zublena, *Livestock Manure Nutrient Assessment in North Carolina*, Raleigh, NC, USA: North Carolina Agricultural Extension Service, North Carolina State Univ., 1995.

[10] D. R. Raman et al., "Estrogen content of dairy and swine wastes," *Environ. Sci. Technol.*, vol. 38, no. 13, pp. 3567–3573, 2004.

[11] J. D. Kaplan, R. C. Johansson, and M. Peters, "The manure hits the land: Economic and environmental implications when land application of nutrients is constrained," *Amer. J. Agricultural Econ.*, vol. 86, no. 3, pp. 688–700, 2004.

[12] M. A. Mallin, "Impacts of industrial animal production on rivers and estuaries," *Amer. Scientist*, vol. 88, no. 1, pp. 26–37, 2000.

[13] A. A. Schultz, P. Peppard, R. E. Gangnon, and K. M. Malecki, "Residential proximity to concentrated animal feeding operations and allergic and respiratory disease," *Environ. Int.*, vol. 130, 2019, Art. no. 104911.

[14] S. T. Sigurdarson and J. N. Kline, "School proximity to concentrated animal feeding operations and prevalence of asthma in students," *Chest*, vol. 129, no. 6, pp. 1486–1491, 2006.

[15] L. Schinasi, R. A. Horton, V. T. Guidry, S. Wing, S. W. Marshall, and K. B. Morland, "Air pollution, lung function, and physical symptoms in communities near concentrated swine feeding operations," *Epidemiology*, vol. 22, no. 2, pp. 208–215, 2011.

[16] J. McMahon, "Meat and agriculture are worse for the climate than power generation, Steven Chu says," 2019. Accessed: Dec. 6, 2021. [Online]. Available: https://www.forbes.com/sites/jeffmcmahon/2019/04/04/meat-and-agriculture-are-worse-for-the-climate-than-dirty-energy-steven-chu-says

[17] L. Miralha, R. L. Muenich, D. Schaffer-Smith, and S. W. Myint, "Spatiotemporal land use change and environmental degradation surrounding CAFOs in Michigan and North Carolina," *Sci. Total Environ.*, vol. 800, 2021, Art. no. 149391.

[18] C. Robinson, A. Ortiz, J. M. L. Ferres, B. Anderson, and D. E. Ho, "Temporal cluster matching for change detection of structures from satellite imagery," in *Proc. ACM SIGCAS Conf. Comput. Sustain. Soc.*, 2021, pp. 138–146.

[19] B. Chugg, B. Anderson, S. Eicher, S. Lee, and D. E. Ho, "Enhancing environmental enforcement with near real-time monitoring: Likelihood-based detection of structural expansion of intensive livestock farms," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, 2021, Art. no. 102463.

[20] C. Handan-Nader and D. E. Ho, "Deep learning to map concentrated animal feeding operations," *Nature Sustainability*, vol. 2, no. 4, pp. 298–306, 2019.

[21] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, doi: 10.1109/TPAMI.2021.3117983.

[22] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[23] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNSS," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 473–480, 2016.

[24] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, 2021, Art. no. 114417.

[25] D. Hordiiuk, I. Oliinyk, V. Hnatushenko, and K. Maksymov, "Semantic segmentation for ships detection from satellite imagery," in *Proc. IEEE 39th Int. Conf. Electron. Nanotechnol.*, 2019, pp. 454–457.

[26] W. Li, C. He, J. Fang, J. Zheng, H. Fu, and L. Yu, "Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data," *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 403.

[27] D. Rashkovetsky, F. Mauracher, M. Langer, and M. Schmitt, "Wildfire detection from multisensor satellite imagery using deep semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7001–7016, 2021.

[28] T. Vukina, "Vertical integration and contracting in the US poultry sector," *J. Food Distrib. Res.*, vol. 32, pp. 29–38, 2001.

[29] S. Maroney, M. McCool-Eye, A. Fox, and C. Burdett, "Using object-based image analysis to map commercial poultry operations from high resolution imagery to support animal health outbreaks and events," *Geospatial Health*, vol. 15, no. 2, pp. 258–266, 2020.

[30] S. Blundell, D. Opitz, M. Morris, and R. Rao, "Feature analyst v5.0," in *Proc. ASPRS Annu. Conf.*, Portland, OR, USA, 2008.

[31] D. Opitz and S. Blundell, "Object recognition and image segmentation: The feature analyst approach," in *Object-Based Image Analysis*. New York, NY, USA: Springer, 2008, pp. 153–167.

[32] K. A. Patyk et al., "Modelling the domestic poultry population in the United States: A novel approach leveraging remote sensing and synthetic data methods," *Geospatial Health*, vol. 15, no. 2, pp. 244–257, 2020.

[33] C. L. Burdett, B. R. Kraus, S. J. Garza, R. S. Miller, and K. E. Bjork, "Simulating the distribution of individual livestock farms and their populations in the United States: An example using domestic swine (sus scrofa domesticus) farms," *PLoS One*, vol. 10, no. 11, 2015, Art. no. e0140338.

[34] L. J. Young, A. C. Lamas, and D. A. Abreu, "The 2012 census of agriculture: A capture–recapture analysis," *J. Agricultural, Biol. Environ. Statist.*, vol. 22, no. 4, pp. 523–539, 2017.

[35] A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich, and T. Tsunoda, "DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture," *Sci. Rep.*, vol. 9, no. 1, 2019, Art. no. 11399.

[36] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution efficientnets with meta data," *MethodsX*, vol. 7, 2020, Art. no. 100864.

[37] G. Boeing, "OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks," *Comput., Environ. Urban Syst.*, vol. 65, pp. 126–139, 2017.

[38] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. New York, NY, USA: Springer, 2018, pp. 3–11.

[39] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[40] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and State of the Art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[41] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020.

[42] X. Yao, Q. Cao, X. Feng, G. Cheng, and J. Han, "Scale-aware detailed matching for few-shot aerial image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5611711.

[43] P. Yakubovskiy, "Segmentation models pytorch," 2020. [Online]. Available: https://github.com/qubvel/segmentation_models.pytorch

[44] D. Marcos, M. Volpi, N. Komodakis, and D. Tuia, "Rotation equivariant vector field networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5048–5057.

[45] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 96–107, 2018.

[46] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[47] S. Gillies et al., "Rasterio: Geospatial raster I/O for Python programmers," *Mapbox*, 2013. [Online]. Available: https://github.com/rasterio/rasterio

[48] H. Freeman and R. Shapira, "Determining the minimum-area encasing rectangle for an arbitrary closed curve," *Commun. ACM*, vol. 18, no. 7, pp. 409–413, 1975.

[49] D. Eberly, *Minimum-Area Rectangle Containing a Set of Points*. Redmond, WA, USA: Geometric Tools LLC, 2015.

[50] B. Huang, D. Reichman, L. M. Collins, K. Bradbury, and J. M. Malof, "Tiling and stitching segmentation output for remote sensing: Basic challenges and recommendations," 2018, *arXiv:1805.12219*.

[51] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[52] T. Fan, G. Wang, Y. Li, and H. Wang, "MA-Net: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179656–179665, 2020.

[53] C. Copeland, *Animal Waste and Water Quality: EPA's Response to the Waterkeeper Alliance Court Decision on Regulation of CAFOs*. Washington, DC, USA: Congressional Research Service, 2010.

[54] V. Golightly-Howell, "Re: Notification of acceptance of administrative complaint," 2015. Accessed: Dec. 6, 2021. [Online]. Available: https://earthjustice.org/sites/default/files/files/EPA%20Notice%20of%20Acceptance.pdf

[55] R. Purdy, "Using Earth observation technologies for better regulatory compliance and enforcement of environmental laws," *J. Environ. Law*, vol. 22, no. 1, pp. 59–87, 2010.

[56] Environmental Protection Agency, "National pollutant discharge elimination system (NPDES) concentrated animal feeding operation (CAFO) reporting rule," *Federal Register*, vol. 76, pp. 65431–65458, 2011.

[57] D. Bigelow and A. Borchers, "Major uses of land in the United States, 2012," U.S. Dept. Agriculture, Econ. Res. Service, Washington, DC, USA, Tech. Rep. EIB-178, 2017.

[58] J. S. McCarthy, K. Ott, H. Ridolfo, P. McGovern, R. Sirkis, and D. Moore, "Combining multiple methods in establishment questionnaire testing: The 2017 census of agriculture testing Bento box," *J. Official Statist.*, vol. 34, no. 2, pp. 341–364, 2018.

[59] S. Graddy, E. Simon, and S. Rundquist, "Exposing fields of filth: Factory farms disproportionately threaten Black, Latino and Native American North Carolinians," 2020. Accessed: Dec. 6, 2021. [Online]. Available: https://www.ewg.org/interactive-maps/2020-fields-of-filth/

[60] S. Rundquist and D. Carr, "Under the radar: New data reveals NC regulators ignored decade-long explosion of poultry CAFOs," 2019. Accessed: Dec. 6, 2021. [Online]. Available: https://www.ewg.org/sites/default/files/u352/EWG_NC-CAFO_Report_C05.pdf

[61] M. Newman, "Here's how bad Coronavirus has been in Delaware's poultry industry," *Delaware News J.*, 2020. Accessed: Dec. 6, 2021. [Online]. Available: https://www.delawareonline.com/story/news/health/2020/08/25/coronavirus-delawares-poultry-industry/3435723001/

[62] USDA NASS, "Appendix A Census of agriculture methodology," 2017. Accessed: Dec. 6, 2021. [Online]. Available: https://www.nass.usda.gov/Publications/AgCensus/2017/Full_Report/Volume_1,_Chapter_1_US/usappxa.pdf

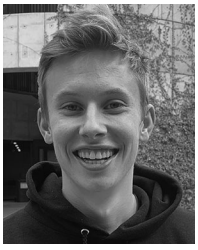
[63] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.

[64] Y. Zhu, Z. Cao, H. Lu, Y. Li, and Y. Xiao, "In-field automatic observation of wheat heading stage using computer vision," *Biosyst. Eng.*, vol. 143, pp. 28–41, 2016.

[65] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *Proc. VLDB Endowment*, vol. 11, pp. 269–282, 2017.
[66] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[67] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

Caleb Robinson received the B.Sc. in computer science from the University of Mississippi, in 2015 and the Ph.D. degree from the School of Computational Science and Engineering at the Georgia Institute of Technology, in 2020.

He is a Research Scientist with the Microsoft AI for Good Research Lab, Redmond, WA, USA. His work focuses on tackling large-scale problems at the intersection of remote sensing and machine learning/computer vision. Some of the projects he works on include estimating land cover from high-resolution satellite imagery across the continent, detecting concentrated animal feeding operations from aerial imagery, and estimating human population density from satellite imagery. He is interested in research topics that facilitate using remotely sensed imagery more effectively in these types of problems in computational sustainability. For example, self-supervised methods for training deep learning models with large amounts of unlabeled satellite imagery, human-in-the-loop methods for creating and validating modeled layers, and domain adaptation methods for developing models that can generalize over space and time.



Ben Chugg received the B.Sc. degree in mathematics and computer science from the University of British Columbia, Vancouver, BC, Canada, in 2018, and the M.Sc. degree in mathematics and theoretical computer science from Oxford University, Oxford, U.K., in 2019. He is currently working toward the Ph.D. degree with Carnegie Mellon University, Pittsburgh, PA, USA.

His is currently the Lead Research Analyst with Stanford's RegLab, Stanford, CA, USA.

Brandon Anderson received the Ph.D. degree in physics from the University of California, Santa Cruz, in 2012.

He was a Research Associate with the Oskar Klein Centre for Cosmoparticle Physics, Stockholm, Sweden, where he focused on dark matter and statistical inference in gamma-ray astronomy. Pivoting into the startup world as a Data Scientist, he designed experiments and developed machine learning techniques for the analysis of large biological datasets with BioElectron Corporation, Mountain View, CA, USA, and led a team to build out real-time unsupervised fault detection systems in autonomous vehicles with Cognomotiv, Menlo Park, CA. Most recently, he has served as the Head of Data Science with Stanford's RegLab, Stanford, CA, where he provided research direction, team management, and technical support for a wide variety of projects in collaboration with government agencies and nonprofits. He is currently putting his considerable experience to public use at the Research, Applied Analytics, and Statistics Division, IRS.

Juan M. Lavista Ferres received dual B.A. degree in information systems and software engineering in 1999 and 2001, respectively, and the M.Sc. degree in computer science in 2005, from the Catholic University in Uruguay, Montevideo, Uruguay, and the graduate degree in data mining and machine learning from Johns Hopkins University, Baltimore, MD, USA.

He is the Chief Data Scientist and the Lab Director of the Microsoft AI For Good Research Lab, Redmond, WA, USA, where he works with a team of data scientists and researchers in artificial intelligence, machine learning, and statistical modeling, working across disinformation, cybersecurity, and all other Microsoft AI For Good efforts. He is involved in working to define the data science discipline within Microsoft, Redmond, WA, USA. Before joining Microsoft, he was the chief technology officer and co-founder of alerts.com.

Mr. Ferres is the Editor for *Microsoft Journal of Applied Research*.

Daniel E. Ho received the J.D. degree from Yale Law School, New Haven, CT, USA, in 2005, and the Ph.D. degree in government from Harvard University, Cambridge, MA, USA, in 2004.

He is currently the William Benjamin Scott and Luna M. Scott Professor of Law with Stanford RegLab, Stanford Law School, Stanford, CA, USA, a Professor of Political Science, and a Senior Fellow with the Stanford Institute for Economic Policy Research, Stanford. He is also the Associate Director of the Stanford Institute for Human-Centered Artificial Intelligence, a Faculty Fellow with the Center for Advanced Study in the Behavioral Sciences, and the Director of the Regulation, Evaluation, and Governance Lab. He serves on the National Artificial Intelligence Advisory Commission, advising the White House on artificial intelligence, and as a Public Member of the Administrative Conference of the United States. He clerked for Judge Stephen F. Williams on the U.S. Court of Appeals, District of Columbia Circuit.