International Journal of Geo-Information

MDPI

Article

Deep Neural Networks and Kernel Density Estimation for Detecting Human Activity Patterns from Geo-Tagged Images: A Case Study of Birdwatching on Flickr

Caglar Koylu [1,*], Chang Zhao [1] and Wei Shao [2]

[1] Geographical and Sustainability Sciences, University of Iowa, Iowa City, IA 52242, USA; chang-zhao@uiowa.edu

[2] Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA; wei-shao@uiowa.edu

* Correspondence: caglar-koylu@uiowa.edu; Tel.: +1-319-335-0161

Received: 6 December 2018; Accepted: 16 January 2019; Published: 18 January 2019

Abstract: Thanks to recent advances in high-performance computing and deep learning, computer vision algorithms coupled with spatial analysis methods provide a unique opportunity for extracting human activity patterns from geo-tagged social media images. However, there are only a handful of studies that evaluate the utility of computer vision algorithms for studying large-scale human activity patterns. In this article, we introduce an analytical framework that integrates a computer vision algorithm based on convolutional neural networks (CNN) with kernel density estimation to identify objects, and infer human activity patterns from geo-tagged photographs. To demonstrate our framework, we identify bird images to infer birdwatching activity from approximately 20 million publicly shared images on Flickr, across a three-year period from December 2013 to December 2016. In order to assess the accuracy of object detection, we compared results from the computer vision algorithm to concept-based image retrieval, which is based on keyword search on image metadata such as textual description, tags, and titles of images. We then compared patterns in birding activity generated using Flickr bird photographs with patterns identified using eBird data—an online citizen science bird observation application. The results of our eBird comparison highlight the potential differences and biases in casual and serious birdwatching, and similarities and differences among behaviors of social media and citizen science users. Our analysis results provide valuable insights into assessing the credibility and utility of geo-tagged photographs in studying human activity patterns through object detection and spatial analysis.

Keywords: deep learning; convolutional neural networks; image object detection; computer vision; kernel density estimation; Flickr; birdwatching

1. Introduction

The availability and widespread use of large-scale, user-generated geo-tagged and time-stamped images offer a unique opportunity to capture human activity patterns. For example, Flickr (www.flickr.com), provides a proxy for capturing human recreational activities from local to global geographic scales [1–3]. Previous studies that utilize online photographs for inferring nature-based human activities have a diverse research focus. Some examples include, but are not limited to, identifying character of landscapes [4], land cover and land use [5], recreational demand in water resources [1], events and tourist hotspots [6–8], impact of rare species in tourism [9], common tourist trajectories [10,11], recreational visitation at national parks [3], the perceived aesthetic value of ecosystems [12,13], as well as the relationship between cultural ecosystem services and landscape features [14,15].

Detecting human activities from geo-tagged social media images is challenging. The most commonly used method for extracting semantics from images is to search the metadata of images generated by the user or the social media application. However, metadata often do not exist, or may portray inaccurate, insufficient, or hard to interpret representations of the image context. An alternative way to capture semantics from images is to use computer vision algorithms that extract information through analyzing the actual content of the image. Recent advances in high-performance computing and deep learning significantly improved efficiency and accuracy of computer vision algorithms in image classification and object detection. However, there are only a handful of studies that evaluate the utility of computer vision algorithms for studying large-scale human activity patterns [5,16,17]. Another major challenge for capturing human activities from geo-tagged images is the lack of ground-truth data for validating the utility of social media posts, and identifying the representativeness and biases that may result from a variety of factors such as date, time, location of posted content, user behaviors, and demographics. In order to assess the credibility and utility of social media sources, few studies compared social media data to those of traditional field surveys and modeling outcomes, and found strong correlation between user generated content and actual human activity patterns [1–3].

In this study, we introduce an analytical framework to: (1) assess the utility of a computer vision algorithm, YOLOv3 (You Only Look Once) [18], in capturing semantics by detecting objects in image content; and (2) identify spatial and temporal footprints of human activities, and better understand the biases and characteristics of social media images and user behaviors. Specifically, we used birdwatching as a case study to demonstrate the analytical framework, which consists of three steps. We first employed a computer vision algorithm, YOLOv3 (YOLO), which uses convolutional neural networks (CNN) to detect objects, i.e., birds, by analyzing the content of images such as color, shape, texture, and any other information related to the actual image itself. We then introduced verification and validation procedures to evaluate the accuracy and utility of the proposed framework for detecting birdwatching activities. For verification, we first compared the results of the computer vision algorithm to concept-based image retrieval, which was based on keyword search on image metadata such as textual description, tags, and titles of images. Second, in order to capture potential biases in geo-tagged images, we compared the patterns in birding activity generated using Flickr bird photographs with patterns identified using eBird data (http://eBird.org/content/eBird/), a freely-available, global, citizen science online bird observation dataset collected and maintained by the Cornell Lab of Ornithology and the National Audubon Society [19]. eBird data have been used in a wide variety of ornithological studies across broad spatial and temporal scales. eBird observations are expert-verified, and are considered to be highly accurate [19,20]. As such, comparing our assessment of birding activity using Flickr images to eBird data for the same period allowed us to identify the biases of each data source, and the potential utility of Flickr in capturing birdwatching behavior.

In the following subsections, we first discuss the limitations and biases inherent in geo-tagged social media photographs. Second, we introduce image object detection methods that allow capturing locational and thematic contexts from geo-tagged images. Finally, we discuss birdwatching as a human activity in order to provide context for our case study.

1.1. Biases of Citizen Science and Social Media Data

User-generated content from social media applications include various sources of bias such as user demographics and behavior, locations, topics, date, and time of produced content [21–23]. Different individuals have different motivations and preferences for sharing information, thus shared data may not be representative of a larger population or even a community. Aside from the issue of representativeness of general population demographics, user contribution bias [24] and urban-rural divide [22] pose significant challenges to accurate portrayal of human activity patterns from user-generated content. User contribution bias results from the fact that a large portion of the content is generated only by a few highly active users. On the other hand, the effect of urban-rural divide, or population density is also very evident in social media data such as Twitter, Flickr, and

Foursquare, as most of the content is produced in urban areas [25,26]. Urban-rural divide in social media usage generates a further limitation, the small area or number problem, which leads to spurious variation in patterns extracted from areas with low density of observations [27]. In order to address the varying density problem, researchers have used expectation surface using the Chi-statistic and density estimation [28]. In an expectation surface the number of observed photograph counts are compared to expected values derived from population density. To keep the most active users from dominating elicited spatial patterns, previous studies normalized the number of photographs by each user based on a threshold determined by distance [4], or distinct user count and density criteria [17]. Other social media-based indicators used for mapping recreational activities include Flickr photograph counts [9], the number of individual Panoramio users participating in specific activities [12], and Flickr generated user-days based on the number of photographs taken by individual users on unique days in a location [1,3,29].

1.2. Computer Vision Algorithms

Computer vision algorithms are used to extract semantics from image properties such as color, shape, texture, or any other information that can be derived from the image itself. Computer vision algorithms are different from concept or description-based image indexing [30,31], which searches keywords in image metadata such as title, tags, and descriptions to infer concepts and semantics from images. Deep learning, which allows extraction of high level abstractions in data by utilizing a hierarchical architecture [32], have been widely used in a variety of applications such as natural language processing, semantic parsing, transfer learning, and computer vision. Extracting semantics from images remains a significant challenge due to the semantic gap problem of extracting high-level semantic concepts from low-level image pixels captured by the deep learning algorithms. However, there has been a variety of successful applications of CNNs for addressing the semantic gap and extracting context from geo-tagged images. For example, Porzi et al. [33] employed a CNN architecture to capture peoples' perception of safety, attractiveness, and uniqueness using Google Street View Images. CNNs have also been widely used in image-based geo-localization and scene recognition. Geo-localization and scene recognition algorithms use image content to identify the location where the image was taken, as well as the characteristics of that location [16,34,35]. Similarly, Tracewski et al. [5] applied neural networks to identify land cover and land use classification of images obtained from various social media sources such as Flickr, Panoramio, Geograph, and Instagram. For a more in-depth discussion of the deep learning methods, readers may refer to Guo et al. [32], Wan et al. [36], and Yang et al. [37].

1.3. Birdwatching

Birdwatching is a non-consumptive outdoor recreational activity that arose in the early 1900s [38]. Birdwatching is a popular activity, with an estimated 46.7 million individuals participating in birding annually in the US [39], and over six million individuals observing birds at least every three weeks in the United Kingdom (CBI 2011). Most birders (88%) view birds around their homes, but many (38%) travel, often great distances, to birdwatch [39]. Birdwatching falls into several, overlapping categories based on expertise and motivation [40,41]. "Birding" is undertaken by hobbyist, professionals, or semi-professionals focused on studying and identifying birds. Birdwatching may also function as a sport through "listing", an often-competitive process whereby individuals maintain checklists of species they have observed. Birders may seek out rare species or species located outside of their typical range, watching blogs for reports of such species, and traveling great distances to add such birds to their lists. Such activities typically involve documenting observations, often via photographs shared through social media, and may involve posting sightings to citizen science applications such as eBird. These activities produce substantial economic impacts related to the purchase of equipment (e.g., telescopes, binoculars, cameras, and bird feeding supplies), and to travel and associated expenses (e.g., air fare, lodging, and dining).

Bird species may be migratory or resident and, if migratory, may be short- (i.e., moving within the same local area), medium- (e.g., moving among US states), or long-distance migrants (e.g., moving across or among continents). Migration follows seasonal variation in resource availability and environmental conditions, and includes movements between breeding and wintering locations in the fall and spring seasons [42]. During migration, birds utilize stopover sites of varied habitat quality for resting and refueling. Stopover sites of high habitat quality may serve numerous species in relatively high densities. These sites may be particularly interesting to birdwatchers, especially if they are publicly accessible (e.g., parks and wildlife refuges) for observing many transitory species in one location. As a result of the competitive nature of listing, social media and citizen science postings may highlight more rare and transient birds during migration at stopover hotspots.

2. Materials and Methods

Figure 1 illustrates the analytical framework proposed in this article. We first collected all geo-located Flickr image metadata and images, and eBird observations given the photograph-taken and observation date, respectively, and the border of conterminous US across a three-year period between December 2013 and December 2016. Flickr metadata consisted of attributes that identify the photograph by id, name and identification number of the user, the location where the photograph was taken (i.e., longitude and latitude coordinates that are either manually added by users or generated by camera/smartphones), the time and date on which the photograph was taken and uploaded, and textual annotations provided by users and the application, including tags, description and title of photo contents. On the other hand, the eBird basic dataset (EBD) was a freely-available, global citizen science online bird observation dataset, collected and maintained by the Cornell Lab of Ornithology and the National Audubon Society (http://eBird.org/content/eBird) [19]. eBird data contained the name, counts, and types of the species observed during a single search event, the location where the search took place, the time, date, and duration of the search, as well as the name of the observer.

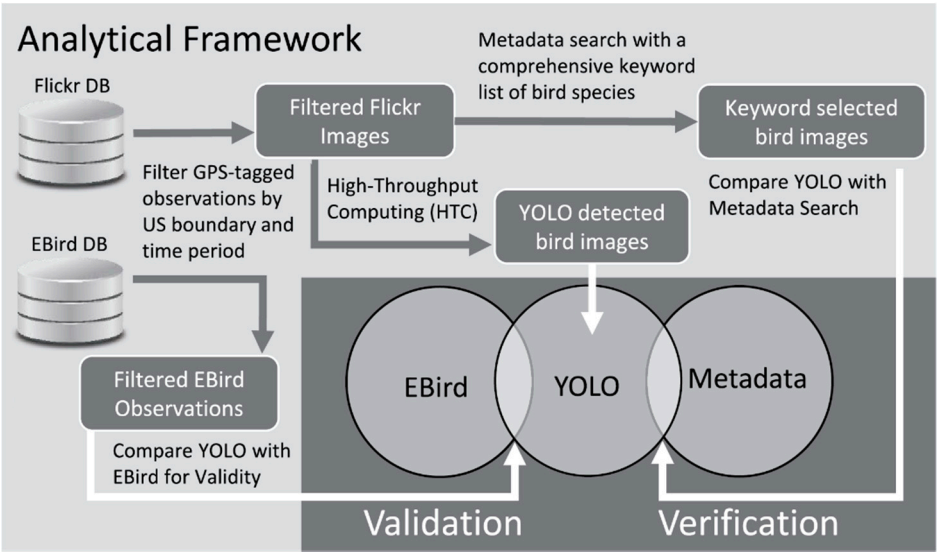

Figure 1. Overview of the analytical workflow.

After the initial download, we filtered both datasets to include only the geo-tagged Flickr images and eBird observations that were within the conterminous US. In the next steps, we extracted bird images using the metadata keyword search and YOLO deep learning library. We first compared the accuracy of objection detection by YOLO with metadata search as well as their spatial patterns. Second, we extracted spatial patterns of birdwatching activity from eBird, and compared them with YOLO-detected Flickr bird photographs in order to identify the similarities and differences between eBird and Flickr in capturing birdwatching behaviors.

2.1. You Only Look Once (YOLO)

YOLO is a state-of-the-art, unified, real-time objection detection system [18,43,44]. Unlike other object detection approaches such as deformable parts model (DPM) [45] and R-CNN [46], YOLO frames object detection as a regression problem. During training, the following sum of squared error function was minimized [44]:

$$\begin{aligned} &\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + \left(\sqrt{w_i} - \sqrt{\hat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2 \right] \\ &+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} \left(C_i - \hat{C}_i\right)^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{noobj} \left(C_i - \hat{C}_i\right)^2 \\ &+ \sum_{i=0}^{S^2} \mathbb{I}_{i}^{obj} \sum_{c \in classes} \left(p_i(c) - \hat{p}_i(c)\right)^2 \end{aligned} \tag{1}$$

where (x_i, y_i), w_i, h_i, and C_i represent the center, width, height, and class value, respectively, of the bounding box relative to the grid cell i; $p_i(c)$ represents the probability that the object in the grid cell i belongs to class c; $(\hat{x}_i, \hat{y}_i)$, $\hat{w}_i$, $\hat{h}_i$, and $\hat{C}_i$ represent the center, width, height, and class value, respectively, of the training object that falls into the grid cell i; $\hat{p}_i(c)$ represents the probability that the training object that falls into the grid cell i belongs to class c; $\mathbb{I}_{ij}^{obj}$ denotes if object appears in the cell i; and $\mathbb{I}_{ij}^{noobj}$ denotes that the jth bounding box predictor in cell i is "responsible" for that prediction; λ_{coord} and λ_{noobj} are weights for localization error and classification error, respectively.

In YOLO, a single neural network is used to simultaneously predict multiple object bounding boxes and the corresponding class probabilities directly from image pixels (Figure 2). Redmon et al. [44] compared the mean average precision (mAP) and frames per second (FPS) of YOLO to other detection algorithms using validation data sets from PASCAL VOC 2007. After evaluating the reported average precision and efficiency of other object detection algorithms including 30Hz DPM [47], Fastest DPM [48], R-CNN Minus R [49], Fast R-CNN [48], Faster R-CNN VGG-16 [50], Faster R-CNN ZF [50], we decided to use YOLO to detect birds in images.

Figure 2. An example of YOLOv3 (You Only Look Once) objection detection result.

The initial 24 convolutional layers of the network extract features from the image, and the two fully connected layers predicted the output bounding boxes and class probabilities. The output of the system was stored as an S × S × (B * 5 + C) tensor. A typical network architecture with S = 7, B = 2, C = 20 is shown in Figure 3 [44].

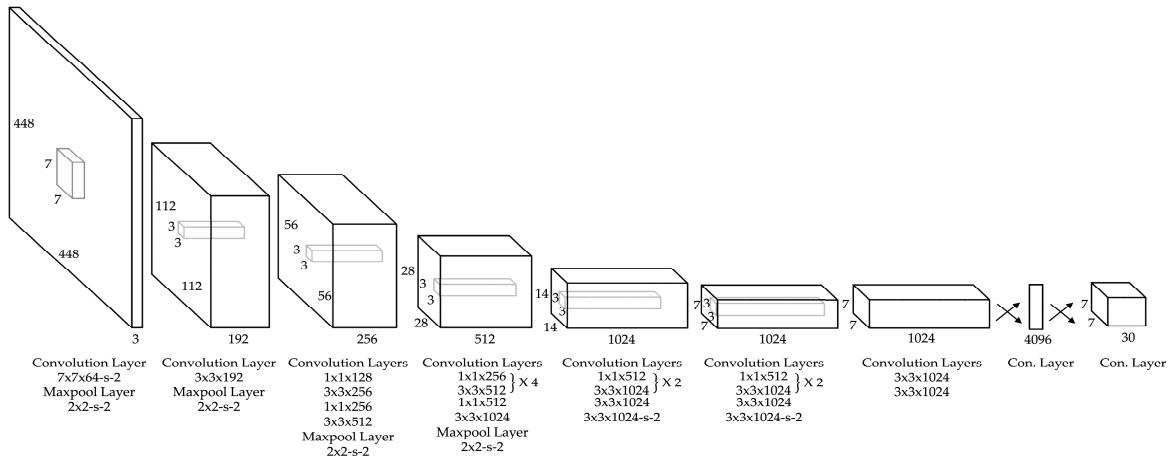

Figure 3. YOLO network architecture (adapted from [44]).

We experimented with the latest version of YOLO, YOLOv3, which took approximately 15 s on the Central Processing Unit (CPU). The total number of images to be processed was ~20 million. We performed high-throughput computing (HTC) on the Argon HPC system at The University of Iowa (UI). The Argon HPC system consists of 366 compute nodes with a range of 40 to 56 cores per node. Due to the substantially larger numbers and availability of CPU units, we performed HTC on CPU nodes. Argon uses Son of Grid Engine (SGE) queuing system for job submissions. Argon system has a limit of 10,000 active jobs per user, which includes currently running and pending jobs waiting to be submitted. We used array jobs to submit YOLO objection detection tasks. An array job consists of identical tasks ordered by a range of index numbers. The maximum amount of tasks an array job can handle is 75,000. Given the user and array job limits, we submitted 267 array jobs to process ~20 million images, which took approximately a week.

2.2. Kernel Density Estimation

In order to compare the spatial distributions of images identified by YOLO and metadata search, and eBird and YOLO-detected observations, we perform two types of kernel density estimation (i.e., fixed-distance and adaptive bandwidth), which were both performed using the formula below [51]:

$$f_h(G_i) = \frac{1}{nh} \sum_{i=1}^{x \in O_i} K\left(\frac{x - x_i}{h}\right) \quad (2)$$

where $x_1, x_2, \ldots, x_n \in O_i$ is a set of observation locations within the bandwidth (neighborhood) of G_i; n is the total number of observations; x is the location of estimation; K is the kernel function; and h is the bandwidth (the distance). Commonly used kernel functions are Uniform, Epanechnikov, Triangular, and Gaussian. The choice of kernel function often does not change the density estimation result. However, the bandwidth h is a key parameter that determines the outcome of fixed-distance kernel estimation. This same formula is applied to both fixed-distance and adaptive kernel density estimation (smoothing). Kernel smoothing can be performed on regular grids or spatial units of different aggregation levels (e.g., counties, census tracts, and block groups). In this study, we divided the study area into a grid of 5 miles (8 km) resolution, which covered the conterminous US. We used this resolution for the fixed-kernel density estimation for comparing spatial patterns of YOLO detection to metadata search, and eBird to YOLO-detected observations. Fixed-distance kernel smoothing requires a distance threshold to determine the bandwidth. We determined the bandwidth as 20 miles, which is approximately equivalent to the second order immediate neighborhood of a grid cell.

Fixed-distance spatial filters in kernel density estimation often result in the loss of geographic detail when the density of observation is much higher. Moreover, smaller filters produce unreliable

estimates in areas with sparse observations. Different from fixed-distance kernel density estimation, adaptive kernel density estimation (smoothing) [52] is a non-parametric method that uses local information in neighborhoods defined by varying kernel sizes to estimate values of specified features at given locations. Adaptive kernel smoothing requires a minimum number of observation threshold (k-nearest observations) to determine the bandwidth, and the observations within the neighborhood of estimation as well as their spatial weights. While fixed-distance kernel allowed us to compare absolute differences between the two datasets, it did not address the user contribution bias and the small area problem, which produced unreliable estimates for areas with lower density of observations.

In addition to using fixed-distance kernel density estimation, we employed an adaptive kernel density estimation in order to derive a smoothed rate of Flickr and eBird observations based on the number of users. Due to the fact that the number of observations together amount to a large dataset with 125 million eBird observations, and 750,000 YOLO-detected bird images, and there were substantial numbers of observations that shared exact coordinates, we introduced an algorithm for efficient computation of the adaptive kernel estimation based on the number of distinct user locations. Definitions and steps of the adaptive kernel smoothing are defined below. In Step 1, the area was divided into a grid of 5 miles (8 km) resolution, which was the same resolution used in fixed-distance kernel smoothing. There was a substantial amount of observations with the same coordinates. In Step 2, we aggregated observations that shared the exact coordinates into distinct observation locations prior to determining the k-nearest users. As a result, we obtained a list of distinct observation locations that included information on the total number of observations and the list of users for both eBird and Flickr. We define k as the minimum number of users to determine the neighborhood. Given a positive neighborhood size threshold k based on the number of users, a k-size neighborhood is derived for each grid $G_i \in G$, which is the smallest k-nearest-neighbors of G_i that meets the size constraint. In Step 3, we employed a Sort-Tile-Recursive Tree algorithm to compute a spatial index of the k-nearest distinct users and their locations for each grid cell in order to improve the computational efficiency for determining the k-nearest distinct users and their locations. We set the threshold k as a combined number of 100 Flickr and eBird users. Once the neighborhood reached the defined threshold k, we determined the list of observations, O_i, bandwidth h (G_i, k), and the weights of distinct observation locations for each grid cell in Step 3. K is the kernel function, and h is the bandwidth for smoothing. Kernel functions determine the weight of each observation within a kernel, and the choice of function often does not have substantial impact on the result. The most commonly used kernel functions are Uniform, Epanechnikov, Triangular, and Gaussian. In this study, we employed the uniform kernel to simplify interpretation of the estimation. Given the list of observations, the spatial weights, and the count of observations for eBird and Flickr, we computed a continuous surface that took into account the number of distinct users in each kernel defined in Step 5.

Definitions:

G: Grid: the total set of grid cells that covers the study area.
G_i: Grid cell i. $G_i \in G$.
k: Adaptive filter (neighborhood) threshold based on the total number of distinct users.
U_i: The list of users within the neighborhood of G_i.
O_i: The list of observations within the neighborhood of G_i.
h (G_i, k): The bandwidth of the k-Size Neighborhood of the grid cell G_i is defined as the smallest KNN (G_i, k) = $\{Gj \in G\}$ that has a total count of distinct users: $\sum k_i \geq k$.
K: Kernel function. Uniform function is used for simple interpretation of the results.

Steps:

(1) Compute G, the grid of the study area given a resolution r. In this study, r = 8 km was used.
(2) Aggregate observation statistics such as the number of observations and keep a list (hash) of users for each distinct observation location for both Flickr and eBird.

(3) Given k = 100, compute a spatial index based on Sort-tile-recursive (STR) tree for finding the k-nearest Flickr and eBird users for each grid-cell.
(4) Determine O_i, h (G_i, k), and the weights of observations for each grid-cell using the adaptive kernel estimation.
(5) Compute the percentage of YOLO-detected Flickr images to eBird observations for each grid-cell.

3. Results and Evaluation

Between December 2013 and December 2016, there were 19,711,242 geo-tagged Flickr images within the conterminous US. Table 1 illustrates the top 48 objects detected by YOLO, and the number of images that contain at least one of these objects. These objects were used to infer human activities as well as environmental characteristics of locations that the photographs were taken at. For example, the presence of bicycles may be useful to quantify biking behavior, sports ball may indicate sports activities, and objects such as sofa, bed, vase, and chair may indicate indoor activities. In this article, we focused on only bird images, and used birdwatching activity as a case study to demonstrate the utility of our analytical framework. The object "bird" was the 5th most frequent object detected in 747 thousand images.

Table 1. The number of images that contain at least one of the top 48 detected objects.

Object	Image Count	Object	Image Count	Object	Image Count
person	8,309,891	tie	277,644	motorbike	139,648
car	2,080,796	traffic light	275,140	sofa	135,411
chair	870,726	train	261,775	cell phone	135,301
truck	764,307	sports ball	259,229	book	133,518
bird	747,015	bench	258,984	bus	130,706
diningtable	452,627	handbag	243,079	cat	129,015
cup	428,193	pottedplant	222,407	horse	125,918
aeroplane	365,747	tvmonitor	219,358	wine glass	117,165
bottle	352,077	bowl	199,816	bed	108,157
dog	317,291	backpack	191,202	cake	104,080
boat	296,850	umbrella	180,092	baseball glove	97,638
bicycle	287,272	clock	171,435	vase	86,496

We organized the results and evaluation under two subsections: verification and validation. We first present our comparative evaluation of metadata search and YOLO to verify the accuracy of both approaches. Second, we compare YOLO-detected birding activity to EBird observations to evaluate the validity and biases of Flickr and eBird data in inferring birdwatching activity.

3.1. Verification

Our objective for verification was to answer the following questions:

- Is object detection more accurate than metadata search for capturing bird images on Flickr?
- Are there any spatial and temporal biases between the results of metadata search and YOLO object detection?

While we detected 747,015 (3.8%) images with birds using YOLO, we detected 534,121 (2.7%) images that contain bird keywords with metadata-based search. Table 2 represents the temporal variability in the detection of birds by metadata search and YOLO. Overall, YOLO allowed increased detection of birds over 50% of what metadata search could detect, and this increase was consistent across different seasons. There was a substantial increase of over 1% in the detected number of bird photographs when using YOLO as compared to the number images captured by the metadata search. Among the 19.7 million images, both YOLO and metadata searches commonly detected birds in 409,779 (2%) images. Since both methods detected birds in these images, we considered the classification as

accurate. In order to identify the mismatch between the two methods, we further compared images detected only by YOLO and those only by metadata search. YOLO detected an additional 1.8% bird images, which were not detected by metadata search. On the other hand, metadata search detected only 0.7% additional images with bird keywords, which were not detected by YOLO.

Table 2. Bird photograph statistics.

Time Periods	Metadata	%	YOLO	%
Winter 2013	44,718	3.31	61,117	4.58
Spring 2014	54,558	2.54	80,164	3.77
Summer 2014	41,216	1.89	65,787	3.05
Autumn 2014	39,243	2.19	57,723	3.28
Winter 2014	43,850	3.49	58,707	4.68
Spring 2015	57,064	3.18	75,026	4.23
Summer 2015	43,901	2.31	61,147	3.45
Autumn 2015	37,279	2.39	49,779	3.23
Winter 2015	38,759	3.74	48,875	4.74
Spring 2016	48,711	3.44	62,816	4.48
Summer 2016	44,945	3.14	66,028	4.22
Autumn 2016	44,877	3.45	59,846	4.33

We assessed the accuracy of bird-detected images by only YOLO and only metadata search, using human classification by the first author. We defined the human classification task with the question: "Is there a real bird in this photograph?" We used a random sample of 1000 bird photographs detected by only YOLO, or by only metadata search. According to the accuracy testing, bird images classified only by the metadata search but not with YOLO resulted in a substantially lower accuracy of 26%, while bird images detected only by YOLO resulted in an accuracy of 89%. Although our sample size for human classification was low at this point, this finding confirmed the increased accuracy of YOLO detection.

Although YOLO detection had an accuracy of 89% in classifying birds, Figure 4 represents a variety of sample cases of accurate and inaccurate classifications of YOLO. Figure 4a,b,d represent accurate classifications of birds. The algorithm detected two birds in Figure 4a with an estimated accuracy of 60% and 59%, although there were obviously more birds (five) in this image. However, since the image was not tagged with bird keywords it was not captured by the metadata search. The algorithm also detected a bench with 60% accuracy, although there were multiple benches in this photograph. Both birds in Figure 4b were accurately detected by 85% and 80% estimated accuracy, and the bird in Figure 4d was accurately detected by an accuracy of 98%. Although the rest of the images in Figure 4c,e,f do not contain birds, they were inaccurately classified by YOLO as containing birds. The shape of a butterfly in Figure 4c and the shape of the flowers resemble features of a bird such as the wings, neck, and beak, which possibly led to misclassification. However, classification accuracy for these two images were low, 54% and 51%, respectively. As we did not include a threshold, we included any classification regardless of the probability value provided by YOLO. Finally, Figure 4f contains a realistic drawing of a hummingbird, which was classified as a bird by YOLO. This classification illustrates the case where classification is algorithmically accurate, but semantically inaccurate as the purpose is to identify real birds.

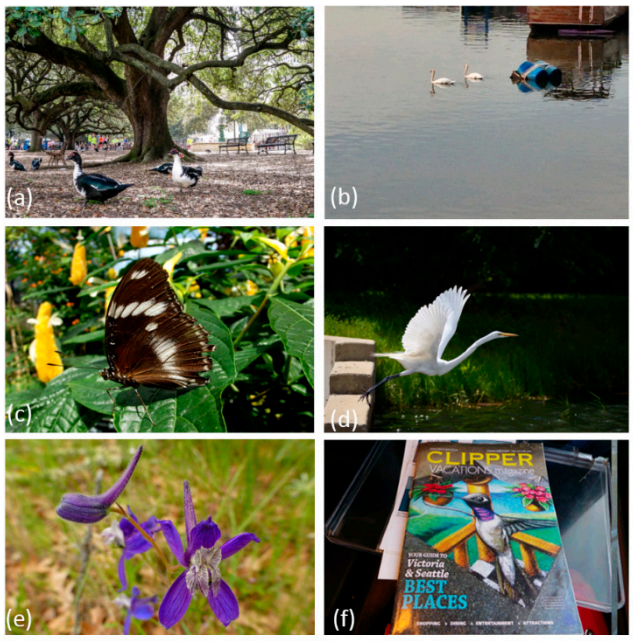

Figure 4. Only YOLO detected bird photographs.

Figure 5 represents bird images detected by only the metadata and not YOLO. Figure 5a is an accurate classification of a woodpecker, a common bird species thanks to the title of this image "Acorn woodpecker". YOLO algorithm was not able to detect the bird in this photograph because of how the bird had blended well with the tree branch, which concealed the major features of the bird for objection detection. On the other hand, Figure 5b–d do not contain real birds but they commonly contain "bird" "keywords".

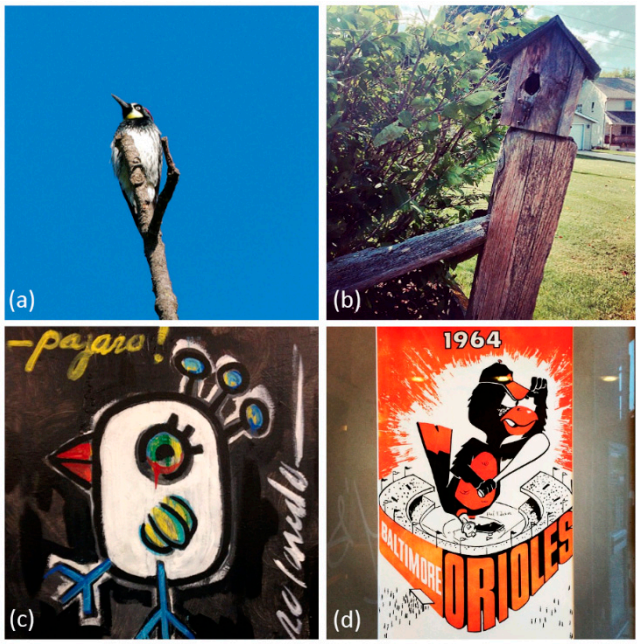

Figure 5. Only metadata detected bird photographs. (**a**) Acorn Woodpecker, (**b**) @fence #birdhouse #wood #lynnfriedman, (**c**) New piece! One if the largest paintings I have done of my birds! 30″ × 40″, and (**d**) #birdland #masnorioles.

A comparison of the density of bird images obtained from metadata search and YOLO are shown in Figure 6. We combined the counts of observations of YOLO and keyword search data, and employed

natural breaks classification to determine the class breaks in Figure 6. While the areas of bird images detected by YOLO and metadata search substantially overlapped, YOLO identified more bird images than metadata search for most of the study area. YOLO identified a much larger quantity of bird pictures in urban areas such as New Orleans, San Francisco, New York, Washington D.C., and Seattle. Moreover, YOLO results represent the continuity of bird habitat regions in coastal areas of Florida, the North East, Lake Michigan, and California. On the other hand, the density of bird images detected by metadata search produced more fragmented spatial patterns across the nation.

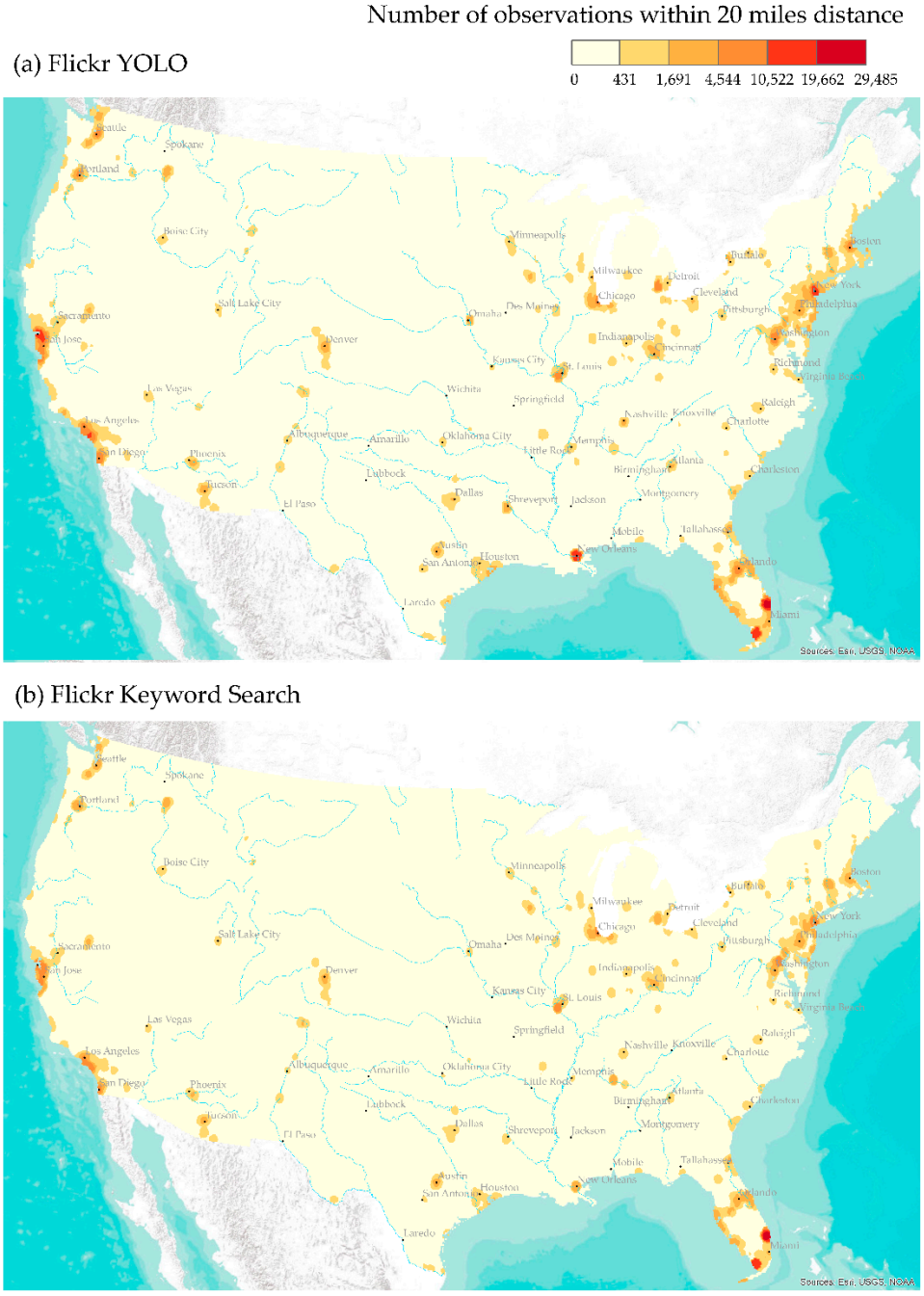

Figure 6. Fixed-distance (20 miles) density of (**a**) YOLO-detected and (**b**) keyword search Flickr bird image counts.

3.2. Validation

Our objective for validation was to answer the following questions:

- To what extent can Flickr be used to infer birdwatching as a human activity pattern?
- Are there any spatial and temporal biases between YOLO-detected birding activities and eBird observations?

To answer these questions, we compared YOLO-detected Flickr bird image statistics with eBird observations. We first calculated Spearman Rank correlation based on the fixed distance counts of observations and distinct users. We found a strong correlation between the count of eBird observations and YOLO-detected Flickr bird images with a correlation coefficient of 79%. Moreover, the count of distinct eBird users and Flickr users produced even a larger coefficient of 85%. These values indicated the strong overlap between eBird observations and Flickr bird images. We then compared the temporal patterns of Flickr image, user, and image-to-user ratios with eBird observation, user and observation-to-user ratios (Figure 7). Overall, Flickr had a declining trend from 2013 to 2016 both in terms of the number of bird photographs and users. This decline was also consistent with the overall decline in Flickr usage. On the contrary, eBird observations and users exhibited an increasing trend over the three-year period. Both Flickr and eBird photographs and users statistics peaked in spring months. Photograph-to-user ratio had an increasing trend for Flickr. On the other hand, eBird observation-to-user ratio was very consistent across the three-year period and peaked around spring and summer months.

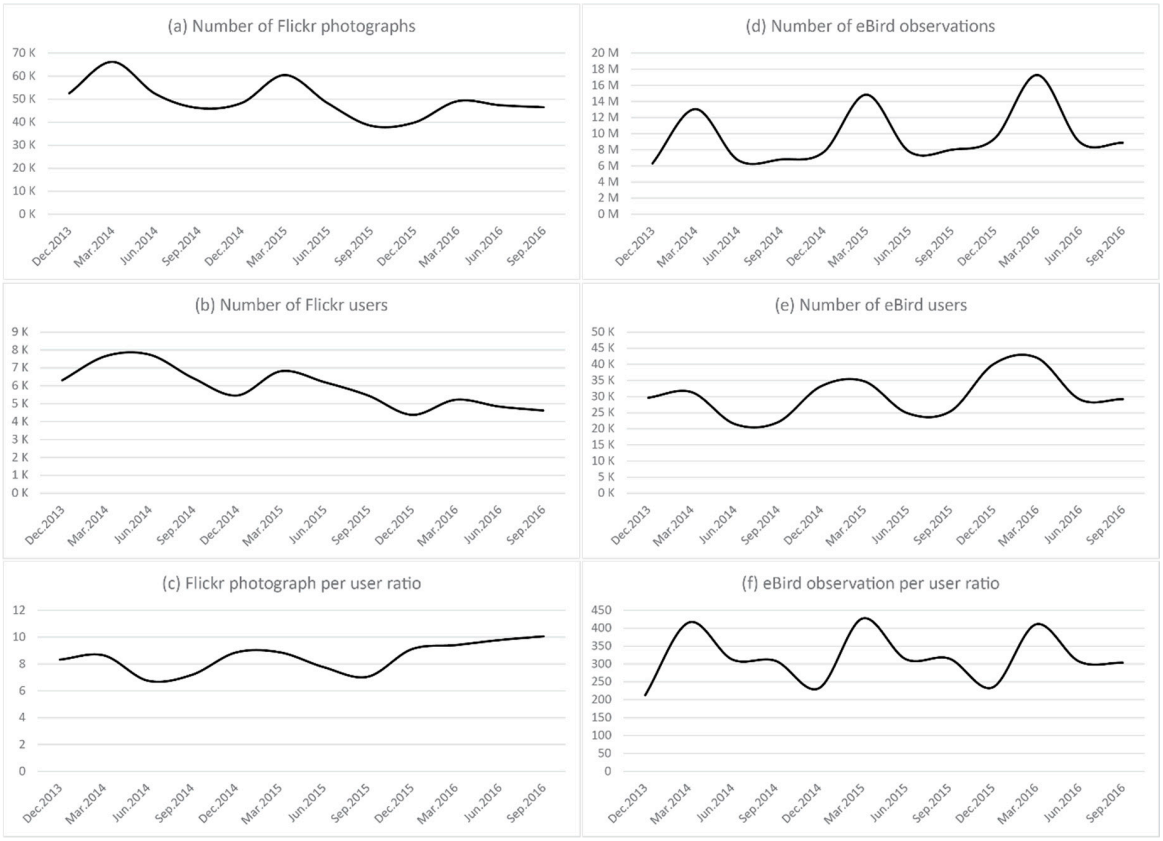

Figure 7. Temporal patterns of YOLO-detected Flickr bird images and eBird observations.

Between December 2013 and December 2016, there were 125,179,161 eBird observations within the bounding box of the conterminous US. Among these observations, 115,682,223 observations were exactly within the conterminous US. There were only 1,422,554 distinct coordinates, which corresponds to 1% of eBird observations in the conterminous US. This was mostly due to the multiple observations made from the same site throughout the day. Among 746,998 Flickr bird images, 346,549 images had distinct coordinates (46%), while the rest of the 54% of the images had coordinates that repeated more

than once. This was also a result of the same user's, or even in rare cases, multiple users' sharing of multiple images from the same coordinates (e.g., habitat observation towers). We attributed this pattern to Flickr users' casual birdwatching behavior as compared to eBird users' serious birdwatching activity.

In order to identify the spatial variation among Flickr bird images and eBird observations, we compared kernel density estimates of YOLO and eBird observations with a fixed-distance threshold of 20 miles (Figure 8). We observed an increased dispersion of the spatial distribution of eBird observations, which can be attributed to the fact that eBird had approximately 167 times more observations than Flickr bird photographs, and approximately 3.7 times more users than Flickr users who took bird photographs. We computed the z-scores for both YOLO and eBird observations in order to compare the two different distributions in which eBird observations had much higher density than YOLO-detected Flickr photographs. We combined the z-scores of the two dataset, and employed natural breaks classification to determine the class breaks for Flickr and YOLO maps in Figure 8. From Figure 8, we confirmed that the spatial distribution of eBird observations and Flickr photographs were similar to each other except few areas in which the magnitude and spatial extent of eBird and Flickr observations showed substantial differences. Both datasets indicated that high birdwatching activities take place around coastal areas and populous regions adjacent to metropolitan areas. While spatial patterns of birdwatching were similar between the two datasets, eBird was relatively more prominent in coastal areas of the North East, South East, West, Gulf Coast, and Great Lakes; national forests, prairie grass lands, wetlands, and areas where there was infrastructure for human access and birdwatching. While the magnitude of eBird density was much higher than Flickr across the nation, Flickr was relatively more prominent around urban areas such as New Orleans, Miami and Detroit.

Figure 9 illustrates the percentage of YOLO-detected Flickr bird images among both Flickr and eBird observations. This figure represents the bi-polar ratio of Flickr to eBird, and highlight areas where YOLO-detected Flickr photographs are above 1% using adaptive kernel smoothing that employs the 100 nearest users (both Flickr and eBird) to identify the neighborhood in the smoothing parameter. Figure 9 highlights prominent areas of Flickr bird photographs in natural lands that provide nesting, stopover, and overwintering habitat for birds. Interestingly, the spatial patterns were very distinct and different from fixed-distance density distribution, and provided a valuable input where Flickr usage was relatively higher in comparison to eBird. Regardless of the difference between the number of observations between Flickr and eBird, Flickr bird photographs were prominent (over 10%) in areas where there was access and infrastructure for birdwatching across the nation. The example areas where Flickr bird photographs were relatively higher are Grand Canyon River and Colorado Plato, Yellow Stone National Park, Southern Colorado, national preserves and wildlife areas in Southern Florida, and the wetland and prairie lands in the Mid-West. These Flickr users likely represent tourists who are not serious birdwatchers.

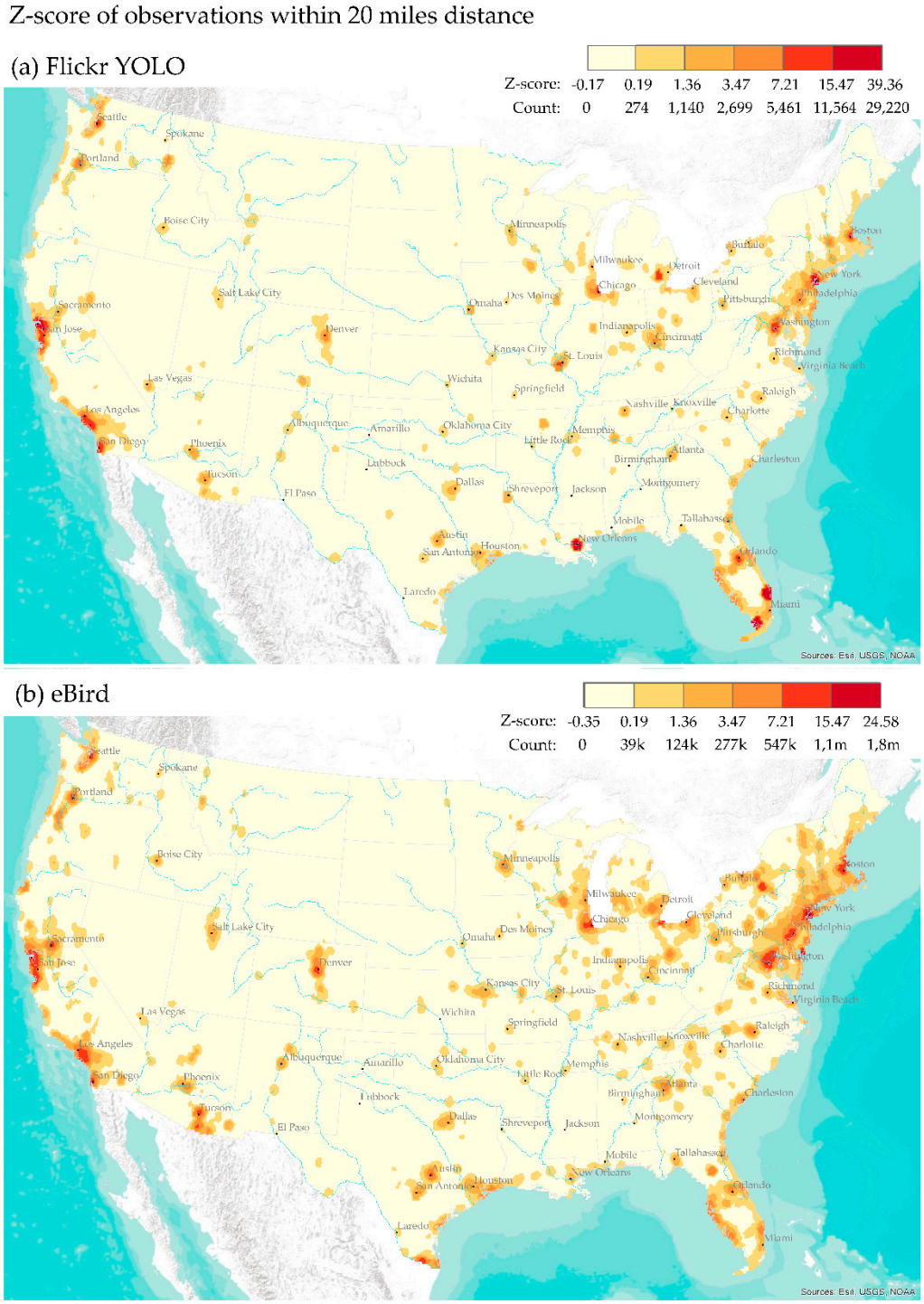

Figure 8. Z-scores of the fixed-distance (20 miles) density of (**a**) YOLO-detected Flickr bird image and (**b**) eBird observations.

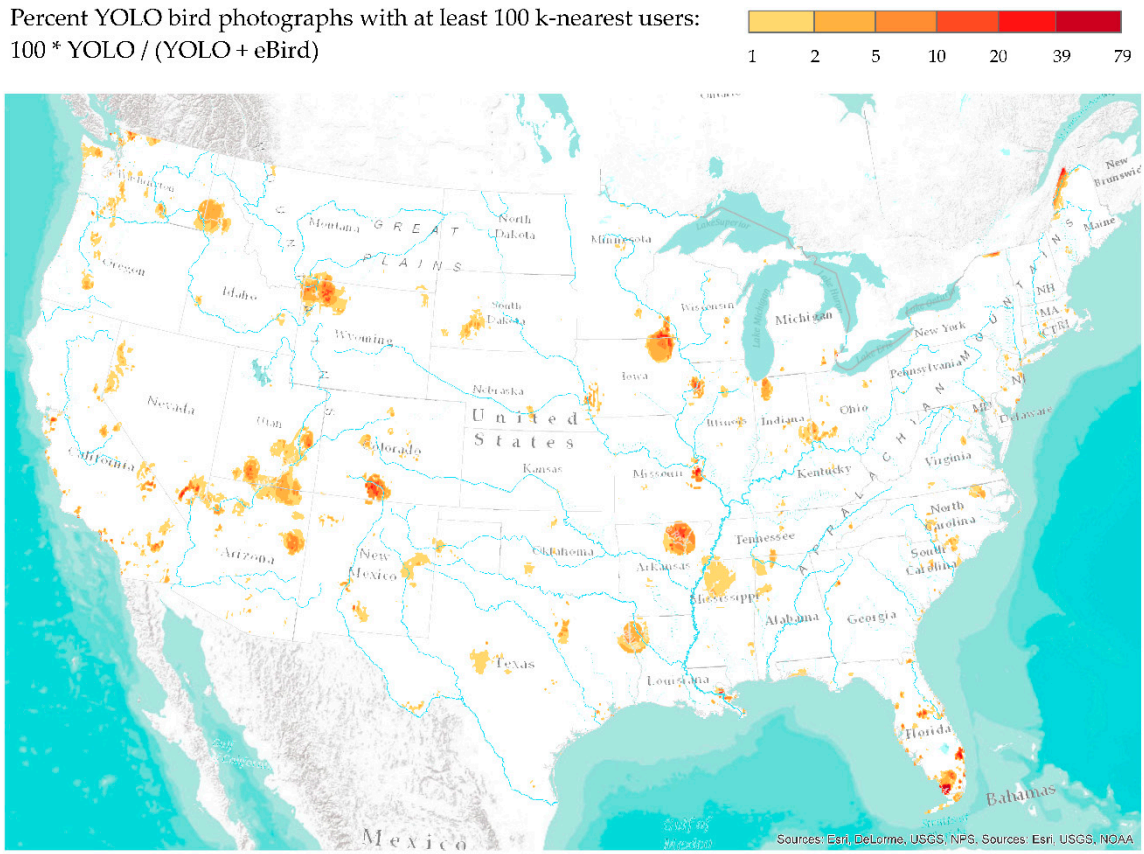

Figure 9. Percent of YOLO-detected Flickr bird images computed by an adaptive kernel based on a minimum threshold of 100 users that contain both Flickr and eBird users.

4. Discussion and Conclusions

In this article, we introduced an analytical framework that integrates a computer vision algorithm based on convolutional neural networks (CNN) with kernel density estimation to identify objects and infer human activity patterns from geo-tagged photographs. To demonstrate our framework, we inferred birdwatching activity by detecting birds from approximately 20 million publicly shared images on Flickr across a three-year period from December 2013 to December 2016. Our comparisons of Flickr and eBird observations highlight behavioral differences among the social media and citizen science users, which we further attribute to casual (Flickr) and serious birdwatching (eBird).

We have shown how the computer vision algorithm, YOLO, can be used for detecting objects and extracting semantics from geo-tagged and time-stamped social media images. Bird images classified only by the metadata but not with YOLO resulted in a substantially lower accuracy of 26%, while bird images detected by only YOLO resulted in an accuracy of 89%. Our case study in birdwatching, and comparisons of patterns captured from Flickr with the patterns from eBird observations highlight the biases in social media and citizen science data sets. While eBird helps identify serious birdwatching behaviors that are focused in particular areas across the US, Flickr patterns suggest more casual and spatially diverse birdwatching activities. The results of our analysis provide valuable insights into the credibility and utility of geo-tagged photos in studying birdwatching activities, and show the potential for studying other human activity patterns through object detection using a large collection of geo-tagged and user-generated images.

While eBird data have been used in a wide variety of ornithological studies across broad spatial and temporal scales [19,20], and the data source has a number of significant biases due to a variety of reasons such as users, locations, and time periods. For example, while in the earlier periods citizen scientists collected information from a diverse set of species, in recent years citizen scientists have

been biased towards collecting information on threatened species and protected areas [53]. On the other hand, Flickr users are usually photographers who are also birdwatchers, who not only upload their images, but also decide to geotag and share them. Our results comparing the spatial distribution of the two datasets highlight similar results as well as some geographic variations, which can be attributed to the potential biases among citizen science and social media applications and users. While eBird users are more likely to travel long distances for bird observations, Flickr users are casual birdwatchers who are likely to take bird photographs around their usual activity spaces. Future studies on extracting the mobility patterns of eBird and Flickr users can help better understand the dynamics of birdwatching activities.

In future work, we plan to complete the accuracy evaluation of all images classified by both the metadata and YOLO deep learning library. In addition, we plan to evaluate other object detection libraries and compare the accuracy results with YOLO. Beyond the scope of our particular focus on birdwatching, we plan to identify characteristics of locations based on the objects detected in an area over a period of time. This way, we will examine whether object detection can be used to advance our understanding of places, and semantics embedded in those places, and identify similarities between places across the world.

Author Contributions: Conceptualization, C.K., C.Z., and W.S.; methodology, C.K., C.Z., and W.S.; software, C.K., C.Z., and W.S.; validation, C.K., C.Z., and W.S.; formal analysis, C.K., C.Z., and W.S.; investigation, C.K., C.Z., and W.S.; resources, C.K., C.Z., and W.S.; data curation, C.K., C.Z., and W.S.; writing—original draft preparation, C.K., C.Z., and W.S.; writing—review and editing, C.K., C.Z., and W.S.; visualization, C.K.; supervision, C.K.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank Dan Holstad, Glenn Johnson, Sai Kumar Ramadugu, and Ben Rodgers for their technical support for data maintenance and the use of the Argon System.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Keeler, B.L.; Wood, S.A.; Polasky, S.; Kling, C.; Filstrup, C.T.; Downing, J.A. Recreational demand for clean water: Evidence from geotagged photographs by visitors to lakes. *Front. Ecol. Environ.* **2015**, *13*, 76–81. [CrossRef]
2. Sessions, C.; Wood, S.A.; Rabotyagov, S.; Fisher, D.M. Measuring recreational visitation at US National Parks with crowd-sourced photographs. *J. Environ. Manag.* **2016**, *183*, 703–711. [CrossRef] [PubMed]
3. Wood, S.A.; Guerry, A.D.; Silver, J.M.; Lacayo, M. Using social media to quantify nature-based tourism and recreation. *Sci. Rep.* **2013**, *3*, 2976. [CrossRef] [PubMed]
4. Dunkel, A. Visualizing the perceived environment using crowdsourced photo geodata. *Landsc. Urban Plan.* **2015**, *142*, 173–186. [CrossRef]
5. Tracewski, L.; Bastin, L.; Fonte, C.C. Repurposing a deep learning network to filter and classify volunteered photographs for land cover and land use characterization. *Geo-Spat. Inf. Sci.* **2017**, *20*, 252–268. [CrossRef]
6. Kisilevich, S.; Krstajic, M.; Keim, D.; Andrienko, N.; Andrienko, G. Event-based analysis of people's activities and behavior using Flickr and Panoramio geotagged photo collections. In Proceedings of the 2010 14th International Conference Information Visualisation (IV), London, UK, 26–29 July 2010; pp. 289–296.
7. Rossi, L.; Boscaro, E.; Torsello, A. Venice through the Lens of Instagram: A Visual Narrative of Tourism in Venice. In Proceedings of the Companion of the Web Conference, Lyon, France, 23–27 April 2018; pp. 1190–1197.
8. Lee, J.Y.; Tsou, M.-H. Mapping Spatiotemporal Tourist Behaviors and Hotspots Through Location-Based Photo-Sharing Service (Flickr) Data. In Proceedings of the LBS 2018: 14th International Conference on Location Based Services, Zurich, Switzerland, 15–17 January 2018; pp. 315–334.
9. Willemen, L.; Cottam, A.J.; Drakou, E.G.; Burgess, N.D. Using social media to measure the contribution of Red List species to the nature-based tourism potential of African protected areas. *PLoS ONE* **2015**, *10*, e0129785. [CrossRef] [PubMed]
10. Jankowski, P.; Andrienko, N.; Andrienko, G.; Kisilevich, S. Discovering landmark preferences and movement patterns from photo postings. *Trans. GIS* **2010**, *14*, 833–852. [CrossRef]

11. Yang, L.; Wu, L.; Liu, Y.; Kang, C. Quantifying Tourist Behavior Patterns by Travel Motifs and Geo-Tagged Photos from Flickr. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 345. [CrossRef]
12. Casalegno, S.; Inger, R.; DeSilvey, C.; Gaston, K.J. Spatial covariance between aesthetic value & other ecosystem services. *PLoS ONE* **2013**, *8*, e68437.
13. Figueroa-Alfaro, R.W.; Tang, Z. Evaluating the aesthetic value of cultural ecosystem services by mapping geo-tagged photographs from social media data on Panoramio and Flickr. *J. Environ. Plan. Manag.* **2017**, *60*, 266–281. [CrossRef]
14. Gliozzo, G.; Pettorelli, N.; Haklay, M. Using crowdsourced imagery to detect cultural ecosystem services: A case study in South Wales, UK. *Ecol. Soc.* **2016**, *21*, 6. [CrossRef]
15. Oteros-Rozas, E.; Martín-López, B.; Fagerholm, N.; Bieling, C.; Plieninger, T. Using social media photos to explore the relation between cultural ecosystem services and landscape features across five European sites. *Ecol. Indic.* **2017**, *94*, 74–86. [CrossRef]
16. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
17. Hu, Y.J.; Gao, S.; Janowicz, K.; Yu, B.L.; Li, W.W.; Prasad, S. Extracting and understanding urban areas of interest using geotagged photos. *Comput. Environ. Urban Syst.* **2015**, *54*, 240–254. [CrossRef]
18. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv*, 2018; arXiv:1804.02767.
19. Sullivan, B.L.; Wood, C.L.; Iliff, M.J.; Bonney, R.E.; Fink, D.; Kelling, S. eBird: A citizen-based bird observation network in the biological sciences. *Biol. Conserv.* **2009**, *142*, 2282–2292. [CrossRef]
20. Walker, J.; Taylor, P. Using eBird data to model population change of migratory bird species. *Avian Conserv. Ecol.* **2017**, *12*, 4. [CrossRef]
21. Tufekci, Z. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM* **2014**, *14*, 505–514.
22. Quattrone, G.; Capra, L.; De Meo, P. There's no such thing as the perfect map: Quantifying bias in spatial crowd-sourcing datasets. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, Vancouver, BC, Canada, 14–18 March 2015; pp. 1021–1032.
23. Tsou, M.-H. Research challenges and opportunities in mapping social media and Big Data. *Cartogr. Geogr. Inf. Sci.* **2015**, *42*, 70–74. [CrossRef]
24. Nielsen, J. Participation inequality: Encouraging more users to contribute. Available online: https://www.nngroup.com/articles/participation-inequality/ (accessed on 1 August 2018).
25. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]
26. Hecht, B.J.; Stephens, M. A Tale of Cities: Urban Biases in Volunteered Geographic Information. *ICWSM* **2014**, *14*, 197–205.
27. Koylu, C.; Guo, D. Smoothing locational measures in spatial interaction networks. *Comput. Environ. Urban Syst.* **2013**, *41*, 12–25. [CrossRef]
28. Antoniou, V.; Morley, J.; Haklay, M. Web 2.0 geotagged photos: Assessing the spatial dimension of the phenomenon. *Geomatica* **2010**, *64*, 99–110.
29. Sonter, L.J.; Watson, K.B.; Wood, S.A.; Ricketts, T.H. Spatial and temporal dynamics and value of nature-based recreation, estimated via social media. *PLoS ONE* **2016**, *11*, e0162372. [CrossRef] [PubMed]
30. Hollenstein, L.; Purves, R. Exploring place through user-generated content: Using Flickr tags to describe city cores. *J. Spat. Inf. Sci.* **2013**, 21–48.
31. Li, L.; Goodchild, M.F.; Xu, B. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartogr. Geogr. Inf. Sci.* **2013**, *40*, 61–77. [CrossRef]
32. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
33. Porzi, L.; Rota Bulò, S.; Lepri, B.; Ricci, E. Predicting and understanding urban perception with convolutional neural networks. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 139–148.
34. Lin, T.-Y.; Cui, Y.; Belongie, S.; Hays, J. Learning deep representations for ground-to-aerial geolocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5007–5015.

35. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning deep features for scene recognition using places database. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 487–495.
36. Wan, J.; Wang, D.; Hoi, S.C.H.; Wu, P.; Zhu, J.; Zhang, Y.; Li, J. Deep learning for content-based image retrieval: A comprehensive study. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 157–166.
37. Yang, L.; MacEachren, A.M.; Mitra, P.; Onorati, T. Visually-Enabled Active Deep Learning for (Geo) Text and Image Classification: A Review. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 65. [CrossRef]
38. Bircham, P.M.M. *A History of Ornithology*; Collins: London, UK, 2007.
39. *National Survey of Fishing, Hunting, and Widllife-Associated Recreation*; United States Fish and Wildlife Service: Arlington, VA, USA, 2012.
40. Sheard, K. A twitch in time saves nine: Birdwatching, sport, and civilizing processes. *Sociol. Sport J.* **1999**, *16*, 181–205. [CrossRef]
41. Oddie, B. *Bill Oddie's Little Black Bird Book*; Pavilion Books: London, UK, 2014.
42. Ramenofsky, M.; Wingfield, J.C. Regulation of migration. *AIBS Bull.* **2007**, *57*, 135–143. [CrossRef]
43. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. *arXiv* **2016**, arXiv:1612.08242.
44. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
45. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]
46. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
47. Sadeghi, M.A.; Forsyth, D. 30hz object detection with dpm v5. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 65–79.
48. Yan, J.; Lei, Z.; Wen, L.; Li, S.Z. The fastest deformable part model for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2497–2504.
49. Lenc, K.; Vedaldi, A. R-cnn minus r. *arXiv*, 2015; arXiv:1506.06981.
50. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
51. Silverman, B.W. *Density Estimation for Statistics and Data Analysis*; Chapman and Hall: London, UK, 1986; 175p.
52. Tiwari, C.; Rushton, G. Using spatially adaptive filters to map late stage colorectal cancer incidence in Iowa. In *Developments in Spatial Data Handling*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 665–676.
53. Boakes, E.H.; McGowan, P.J.; Fuller, R.A.; Chang-qing, D.; Clark, N.E.; O'Connor, K.; Mace, G.M. Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biol.* **2010**, *8*, e1000385. [CrossRef] [PubMed]