Surveilling Surveillance: Estimating the Prevalence of Surveillance Cameras with Street View Data

Hao Sheng
haosheng@stanford.edu
Stanford University
Stanford, California, USA

Keniel Yao
keniel.yao@stanford.edu
Stanford University
Stanford, California, USA

Sharad Goel
scgoel@stanford.edu
Stanford University
Stanford, California, USA

arXiv:2105.01764v1 [cs.CY] 4 May 2021

ABSTRACT

The use of video surveillance in public spaces—both by government agencies and by private citizens—has attracted considerable attention in recent years, particularly in light of rapid advances in face-recognition technology. But it has been difficult to systematically measure the prevalence and placement of cameras, hampering efforts to assess the implications of surveillance on privacy and public safety. Here we present a novel approach for estimating the spatial distribution of surveillance cameras: applying computer vision algorithms to large-scale street view image data. Specifically, we build a camera detection model and apply it to 1.6 million street view images sampled from 10 large U.S. cities and 6 other major cities around the world, with positive model detections verified by human experts. After adjusting for the estimated recall of our model, and accounting for the spatial coverage of our sampled images, we are able to estimate the density of surveillance cameras visible from the road. Across the 16 cities we consider, the estimated number of surveillance cameras per linear kilometer ranges from 0.1 (in Seattle) to 0.9 (in Seoul). In a detailed analysis of the 10 U.S. cities, we find that cameras are concentrated in commercial, industrial, and mixed zones, and in neighborhoods with higher shares of non-white residents—a pattern that persists even after adjusting for land use. These results help inform ongoing discussions on the use of surveillance technology, including its potential disparate impacts on communities of color.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Applied computing** → **Law, social and behavioral sciences**.

KEYWORDS

Computer vision, privacy, urban computing

ACM Reference Format:
Hao Sheng, Keniel Yao, and Sharad Goel. 2021. Surveilling Surveillance: Estimating the Prevalence of Surveillance Cameras with Street View Data. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AIES '21, May 19–21, 2021, Virtual Event, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Surveillance cameras, also known as closed-circuit television (CCTV) systems, have proliferated in the last several decades as the costs to record and store video have fallen dramatically. As of 2016, there were an estimated 350 million surveillance cameras worldwide [25]. The United States, with an estimated 50 million CCTV cameras installed, is believed to have the highest per capita number of surveillance cameras (15.3 CCTV cameras per 100 people) in the world [31].

Past work has found that surveillance cameras may play an important role in crime prevention and investigation, but there is also growing concern about the dangers cameras pose to privacy and equity. Further, recent advances in facial recognition technology significantly amplify both the potential costs and the potential benefits of widespread surveillance, as it is now possible to identify and track specific individuals across space and time. While these technical advances promise to aid law enforcement efforts, they may also unjustly concentrate policing on more heavily monitored communities. This surveillance may also hinder longstanding freedoms of speech and association, as it becomes easier to identify those participating in public gatherings, potentially dissuading dissent.

Despite the wide-ranging implications of surveillance cameras on public safety, police enforcement, and democratic governance, relatively little is known about the precise number and placement of cameras, hampering efforts to assess their impacts. Past work to gauge the prevalence and spatial distribution of surveillance cameras has either examined aggregate production or shipping numbers, or relied on public disclosures in select jurisdictions—approaches that suffer from limitations of scale and scope.

We present a new, scalable approach to measuring the distribution of outdoor surveillance cameras across the United States, and, more generally, across the world. In particular, we apply computer vision algorithms—coupled with expert human annotators—to analyze large-scale corpora of street view images that provide nearly complete visual coverage of many cities. In this manner, we leverage the proliferation of cameras and image data themselves to quantify the prevalence of surveillance technology.

We specifically use the public repository of images collected as part of Google's Street View service, launched in 2007. Since its inception, millions of 360-degree panoramas have been collected by cameras mounted on the roof rack of Google Street View cars, covering more than 10 million miles across 83 countries [33]. The rich archive of historical street view images provides opportunities to understand the evolution of the built environment, particularly the adoption of surveillance cameras on a global scale. However, it is still extremely challenging—if not impossible—for humans to eyeball millions of images and spot cameras from the diverse street

view context: a camera usually consists of 30–50 pixels out of over 400,000 pixels in a standard 640 × 640 street view image.

To scour this collection of images, we train and apply a computer vision algorithm to first filter street view data to those candidate images likely to contain a surveillance camera. We specifically start with a random selection of 1.6 million street view images from 10 large U.S. cities and 6 other major cities, which contains approximately 6,000 positive model detections. This curated set of candidate images is then examined by human experts for verification. To go from verified camera detections in our sample to city-wide estimates, we further estimate both the recall of our model (which we find is 0.63), and the proportion of the city covered by our sample. This latter quantity is computed based on the recorded camera position and angle, coupled with high-precision data on the road network and building footprints.

We find substantial variation in the density of visible surveillance cameras across the 16 cities we consider, ranging from 0.07 cameras per linear kilometer along the road network in Seattle to 0.95 cameras per kilometer in Seoul. Examining the 10 U.S. cities in greater detail, we find that surveillance cameras are concentrated in commercial, industrial, and mixed city zones, and also in areas with higher shares of non-white residents. This concentration of cameras in majority-minority neighborhoods persists even after adjusting for zone, pointing to the potential disparate impacts of surveillance technology on communities of color.

2 RELATED WORK

Our work connects to several interrelated strands of research in computer vision, urban computing, and privacy, which we briefly summarize below.

2.1 Street View Understanding

Visual scene understanding [14] is one of the most fundamental and challenging goals in computer vision. In part because of its potential to support self-driving vehicles, both the industrial and scientific communities have put considerable effort and investment into designing and creating labeled street view datasets for training and evaluating deep learning models, such as CamVid [3], the KITTI Vision Benchmark Suite [12], Cityscapes [9], and Mapillary Vistas [27]. Based on these datasets, several studies exploit the characteristics of urban-scene images and propose object segmentation [8, 23] and change detection [1] algorithms for general street view understanding.

Related research has focused on detecting specific elements from street images, including greenery [22], buildings [20], and city infrastructure such as utility poles [41]. In contrast, the problem of camera detection that we consider in this work is relatively understudied and is particularly challenging given the small size of the objects in question. Further, the Mapillary Vistas dataset is one of the only public resources that provide street view images with labeled cameras—and that dataset has fewer than 20 examples in the U.S. As a result, we construct our own camera dataset through manual labeling.

2.2 Urban Computing

Urban computing aims to tackle major issues in cities—such as traffic control, public health, and economic development—by modeling and analyzing urban data. A large body of research has shown it is possible to infer socioeconomic information from satellite images [17, 36], monitor human mobility [40], and identify geo-tagged social network activities [35]. Recent studies using street view images have dramatically increased the accuracy of processed data, as well as the geographic resolution analyzed. By manually scoring street view images from 2,709 city blocks, Hwang and Sampson [16] find that gentrification in Chicago from 2007 to 2009 was negatively associated with the concentration of minority groups. Mooney et al. [26] labeled the characteristics of 532 intersections in New York City, such as curb cuts and crosswalks, to assess environmental contributions to pedestrian injury.

As an alternative to relying on human experts to annotate street view images, modern computer vision algorithms have a much higher throughput with close to zero cost, enabling researchers to scale to multiple cities. For example, Gebru et al. [11] enumerated 22 million automobiles (8% of all vehicles in the U.S.) in 50 million street view images to accurately estimate local income, race, education, and voting patterns. In our work, we draw on the merits of both approaches, combining high-recall computer vision algorithms with high-precision human verification in a unified estimation pipeline.

2.3 Surveillance and Privacy

While past work has found that surveillance cameras play an important role in crime investigation [21] and deterrence [39], cameras also pose significant challenges to privacy. Legal scholars have long considered the ramifications of cameras on First Amendment freedoms and the constitutional right to privacy [34]. More recently, scholars have been concerned with the role of surveillance cameras in predictive policing [19], in enabling the adverse effects of facial recognition and computer vision [4, 5, 28, 37], and with the threat of surveillance hacking [13, 32]. These concerns have led to bans on facial-recognition technology by law enforcement in San Francisco, Boston, and Portland [10], as well as the drafting of federal legislation [15].

Despite these concerns, there has been limited success in identifying the number and geospatial distribution of cameras. The Electronic Frontier Foundation (EFF) recently acquired the locations of cameras accessible by prosecutors in San Francisco [24]. Other private-market researchers have estimated the prevalence of installed cameras at a national level through unit shipments [18]. However, neither of these approaches are able to estimate the prevalence and specific locations of public and private cameras at scale, hindering downstream analysis on the impacts of surveillance.

3 DATA AND METHODS

For 16 major cities, we estimate the total number and spatial distribution of surveillance cameras visible from the street. We specifically consider the 10 cities with the highest urban density in the U.S., among those with at least 500,000 residents, and 6 other major cities in Asia and Europe. Our statistical estimation procedure involves three key steps. First, we compile a dataset of street view images both with and without cameras and label these images with

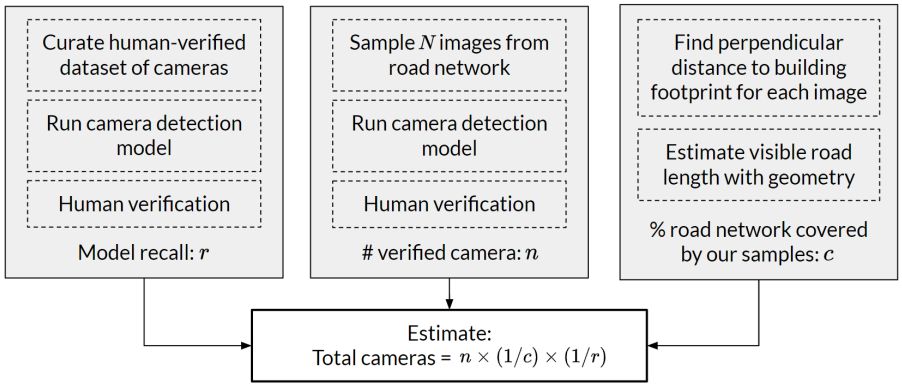


Figure 1: Camera estimation workflow. We take a three-step approach to estimating the prevalence and spatial distribution of surveillance cameras across 16 major cities. First, we create a labelled dataset of street view images to estimate the recall of our camera-detection algorithm. Next, we run this camera-detection algorithm on a random sample of street images, and then use human annotators to identify images with actual cameras. Finally, we estimate the proportion of the road network covered by our random sample of street images.

segmentation masks. We then train a camera segmentation model on this dataset, and, importantly, estimate the recall of our detection algorithm on a held-out validation dataset. Second, we run our camera detection algorithm on a random sample of street view images. All positive camera detections are then reviewed by human experts to remove false positives. Finally, by combining the geometry of the camera angle, the road network, and building footprints, we calculate our sample's coverage of the road network. These three steps are outlined in Figure 1. In the following sections, we describe the data used in our analysis and more fully detail each step in our estimation pipeline.

3.1 Data

We analyze the 16 cities listed in Table 1. For each city, we obtained the road network and building footprints from OpenStreetMap [2, 29]. U.S. Census maps were used to restrict the geospatial data to the city's administrative borders. All street view images used for model training and camera detection were accessed through the Google Static Streetview API.[1] We further used San Francisco camera location data from the EFF [24] to construct training and evaluation datasets for our model.

3.2 Step 1: Model Training and Evaluation

We start by creating training and evaluation datasets for our camera detection model. For each of the 2,660 geo-tagged cameras in San Francisco identified by the EFF, we pulled the closest street view images from 2012–2019 (if there is a scene available within 30 meters). Manually labeling the resulting 13,240 images yielded 861 positive instances containing 977 cameras. We note that many of the cameras listed in the EFF dataset appear to be indoors or otherwise are not visible from the street. In Figure 2, we show several labeled examples.

[1] https://developers.google.com/maps/documentation/streetview

City	Population	Area (sq. km)	Road length (km)
Los Angeles	3,793,000	1,213	21,095
New York City	8,175,000	783	16,362
Chicago	2,696,000	589	10,449
Philadelphia	1,526,000	347	6,759
Seattle	609,000	217	5,569
Milwaukee	595,000	248	4,899
Baltimore	621,000	209	3,746
Washington, D.C.	602,000	158	3,262
San Francisco	805,000	121	3,101
Boston	618,000	125	2,589
Tokyo	13,159,000	2,194	46,688
Bangkok	8,305,000	1,569	34,692
London	8,174,000	1,572	28,907
Seoul	9,630,000	605	14,748
Singapore	3,772,000	728	5,794
Paris	2,244,000	106	1,853

Table 1: We estimate the prevalence and location of surveillance cameras in the 10 large U.S. cities and 6 other major cities around the world. Above we list the population and land area of these cities, and the estimated length of the road network within the city bounds, in descending order of road length.

We frame our camera detection problem as a binary image segmentation problem to maximize learning from a limited number of samples. We split the positive images by location into 70%/15%/15% training/validation/test sets, making sure images from the same site always belong to the same split. We further include all camera instances from Mapillary Vista into our training dataset. By mixing with the negative images, we end up with 5,298 images for training, 1,040 for validation, and 1,040 for testing.

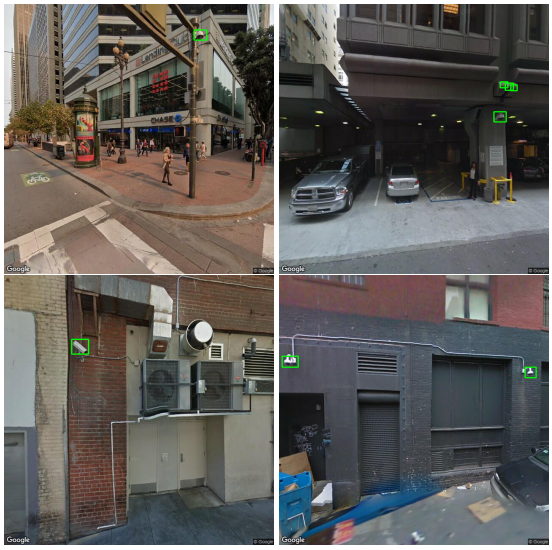


Figure 2: Examples of surveillance cameras in San Francisco. Upper left: Dome camera mounted on a traffic pole. Upper right: Dome camera mounted on the wall of a parking structure. Lower left: A wall mounted directed camera. Lower right: Two wall-mounted dome cameras.

Data Split	$\text{IoU}_{\text{camera}}$	Accuracy	F1-score
Validation Set	0.71	0.94	0.90
Test Set	0.69	0.93	0.89

Table 2: Performance of our camera segmentation model in pixel space.

Our segmentation model follows the architecture of DeepLab V3+ [6, 7] with an EfficientNet-b3 [38] backbone. We apply a random horizontal flip and randomly crop the original image (640 x 640 pixels) to 320 x 320 before feeding it into the model during training. In the inference phase, we first crop the input image into four patches and then merge the output segmentation maps back to the original size. The segmentation model's performance is shown in Table 2. To aggregate the pixel level prediction to the instance level, we first apply a morphological dilation with a 3x3 kernel to merge detected areas and filter false detections by size. After validating several combinations of pixel-level probability thresholds and size filters, we decided to use a probability threshold of 0.75 and a size threshold of 50 pixels, which yields precision and recall equal to 0.58 and 0.63, respectively (see Figure 3).

In Figure 4, we present several illustrative failures of our detection model. The model is occasionally confused by objects that share some of the visual features of cameras, such as building structures, parking meters, and street lamps. In some instances, our model also merges multiple cameras into one detection. These problems are mitigated by the human verification step, as described in the next section.

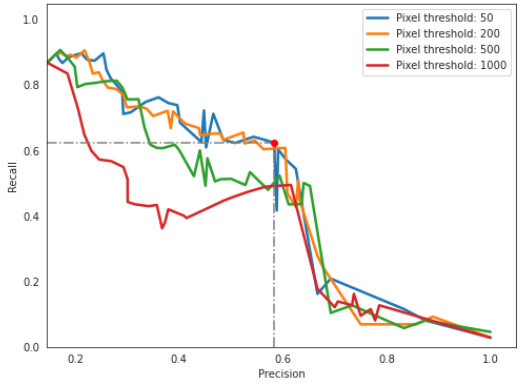


Figure 3: Model precision and recall at camera instance level as we vary the classification threshold, for different size filters (in pixels). We decided to use a probability threshold of 0.75 and a size threshold of 50 pixels, which yields precision and recall equal to 0.58 and 0.63, respectively.

Figure 4: Examples of false positives and other errors from our camera detection model. Upper left: Building structure. Upper right: Parking meter. Lower left: Street lamp. Lower right: Our algorithm incorrectly merged two cameras into one detection.

3.3 Step 2: Camera Detection and Verification

For each city, we sampled street view images at $N = 100,000$ points chosen uniformly at random from the road network.[2] For approximately 3% of the selected points, there was no street view coverage within 10 meters, in which case we discarded and then

[2]For reference, there are more than 400,000 points covered with distinct street view panoramas in San Francisco.

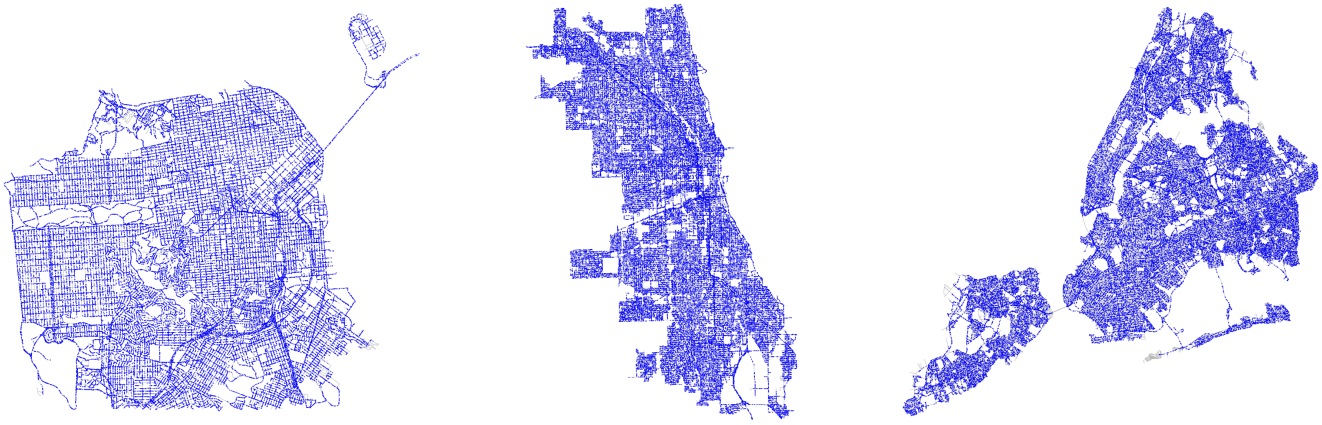


Figure 5: The spatial distribution of sampled points on the road network for three illustrative cities. Left: San Francisco; Middle: Chicago; Right: New York City.

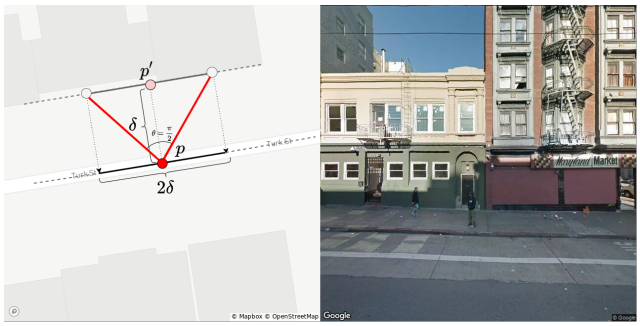


Figure 6: Illustration of how we estimate the road segment coverage for one image. Right: The Google Street View image. Left: Corresponding field of view on map. We obtain the point location of each image, p, and find the closest point p' within the footprint of the nearby buildings. The length of the road in view is twice the distance d between p and p', as each image has a 90-degree field facing the road. We estimate the road segment covered in the shown image has a length of 19.55 meters.

re-sampled the location. Figure 5 shows the spatial distribution of the sampled points for three example cities: San Francisco, New York, and Chicago.

For each location, we then selected a 360-degree street view panoramas from 2016–2020, selecting the most recent available panorama in each period. Finally, for each location sampled, we randomly selected one out of the two 90-degree views with a midpoint perpendicular to the orientation of the road (see Figure 6). This approach provides the maximum view of the roadside.

For $N = 5,000$ of the randomly-selected locations from U.S. cities, we also selected a 360-degree street view panorama from 2011-2015, selecting the most recent panorama available. This allowed us to assess change in camera density over time.

We ran our camera detection model on the resulting set of 100,000 images from 2016–2020 for each of the 16 cities and the additional 5,000 images from 2011-2015 for each of the U.S. cities. Annotators received the raw image and bounding boxes highlighting the predicted cameras, which were automatically generated from the model segmentation outputs. This process yielded 6,281 positive images with a total of 6,469 camera detections, all of which were then verified by a human annotator.

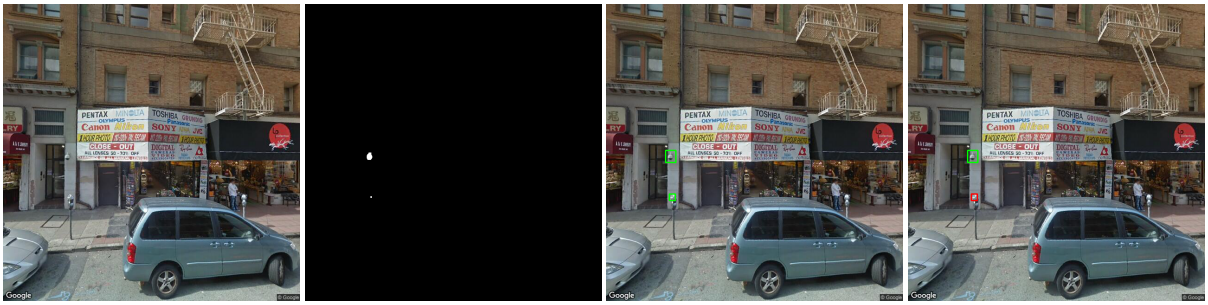
Figure 7 illustrates the pipeline from the raw image to segmentation and bounding boxes to human verification. In our subsequent analysis, we only consider these human-verified camera detections.

3.4 Step 3: Road Network Coverage Estimation

The final step in our procedure is to estimate the fraction of the visible area covered by our randomly sampled images. To estimate how much of the total length of a city's road network (D) has been covered by our sample (D_{covered}), we estimate the average length of road covered by one street view image ($\bar{d}$), which we can then multiply by the number of images sampled (N).

We estimate $\bar{d}$ based on the geometry illustrated in Figure 6. Each street view image comes with the exact latitude and longitude of where it was taken. Given an image's point location p, we find the closest point p' within the nearby buildings' footprint, and denote the distance between p' and p with δ. As discussed above, we chose the street view's heading to be perpendicular to the road orientation, and restricted it to a 90-degree view. As a result, we estimate the length of road segment covered by the image to be $d = 2\delta$.

We repeat this procedure for each sampled street view image. We remove the relatively small number of images taken at locations more than 30 meters from a building—corresponding to 60 meters of street coverage—since at further distances, cameras become too small to be reliably detected by either humans or computer vision algorithms. This results in images that cover about 25–30 meters of street. For example, the mean road segment covered by an image $\bar{d}$ is 24, 29, and 28 meters in San Francisco, Chicago, and New York City, respectively, as shown in Figure 8.



(a) Raw Image (b) Segmentation (c) Bounding Boxes (d) Verification

Figure 7: Illustration of our machine detection and human verification pipeline. For each image yielding a positive model detection, human coders verify the detection based on the raw image (a) and bounding boxes (c) generated from the segmentation (b). In this example, the human coder confirmed the upper box is a camera but not the lower box (d).

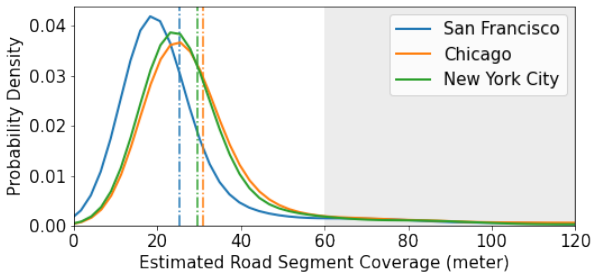


Figure 8: Distribution of the length of road covered by a street view image in San Francisco, Chicago and New York City. We limit our analysis to images taken within 30 meters of the nearby buildings—corresponding to 60 meters of street coverage—since we cannot reliably detect cameras at further distances.

Now, we estimate the proportion of a city's road network covered by our sample as $c = (N\bar{d})/(2D)$, where N is the total number of samples for a city (within a given time period) and D is the total length of the city's road network. Note that the factor of 2 is to account for the fact that our street view images only cover one of the two sides of a street at any sampled point.

Finally, putting all the above pieces together, we can estimate the number of cameras K_{it} in city i in time period t:

$$\hat{K}_{it} = \frac{n_{it}}{c_{it} r}, \tag{1}$$

where r is the recall of our model, n_{it} is the number of verified camera detections, and c_{it} is the proportion of the road network of a city covered by our sample.

Similarly, we model variance by treating each sampled instance as a draw from a Bernoulli distribution with detection probability $p_{it} = n_{it}/N_{it}$. Assuming that recall and coverage are both exact, we can estimate the standard error of the number of cameras K_{it} as:

$$\hat{\text{se}}(\hat{K}_{it}) = \frac{\sqrt{N_{it} \cdot p_{it} \cdot (1 - p_{it})}}{c_{it} r}. \tag{2}$$

4 RESULTS

Applying the methods described above, we now estimate the total number and spatial distribution of cameras on the road network for all 16 cities, and the change in density between the two time periods for the 10 U.S. cities. In addition, for the U.S. cities, we estimate the prevalence of cameras across city zones, and examine the racial composition of the neighborhoods in which cameras are concentrated.

4.1 Camera Prevalence

Table 3 shows the number of identified cameras for each city and period—after human verification—along with point estimates and 95% confidence intervals for camera density and for the total number of cameras, following Eqs. (1) and (2). The same density estimates for 2016–2020 are also depicted in descending order in Figure 10. We find that camera density varies widely between cities: For example, Boston and New York City, the U.S. cities with the highest camera density, have almost four times as many cameras per kilometer than Seattle and Los Angeles.[3] We note that our estimates exclude indoor cameras, as well as outdoor cameras not captured by street view images. Perhaps due to these limitations, our estimate of 10,100 cameras in New York City is lower than the 18,000 cameras that the NYPD reportedly has access to [30].

Interestingly, as shown in Table 3, our camera estimates do not vary substantially between the two time periods we consider (2011–2015 and 2016–2020), suggesting that the installation of surveillance cameras in these cities may have reached a plateau.

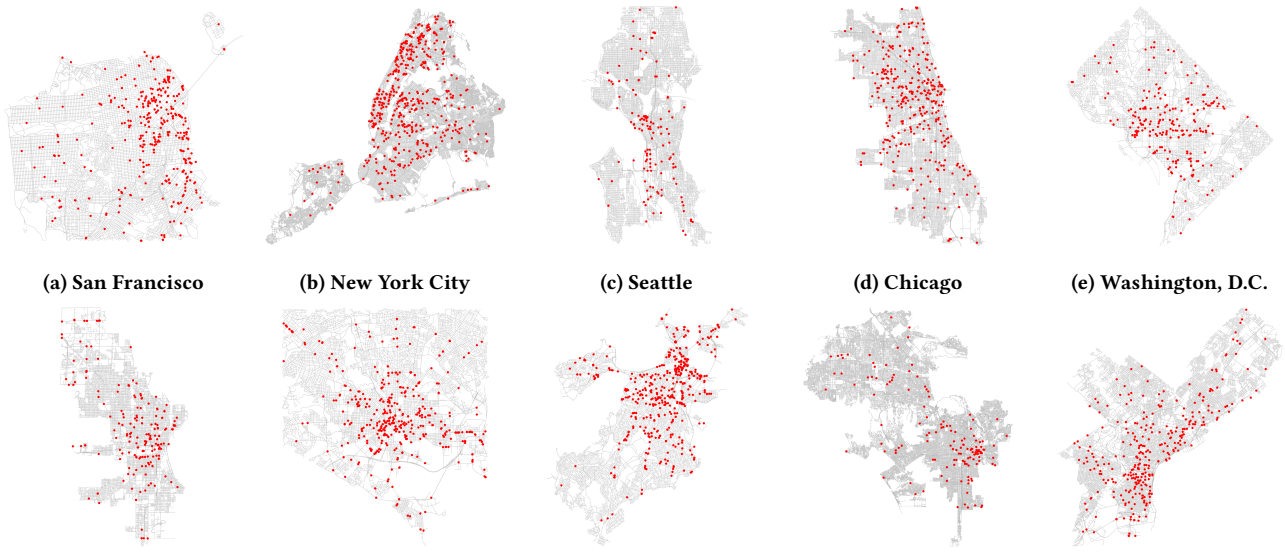
4.2 Camera Placement

The detection maps in Figure 9 show that cameras are not distributed uniformly across a city. Despite sampling uniformly over the road network, we find densely covered regions in each city, representing neighborhoods with a high concentration of cameras. We examine these patterns in more detail for the 10 U.S. cities we consider, analyzing the rate of (verified) camera identifications per

[3] Our computer vision model was trained on San Francisco data, and so it is possible that the camera identification rate in San Francisco is inflated due to over-fitting. We note, however, that while model precision varies between cities, Philadelphia and Boston both have higher precision than San Francisco, which suggests our model does indeed transfer well across contexts.

City	Road length (km)	Mean road coverage (m)	No. of detections		Estimated density (cameras/km)		Estimated number of cameras	
			2011-2015	2016-2020	2011-2015	2016-2020	2011-2015	2016-2020
Boston	2,589	26	26	516	0.63 (0.12)	0.63 (0.03)	1,600 (300)	1,600 (100)
New York	16,362	29	29	556	0.65 (0.12)	0.62 (0.03)	10,600 (2,000)	10,100 (400)
Baltimore	3,746	30	35	512	0.74 (0.12)	0.54 (0.02)	2,800 (500)	2,000 (100)
San Francisco	3,101	24	25	398	0.66 (0.13)	0.52 (0.03)	2,000 (400)	1,600 (100)
Chicago	10,449	30	19	382	0.41 (0.09)	0.41 (0.02)	4,300 (1,000)	4,300 (200)
Philadelphia	6,759	29	22	348	0.48 (0.10)	0.38 (0.02)	3,200 (700)	2,600 (100)
Washington	3,262	33	13	237	0.25 (0.07)	0.23 (0.01)	800 (200)	700 (50)
Milwaukee	4,899	33	9	202	0.17 (0.06)	0.19 (0.01)	800 (300)	900 (100)
Seattle	5,569	29	3	155	0.07 (0.04)	0.17 (0.01)	400 (200)	1,000 (100)
Los Angeles	21,095	29	7	144	0.15 (0.06)	0.16 (0.01)	3,200 (1,200)	3,300 (300)
Seoul	14,748	29	-	869	- -	0.95 (0.03)	- -	13,900 (500)
Paris	1,853	24	-	590	- -	0.76 (0.03)	- -	1,400 (100)
Tokyo	46,688	29	-	428	- -	0.47 (0.02)	- -	21,700 (1,000)
London	28,907	32	-	448	- -	0.45 (0.02)	- -	13,000 (600)
Bangkok	34,692	29	-	324	- -	0.35 (0.02)	- -	12,200 (700)
Singapore	5,794	29	-	172	- -	0.19 (0.01)	- -	1,100 (100)

Table 3: Number of detections and estimates of camera density and total cameras in each city for 2011–2015 (N = 5,000 images) and 2016–2020 (N = 100,000 images), arranged in descending order of camera density in 2016–2020. Standard errors are listed in parentheses. We impute a mean road coverage of 29m for cities without adequate building footprint data.



(a) San Francisco **(b) New York City** **(c) Seattle** **(d) Chicago** **(e) Washington, D.C.**

(f) Milwaukee **(g) Baltimore** **(h) Boston** **(i) Los Angeles** **(j) Philadelphia**

Figure 9: Locations of verified cameras in 10 large U.S. cities for the period 2016–2020. Densely clustered areas of points indicate regions with high camera density in each city. Camera density varies widely between neighborhoods. Note: Scale varies between cities.

street image across zoning designations and neighborhood racial composition, using our dense sample of images from 2016–2020.

Figure 11 shows the camera identification rate for different zoning designations aggregated over the 10 U.S. cities we analyze. We find that images from mixed, industrial, and commercial zones are more likely to contain an identified camera than images from public (such as parks and other public facilities) and residential areas. For example, the identification rate in mixed zones (2.1%) is more than three times the rate in residential zones (0.6%). This pattern holds for the majority of our chosen cities.

To compute the camera identification rate, we assigned each sampled point to the zoning designation of the closest parcel of land. To do so, we collected land use and zoning designation data for all 10 cities, and then standardized the zoning code into one of the following five categories: mixed, industrial, commercial, public, and residential. Zones with codes that represent planned development or did not clearly fit into the aforementioned categories are labeled

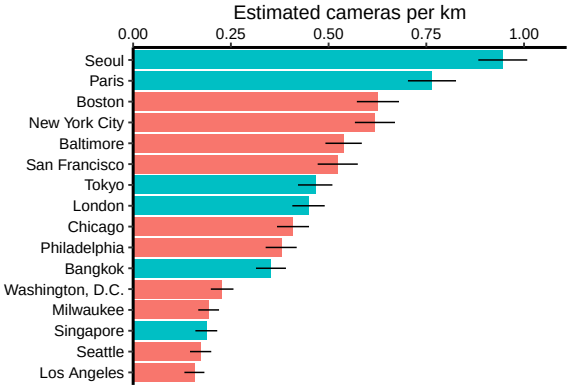


Figure 10: Estimated camera density (cameras per km) for ■ 10 large U.S. cities and ■ 6 other major cities for the period 2016–2020. Horizontal lines indicate 95% CIs for the estimate.

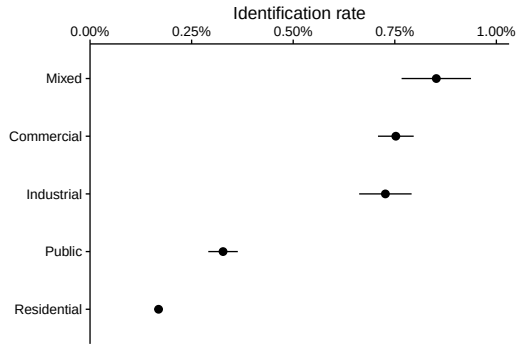


Figure 11: Camera identification rates (cameras per image) for different zoning designations across 10 large U.S. cities for the period 2016–2020. Horizontal lines indicate 95% CIs for the estimate. Cameras are more likely to be detected in mixed, industrial, and commercial zones than in public and residential zones.

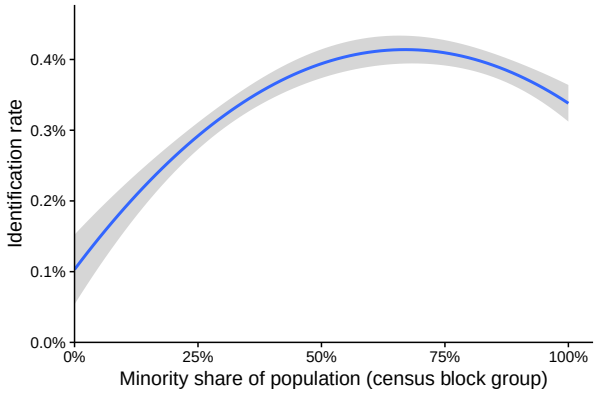


Figure 12: The relationship between camera identification rate (cameras per image) and minority share of population, aggregated across 10 large U.S. cities for the period 2016–2020. The blue line shows a regression fit to the data, and indicates that cameras are more likely to be detected in majority-minority neighborhoods than in predominantly non-Hispanic white neighborhoods.

	Identification Rate
Public	0.0014^{***} (0.0002)
Commercial	0.0055^{***} (0.0002)
Industrial	0.0053^{***} (0.0002)
Mixed	0.0060^{***} (0.0003)
Percentage minority	0.0059^{***} (0.0009)
Percentage minority^2	-0.0044^{***} (0.0008)
Observations	787,418
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Coefficients and standard errors for a linear probability model of camera identification as a function of city, zone, and minority share (with a quadratic term), fit on sampled points in the period 2016–2020. The reference group for zone is residential neighborhoods.

as unknown and omitted in the following analysis. We find that 60% of sampled points are classified as residential, and unknown codes comprise less than 3% of sampled points.

We next examine the relationship between camera identification rate and the share of residents in the surrounding area that identify as belonging to a minority racial or ethnic group, aggregated over our 10 U.S. cities. To compute this relationship, we assigned each sampled image to the minority proportion of the census block group in which it is located, as estimated by the 2018 American Community Survey. For purposes of this analysis, we define "minority" as comprising those individuals who identify either as Hispanic (regardless of their race) or who do not identify as white.

Figure 12 shows the results. The blue line is a regression (with both linear and quadratic terms) fit to the data, and indicates that an increase in the share of minority residents in a neighborhood is associated with an increase in camera identification rate. For example, the identification rate in census blocks with a 50% minority share (0.38%) is twice as high as in those blocks with a 10% minority share (0.2%). We see qualitatively similar results with higher-order polynomial fits.

The observed concentration of cameras in majority-minority neighborhoods persists even after adjusting for zone category. Specifically, Table 4 shows the results of a linear probability model that predicts camera detections as a function of city, zone, and racial composition—where we again use a quadratic term to account for the curvature seen in Figure 12. The fitted model confirms that camera identifications increase with the minority share of residents, plateauing at approximately 60% share and then remaining relatively flat, as in Figure 12. It is unclear what is driving the apparent concentration of cameras in high minority neighborhoods. However, regardless of the underlying mechanism, these results

point to the potential impacts that video surveillance can have on communities of color.

5 DISCUSSION AND CONCLUSION

By applying state-of-the-art computer vision techniques to large-scale, geo-tagged image data, we are—for the first time—able to estimate the number and spatial distribution of outdoor surveillance cameras in 16 major cities around the world. Further, the approach we have developed has the potential to scale to even more cities across the country and the world, providing a new perspective on the state of video surveillance.

In the 16 cities we analyzed, we found considerable variation in the estimated number of surveillance cameras. Interestingly, though, the prevalence of detected cameras in every city is similar across the two time periods we consider (2011–2015 and 2016–2020), which suggests we may have reached a (perhaps temporary) plateau. Among U.S. cities, our analysis also shows that cameras are more likely to be found in industrial, commercial, and mixed zones as compared to residential areas. Finally, even after adjusting for zone category, we find a greater concentration of cameras in majority-minority neighborhoods, highlighting the need to carefully consider the potential disparate impacts of surveillance technology on communities of color.

While our computational approach is able to provide a novel quantitative perspective into the state of surveillance, it is still subject to several important limitations, which we outline below. First, our method relies on being able to see cameras from the street, and, more specifically, from street view images. Indoor cameras, as well as outdoor cameras obscured from view, are not counted by our estimation pipeline. Further, due to the limited resolution of street view images, small cameras—such as increasingly popular doorbell cameras—are difficult to detect by either humans or algorithms. Higher resolution and higher coverage image data could mitigate these issues in the future. However, our current results likely underestimate the density of cameras in a city.

Second, our human annotators may not perfectly label cameras in the candidate images selected by the model, skewing our final estimates. For example, it is possible that they rule out an actual camera (leading to an undercount) or, conversely, that they report a camera that is not in fact there (leading to an overcount). To minimize these errors, every candidate image is independently labeled by three human annotators, but at least some errors likely remain.

Third, errors in the estimated recall of our computer vision model—and, similarly, errors in the estimated coverage of our images—can bias our final estimates. Estimating city-specific model recall is particularly challenging, as it requires city-specific labeled datasets. In our analysis, we thus estimated recall for a single city, San Francisco, in which the locations of some surveillance cameras had already been compiled, which we then apply to other jurisdictions. Further, our variance estimates treat the recall and coverage as known quantities. Accounting for errors in their measurement would increase the variance of our final estimates.

Finally, our method does not provide any information about the cameras other than what can be inferred from their appearance. For example, we cannot determine whether identified cameras are decoys, are malfunctioning, or otherwise are not in use. We likewise cannot always tell who owns the cameras (e.g., a government agency or a private citizen), who has access to the video, and whether the camera footage is stored. All of these factors are critical in assessing the downstream consequences of video surveillance. Although difficult, future work may be able to answer some of these questions by conducting a more intensive audit of a sample of the identified cameras.

Despite these limitations, we believe our approach and results constitute an important step toward understanding the use of surveillance technology across the world. More broadly, our general statistical estimation pipeline can be extended and applied to characterize the prevalence and spatial distribution of a variety of other city elements detectable from street images. Looking forward, we hope this work spurs further theoretical and empirical research at the intersection of computer vision, urban computing, and public policy.

REFERENCES

[1] Pablo F Alcantarilla, Simon Stent, German Ros, Roberto Arroyo, and Riccardo Gherardi. 2018. Street-view change detection with deconvolutional networks. *Autonomous Robots* 42, 7 (2018), 1301–1322.

[2] Geoff Boeing. 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* 65 (2017), 126–139.

[3] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. 2008. Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*. Springer, 44–57.

[4] J. Buolamwini and T. Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR*, Vol. 81. 77–91.

[5] M. Ryan Calo. 2010. People Can Be So Fake: A New Dimension to Privacy and Technology Scholarship. *114 Pennsylvania State Law Review* 809 (2010), 828–829.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[8] Sungha Choi, Joanne T Kim, and Jaegul Choo. 2020. Cars Can't Fly Up in the Sky: Improving Urban-Scene Segmentation via Height-Driven Attention Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9373–9383.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.

[10] Fight for the Future. 2020. Ban Facial Recognition. https://www.banfacialrecognition.com. Accessed: 2021-01-30.

[11] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences* 114, 50 (2017), 13108–13113.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.

[13] Peter Hermann. 2018. Hack of D.C. police cameras was part of ransomware scheme, prosecutors say.

[14] Derek Hoiem, James Hays, Jianxiong Xiao, and Aditya Khosla. 2015. Guest editorial: Scene understanding. *International Journal of Computer Vision* 112, 2 (2015), 131–132.

[15] H.R. 7235. 2019. Stop Biometric Surveillance by Law Enforcement Act.

[16] Jackelyn Hwang and Robert J Sampson. 2014. Divergent pathways of gentrification: Racial inequality and the social order of renewal in Chicago neighborhoods. *American Sociological Review* 79, 4 (2014), 726–751.

[17] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.

[18] Niall Jenkins. 2015. Video Surveillance: New Installed Base Methodology Yields Revealing Results.

[19] Elizabeth E. Joh. 2016. The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing. *Harvard Law and Policy Review* 10, 1 (2016), 15–42.

[20] Jian Kang, Marco Körner, Yuanyuan Wang, Hannes Taubenböck, and Xiao Xiang Zhu. 2018. Building instance classification using street view images. *ISPRS journal of photogrammetry and remote sensing* 145 (2018), 44–59.

[21] Jennifer King, Deidre K. Mulligan, and Steven Raphael. 2008. *CITRIS Report: The San Francisco Community Safety Camera Program - An Evaluation of the Effectiveness of San Francisco's Community Safety Cameras.* Technical Report.

[22] Xiaojiang Li, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, and Weixing Zhang. 2015. Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry & Urban Greening* 14, 3 (2015), 675–685.

[23] Ming-Yu Liu, Shuoxin Lin, Srikumar Ramalingam, and Oncel Tuzel. 2015. Layered interpretation of street view images. *arXiv preprint arXiv:1506.04723* (2015).

[24] Dave Maass. 2019. The San Francisco District Attorney's 10 Most Surveilled Neighborhoods.

[25] IHS Markit. 2016. *Video surveillance: How technology and the cloud is disrupting the market.* Technical Report.

[26] Stephen J Mooney, Charles J DiMaggio, Gina S Lovasi, Kathryn M Neckerman, Michael DM Bader, Julien O Teitler, Daniel M Sheehan, Darby W Jack, and Andrew G Rundle. 2016. Use of Google Street View to assess environmental contributions to pedestrian injury. *American journal of public health* 106, 3 (2016), 462–469.

[27] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*. 4990–4999.

[28] Mutale Nkonde. 2019. Automated Anti-Blackness: Facial Recognition in Brooklyn, New York. *Harvard Kennedy School Journal of African American Policy* (2019), 30–36.

[29] OpenStreetMap contributors. 2017. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org.

[30] Rocco Parascandola. 2019. New NYPD surveillance cameras to cover stretch of Upper East Side not easily reached by patrol cars. https://www.nydailynews.com/new-york/nyc-crime/ny-metro-argus-cameras-east-20181024-story.html. Accessed: 2021-01-30.

[31] PreciseSecurity.com. 2020. Top 10 Countries and Cities by Number of CCTV Cameras. https://www.precisesecurity.com/articles/Top-10-Countries-by-Number-of-CCTV-Cameras. Accessed: 2021-01-30.

[32] Cooper Quintin and Dave Maass. 2015. License Plate Readers Exposed! How Public Safety Agencies Responded to Major Vulnerabilities in Vehicle Surveillance Tech.

[33] Arjun Raman. 2017. Cheers to Street View's 10th birthday! https://www.blog.google/products/maps/cheers-street-views-10th-birthday/.

[34] Gary C. Robb. 1980. Police Use of CCTV Surveillance: Constitutional Implications and Proposed Regulations. *University of Michigan Journal of Law Reform* 13, 3 (1980), 571–602.

[35] Raz Schwartz and Nadav Hochman. 2014. The social media life of public spaces: Reading places through the lens of geo-tagged data. , 52–65 pages.

[36] Hao Sheng, Xiao Chen, Jingyi Su, Ram Rajagopal, and Andrew Ng. 2020. Effective Data Fusion With Generalized Vegetation Index: Evidence From Land Cover Segmentation in Agriculture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[37] Jay Stanley. 2019. *The Dawn of Robot Surveillance: AI, Video Analytics, and Privacy.* Technical Report. American Civil Liberties Union.

[38] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[39] Brandon C. Welsh, David P. Farrington, and Sema A. Taheri. 2015. Effectiveness and Social Costs of Public Area Surveillance for Crime Prevention. *Annual Review of Law and Social Science* 11, 1 (2015), 111–130.

[40] Yang Xu, Alexander Belyi, Iva Bojic, and Carlo Ratti. 2018. Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Computers, Environment and Urban Systems* 72 (2018), 51–67.

[41] Weixing Zhang, Chandi Witharana, Weidong Li, Chuanrong Zhang, Xiaojiang Li, and Jason Parent. 2018. Using deep learning to identify utility poles with crossarms and estimate their locations from google street view images. *Sensors* 18, 8 (2018), 2484.