

Fine-Grained Land Use Classification at the City Scale Using Ground-Level Images

Yi Zhu, Xueqing Deng and Shawn Newsam

Abstract—Multimedia researchers have exploited large collections of community-contributed geo-referenced images to better understand a particular image, such as its subject matter or where it was taken, as well as to better understand a geographic location, such as the most visited tourist spots in a city or what the local cuisine is like. The goal of this paper is to better understand location. In particular, we use geo-referenced image collections to better understand what occurs in different parts of a city at fine spatial and activity class scales. This problem is known as land use mapping in the geographical sciences.

We propose a novel framework to perform fine-grained land use mapping at the city scale using ground-level images. Mapping land use is considerably more difficult than mapping land cover and is generally not possible using overhead imagery as it requires close-up views and seeing inside buildings. We postulate that the growing collections of georeferenced, ground-level images suggest an alternate approach to this geographic knowledge discovery problem. We develop a general framework that uses Flickr images to map 45 different land-use classes for the City of San Francisco. Individual images are classified using a novel convolutional neural network containing two streams, one for recognizing objects and another for recognizing scenes. This network is trained in an end-to-end manner directly on the labeled training images. We propose several novel strategies to overcome the noisiness of our user-generated data including search-based training set augmentation and online adaptive training. We derive a ground truth map of San Francisco in order to evaluate our method. We demonstrate the effectiveness of our approach through geo-visualization and quantitative analysis. Our framework achieves over 29% recall at the individual land parcel level which represents a strong baseline for the challenging 45-way land use classification problem especially given the noisiness of the image data.

Index Terms—Geo-Referenced Images, Land Use Classification, Convolutional Neural Networks, Proximate Sensing

I. INTRODUCTION

THE proliferation of camera-equipped mobile devices, primarily smart-phones, has generated large collections of geo-referenced ground-level images and videos. Multimedia researchers have recognized the value of these collections and exploited them for two broad categories of applications: 1) understanding or annotating the images and videos themselves, and 2) understanding or annotating the locations where the images and videos were captured.

This work was supported in part by a National Science Foundation (NSF) CAREER grant, No. IIS-1150115, and a seed grant from the Center for Information Technology in the Interest of Society (CITRIS). We gratefully acknowledge the support of NVIDIA Corporation through the donation of the Titan X GPUs used in this work.

Yi Zhu, Xueqing Deng and Shawn Newsam are with the Department of Electrical Engineering and Computer Science, University of California, Merced, Merced, CA, 95343 USA. (email: {yzhu25, xdeng7, snewsam}@ucmerced.edu)



Fig. 1. Traditional approaches to land use mapping are restricted to a limited number of classes as shown in the map on the top. Our goal in this paper is to produce fine-grained maps as shown on the bottom using ground-level images. The figure is best viewed in color.

With regards to the first category of applications, multimedia researchers have leveraged large collections of community-contributed geo-referenced images and videos to perform tag

recommendation [1], video search [2], event recognition [3], location estimation [4], [5], [6], [7], summarize personal collections and perform story telling [8], and even provide guidance on taking better photos [9].

The second category of applications can be thought of as performing geographic discovery, and includes tasks such as visual summarization of geographic areas [10], travel [11] and point of interest [12], [13] exploration and recommendation, food recognition [14], clothing recommendation for anticipated journeys [15], and learning generic facial attributes [16].

Our work in this paper falls into the second category in that its goal is geographic discovery. Specifically, we leverage large collections of community-contributed photos to map land use.

Mapping land cover and land use, and their changes, are two fundamental geographic tasks. While land cover and land use are related and often overlap, their distinctions are important. Land cover “is the physical material at the surface of the Earth. It is the material that we see and which directly interacts with electromagnetic radiation and causes the level of reflected energy that we observe as the tone or the digital number at a location in an aerial photograph or satellite image. Land covers include grass, asphalt, trees, bare ground, water, etc. Land use, by contrast, is a description of how people use the land. Urban and agricultural land uses are two of the most commonly recognized high-level classes of use. Institutional land, sports grounds, residential land, etc. are also all land uses” [17]. Land cover can be mapped using overhead imagery since this imagery is essentially the reflected energy discussed above. Mapping land use is much more difficult.

Detailed and accurate land use information is important for building smart cities [18], [19], [20], [21], [22] as it can help with tasks such as environmental monitoring, urban planning, resource allocation, traffic control and governmental management. In particular, the transformation of land use over time provides a wealth of information for both the government and individuals to make informative decisions. Traditionally, land use maps are generated using survey-based approaches which requires enormous human effort. Further, these maps are only updated every 5 to 10 years and thus do not convey important information of how urban functional structures are changing. There is a great need to develop systems which can automatically generate accurate and up-to-date land use maps on a large scale.

Most research on automated land use classification utilizes high-resolution overhead (remote sensing) imagery [23], [24], [25]. However, while it might be possible to distinguish some land use classes using overhead imagery, such as airports from residential areas as is done in [26], it is much more difficult to determine land use in complex urban areas from above. In contrast, images taken at ground level are potentially more indicative. For example, overhead imagery is unlikely to be effective for determining whether a building is a restaurant or a barber shop whereas ground-level images taken inside the building can help decide this easily. In general, overhead imagery has limited ability to perform land use mapping at fine class granularity.

There have been recent efforts [27], [28] to utilize other data sources that are informative on how the land is used. This

includes point of interest (POI) data [29], street view images [30], [31], mobile phone data [32] and social multimedia [33]. However, these data sources are also limited in how well they can see inside buildings.

This motivates our work in this paper on exploring large online photos collections for performing large-scale fine-grained land use mapping. Popular social photo sharing websites present an all-around view of the world and contain a wealth of information. With more than 400 million geotagged images at Flickr alone, there is an exciting opportunity to automatically generate up-to-date city-scale land use maps. However, this endeavor faces many technical challenges including:

- The lack of ground truth land use maps makes it difficult to evaluate the proposed methods.
- It is difficult to manually label a large collection of images to train the classifiers, particularly deep learning based ones, so that weakly supervised or unsupervised learning is necessary.
- The photos at the online sharing websites are very noisy in terms of image quality, inaccurate geotags, uneven spatial distribution, etc.

To address these challenges, we first introduce a ground truth land use map of the City of San Francisco for evaluating the proposed methods. The ground truth map has a three level hierarchy: 5 top classes, 16 middle classes and 45 fine-grained classes. We then train a novel convolutional neural network (CNN) in a weakly supervised and end-to-end manner to classify individual images as depicting one of the land use classes. Finally, we propose several novel strategies to overcome the noisiness of the online photos, including search-based training dataset augmentation, online adaptive training and classification networks with two streams, one for objects and another for scenes.

Our work in this paper represents a thorough investigation into mapping fine-grained land use on a large-scale using online photos. There are several distinctive aspects of our work. First, is the fine granularity of our classes. As shown in Figure 1, traditional approaches usually perform only coarse-level land use classification with fewer than 10 classes for example. However, our approach considers 45 classes and can be easily extended to more classes. The fine granularity makes the problem more challenging yet more applicable. Our work also makes several algorithmic advances. The salient, novel contributions of our work include:

- To the best of our knowledge, our work is the first to conduct fine-grained land use mapping using ground level images available at photo sharing websites.
- We combine the Land Based Classification Standards (LBCS) and Google places API to create a ground truth map of the City of San Francisco. This map can be used to quantitatively evaluate proposed approaches to land use classification.
- We introduce online adaptive training to help address how noisy the online photos are. The strategy not only increases the classification accuracy, but also makes our trained model more robust to domain adaptation.
- We propose a two-stream classification network, with

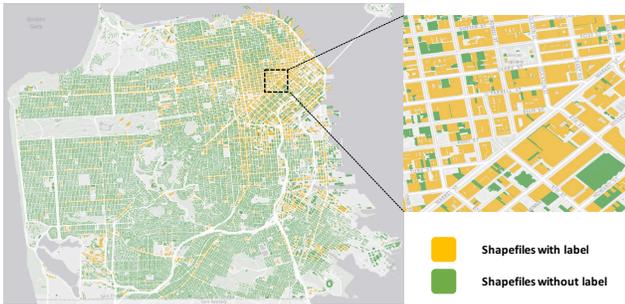


Fig. 2. Fine-grained land use maps do not exist for our study area which is one of the motivations of our work. We therefore construct a ground truth map using POI data from Google Places in order to evaluate our methods. This data is fairly sparse, though, so there are a lot of locations for which we do not have labels. We do not use these locations in our evaluation. This figure is best viewed in color.

object- and scene-centric models, to further improve the performance.

Our paper is organized as follows. After introducing related work in Section II, we describe the construction of our ground truth land use map as well as the details of our classification framework in Section III. In Section IV, we describe our dataset, implementation details, results and geo-visualizations. We then discuss how various designs impact the results in Section V, and finally conclude the paper in Section VI.

II. RELATED WORK

Our work has several lines of related research.

Large-scale geotagged photo collections Computer vision and multimedia researchers have been leveraging large collections of geotagged photos for geographic discovery for around a decade. This includes mapping world phenomenon [34], multimedia geolocation [35], landmark recognition and 3D modeling [36], smart city and urban planning [37], land cover and land use classification [23], [31], sentiment hotspot detection [38] and mapping human activity [39]. The exponential growth of photo sharing and related websites along with ever more publicly available multimedia sources make this research paradigm a promising direction for a range of interesting problems. Although such open-access multimedia represents a wealth of information, analyzing it is challenging due to how noisy it is. Challenges specific to using this data for geographic discovery include inaccurate location information, uneven spatial distribution, varying photographer intent and license limitations. We are mindful of these challenges and recognize they likely temper our results.

Our work is novel in that it uses a large collection of geotagged photos to perform fine-grained land use classification. We address several challenges mentioned above. We use region shapefiles to reduce geolocation error. In addition, we create a large training set (more than two million images) consisting of Google and Flickr images to learn a robust CNN model.

Convolutional neural networks Deep learning is advancing a number of pattern recognition and machine learning areas. Deep convolutional neural networks (CNNs) have resulted in often surprising performance gains in a range of computer vision problems [40], [41], [42]. Key to CNNs' performance

is their ability to learn high-level or semantic features from the data as opposed to the hand-crafted low- to mid-level features traditionally used in image analysis. Visualization of the feature maps learned by the convolutional layers during training [43] shows how the features become increasingly semantic, progressing from pixels, edges, color and texture, to motifs, parts, objects, scenes and concepts. Another significant benefit of the features learned by CNNs is their ability to generalize to problems involving image datasets other than the ones they were trained on [44]. This avoids having to retrain the networks which can take from hours to days even on powerful GPUs. Hence, several works have applied deep learning to advance the state-of-the-art in land use classification [45], [46], [47], [48], [49], [50], [51], [52], [53]. However, these works either use overhead imagery or only perform coarse-grained land use classification.

In this paper, we also use CNNs to classify the land use depicted in an image but introduce a novel learning technique termed online adaptive training to reduce the effect of the noisy web images during fine-tuning. We also propose a two-stream network, consisting of object- and scene-centric models, to further improve the land use classification performance.

Land cover and land use classification Land cover and land use classification are important tasks in geographic science. The maps they produce are critical for a range of important societal problems. However, land cover is distinct from land use, despite the two terms often being used interchangeably. Land cover is the physical material at the surface of the earth, which includes grass, trees, bare ground, water, etc. Land use is a description of how people utilize the land and of social-economic activity. Land cover classification is typically performed through the automated analysis of overhead imagery [54], [55]. However, land use classification is more difficult since it is often not possible from an overhead vantage point. We need to see inside buildings to determine their use(s). We also need to resolve details which are not discernible in overhead imagery or are only observable from ground level.

Researchers have performed some initial investigation into using ground-level photo collections for land cover [56], [57] and land use [58], [59], [60], [31] classification. Here, we only consider land use classification. [59] considered a two-class land use problem: developed and undeveloped, and [60] considered a limited number of land use classes on university campuses. Both of these works are able to use existing ground truth maps based on existing city zoning and campus maps for evaluation. However, there are no existing ground truth maps for our fine-grained problem. There is also some work [61], [62], [63] which focuses on other fine-grained aspects of the land use problem. For example, [61] focuses on fine-grained time scales and [62] focuses on fine-grained levels of damage. None of these works consider fine-grained land use classes.

III. FINE-GRAINED LAND USE CLASSIFICATION

This section provides the details of our approach. Section III-A describes our dataset which includes our hierarchical taxonomy of land use classes, our ground truth map and our training and mapping image sets. Sections III-B and

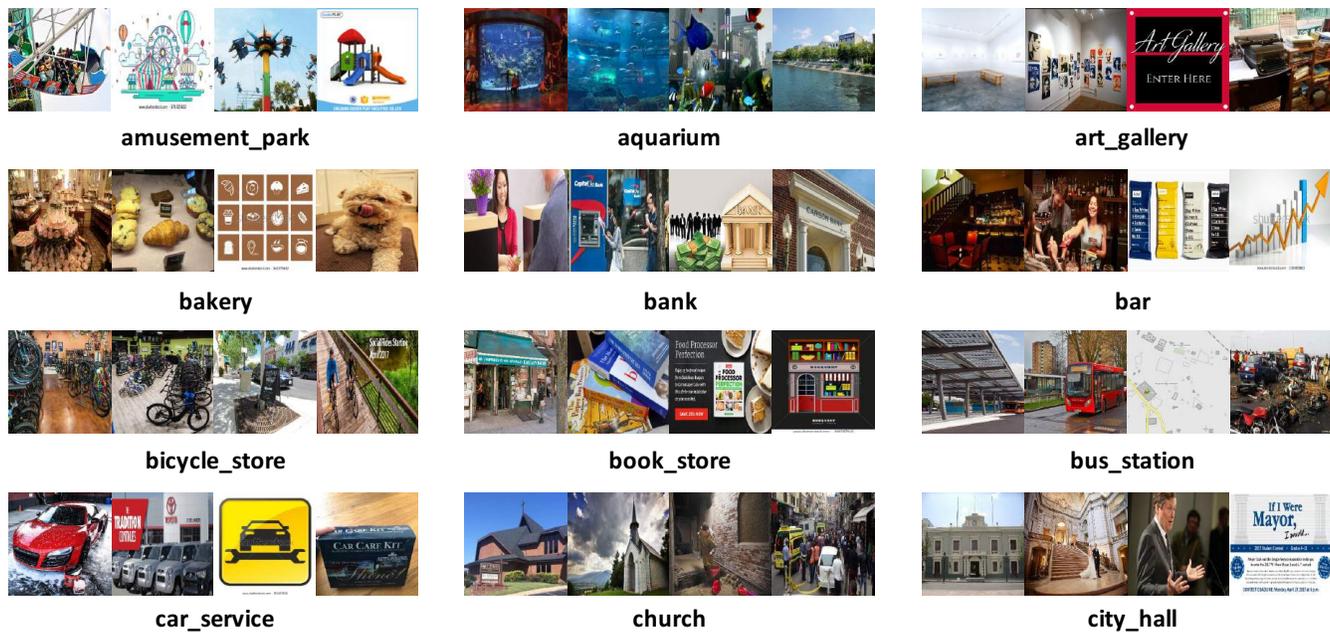


Fig. 3. Sample images from our training dataset. For each class, we display 4 images. The images are (manually) arranged from left to right based on how accurately they depict the land use class to demonstrate how noisy the online photos are. The figure is best viewed in color.

III-C describe our end-to-end training framework and online adaptive training strategy. Finally, Section III-D describes our two-stream classification network.

A. Dataset

Land Use Taxonomy

The first challenge we face is deriving a fine-grained land use taxonomy. Existing land use maps are limited to coarse taxonomies with 10 classes or fewer. This is probably due to the difficulty of the problem. After some investigation, we found the Land Based Classification Standards (LBCS)¹ of the American Planning Association which “extends the notion of classifying land uses by refining traditional categories into multiple dimensions, such as activities, functions, building types, site development character, and ownership constraints. Each dimension has its own set of categories and subcategories.” We adopt the “Function” dimension which refers to the economic function or type of establishment using the land, and consists of four levels with 9, 56, 212 and 154 classes each (not all classes have subclasses).

However, not all of the LBCS classes are pertinent to our problem since they 1) are not found in urban environments, such as the top-level class “Mining and extraction establishment”, 2) are not observable in shared online photos, such as the third-level class “Arts, entertainment, and recreation:Performing arts or supporting establishment:Agent for management services”, or 3) are not distinguishable in shared online photos, such as the two third-level classes “Assisted-living services” and “Life care or continuing care services”. Again, our focus is on land use classification in urban areas using shared online photos. We therefore prune and aggregate the LBCS taxonomy to three levels with 5, 16 and 45

classes each. The five top-level classes are: (1) Residence or accommodation functions; (2) General sales or services; (3) Transportation, communication, information, and utilities; (4) Arts, entertainment and recreation; (5) Education, public admin, health care and other institution. The full taxonomy can be found in Appendix A.

Ground Truth map

Fine-grained land use maps do not exist for our study area—this is one of the primary motivations of our work. We do not have an official ground truth to evaluate our results. We therefore create a ground truth map² using points of interest (POIs) as indexed by Google Places. The Google Places application programming interface (API)³ contains a large number of place types that are correlated with our land use classes. Based upon our previous work [28], we aligned the relevant place types with our land use taxonomy and used a large number of POIs from Google Places to create a land use map. As shown in Figure 2, the Google Places data is fairly sparse and so we do not have ground truth labels for many of the locations in our study area. We do not use these locations in our evaluations.

Training and Mapping Image Datasets

We downloaded 96,382 geotagged Flickr images for San Francisco from 2016. These images will be labeled by our trained classifier to generate the predicted land use map and we thus refer to them as the mapping images.

Our classifier is trained in a supervised fashion so we need labeled training images. Further, as a deep CNN, our classifier needs a large amount of training images. We want to preserve as much of the San Francisco Flickr images for mapping so we need another source of labeled training data. We make the observation that we do not need to know the locations of the training images; all we need are the class labels. We therefore

¹<https://www.planning.org/lbcs/>

²We will make this map publicly available in GIS compatible format.

³<https://developers.google.com/places/>

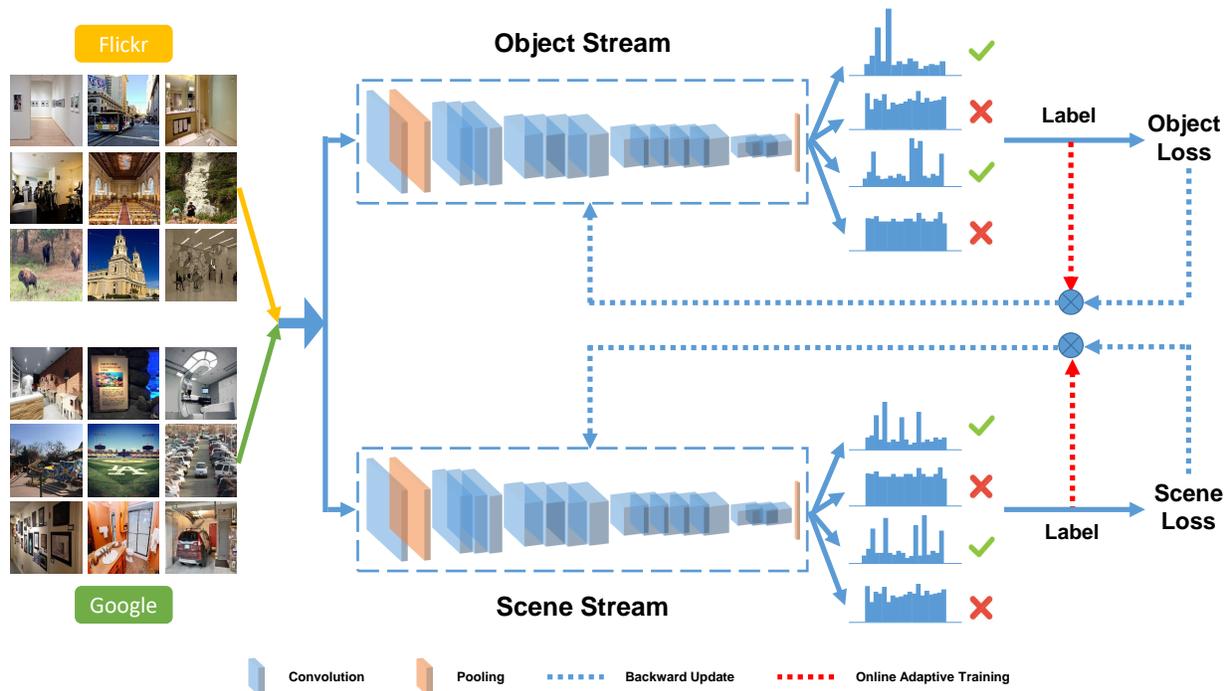


Fig. 4. Overview of our two-stream land use classification framework. During training, the input is mixed batches of images from Google Images and Flickr labeled with land use classes. Separate object- and scene-streams are used since each can be informative about the land use depicted in an image. During inference, the prediction scores of the two streams are fused with equal weights. The red dashed arrows indicate how the online adaptive training determines which training samples are used to perform network updates. Once trained, the network is applied to geotagged Flickr images of San Francisco in order to map land use. This figure is best viewed in color.

perform a keyword search at Google Images using our fine-grained class labels as keywords. We also perform keyword expansion. For example, in addition to searching for images using our land use class “school”, we also search using the keywords “elementary school”, “high school”, “adult school”, etc. This results in a large number of images returned from Google Images, 35,478 images for the class “school” and over 1 million images for all 45 classes.

We notice, though, that there is a potential domain shift between the training images from Google and the mapping images from Flickr which might limit the generalization of our trained network. In particular, the Google images tend to be simpler without complex backgrounds whereas the Flickr images often have faces and people, have been manipulated using photo editing software and/or have complex backgrounds. Therefore, we augment the Google training images with Flickr images downloaded using keyword searchers but from locations other than San Francisco, such as Atlanta, New York, Dallas, etc., in order to avoid overlap between our training and mapping sets. Our final training set consists of over 2 million images. We randomly split this into training and validation sets using a 0.8 to 0.2 ratio. Sample training images can be seen in Figure 3.

This search-based strategy for deriving the training images has three benefits: (1) it results in a balanced, rich training set; (2) it preserves all the San Francisco Flickr images for mapping; and (3) the data labeling procedure is automated and efficient. While it does result in noise in the training set, we rely on the findings that deep neural networks are immune to certain levels of noise and that their generalization capabilities

can sometimes actually be improved by it [64].

Land Parcel Shapefiles

Our land use maps, both the ground truth and predicted, are based on the irregularly shaped polygons corresponding to the land parcel footprints. These polygons are represented as shapefiles in a geographical information system (GIS). Parcel shapefiles are generally available. We download the parcel shapefiles for San Francisco from DataSF [65].

Exploiting parcel shapefiles has several distinct advantages over performing a regular gridding of the target area as has been done in previous work [66], [67], [68]:

- It allows us to ignore images that are not located in the regions we want to classify. In our case, those regions are the parcel footprints. In order to tolerate some geolocation error, we dilate the shapefile regions and regard a photo that is within 5 meters from a region boundary as being associated with that region. This geo-filtering of spatially unrelated images refines our mapping dataset from 96,382 to 58,418 images.
- The region boundaries of the resulting land use maps are very precise. They are also georegistered and so can be distributed for overlay on other GIS elements such as street networks. Figure 2 shows the ground truth map created using the parcel shapefiles.

B. End-to-End Learning

Standard approaches to land use/cover classification typically adopt a two stage pipeline [26] which first extracts the image features, such as color histogram, shape, texture, Scale Invariant Feature Transform (SIFT), GIST, or deep CNN



Fig. 5. Sample geo-visualizations in downtown San Francisco. Though each shapefile (building) may contain multiple land use types, we show the one with majority votes for clearer visualization. Slashed regions are correct predictions with respect to Google Map. The figure is best viewed in color.

features, and then trains an image classifier, such as a logistic regression, support vector machine (SVM), or shallow neural network classifier. This two stage framework has several shortcomings though: (1) the two stages are independent of each other and might not learn the optimal combination of feature extractor and classifier; and (2) the image features need to be cached, at least during training, which can be computationally and storage prohibitive for large-scale applications.

We therefore adopt an end-to-end deep learning framework in which the image features and classifier are learned jointly in an optimal fashion using CNNs. In particular, we train a 45-way classifier using the labeled training images. In order to reduce the effects of the possible domain gap between the Google and Flickr images, we use a batch size of 256 images during training, half of which are from the Google Images training set, and the other half are from the Flickr training set.

We adopt ResNet101 [69] as our network architecture due to its good trade-off between accuracy and efficiency. Implementation details can be found in Section IV-B. We also explore other model architectures, such as AlexNet [40], VGG16 [70], GoogleNet [71], ResNet34 [69] and DenseNet121 [72]. Comparisons between ResNet101 and the other architectures can be found in Section V-A Discussion.

C. Online Adaptive Training

We propose a novel method for improving the performance of our image classifier. We term this method online adaptive training and use it overcome the noisiness of our training data.

Deep learning requires large amounts of training data. One popular method for improving the performance of deep learning classifiers, especially when there is limited training data available for the task at hand, is to perform transfer learning where classifiers that have been (pre-)trained on a related task are fine-tuned rather than learned from scratch.

However, our training data is fairly noisy having been derived through keyword searches. Noisy labels can lead to poor local minima or model collapse during training. Manually cleaning our training dataset would be ideal but is not possible due to its size. We thus need to learn useful visual representations in the presence of label noise.

We thus propose online adaptive training as an unsupervised dataset cleaning procedure. During training, we use the

distribution of the class prediction scores to determine which samples to use for back propagating the network updates. We only use samples with distinct prediction scores and discard samples with uniform distributions. The intuition is that samples with distinct prediction scores are easier for our model to classify (correctly or incorrectly) whereas samples with more uniform distribution scores are ambiguous and thus can confuse our model. Note that, we use the same training dataset in online learning as in the previous end-to-end training. We just automatically ignore the effect of noisy labels based on the prediction distribution during loss computation.

Let $y_i = [y_{i1}, y_{i2}, \dots, y_{in}]$ represent the prediction (softmax) scores of training sample i and let n denote the number of classes which is 45 in our situation. We calculate the probability of *discarding* sample i as

$$p_i = \max(0, 2 - \exp |\max(y_i) - \bar{y}_i|). \quad (1)$$

Here, \bar{y}_i is the mean of the prediction scores y_i . When the difference between the maximum and average of the prediction scores of a training sample is large, p_i will be small, and so the probability of using it to update the network weights will be high. As the distribution of the prediction scores becomes more uniform, p_i will become larger, and so will the probability of discarding the sample. We use a threshold to decide whether to discard a sample and set this threshold to 0.5 in the experiments below. We empirically found 0.5 to be most effective through a grid search from 0.1 to 0.9. We could alternately perform soft weighting where we use the probability p_i to weight the importance of each sample instead of using a threshold to discard it. We explored both the soft and hard selection schemes but did not observe much difference in performance between them.

This sampling strategy is somewhat similar in motivation to the idea of hard negative mining [73], [74], a useful strategy for optimizing the training of machine learning models without leveraging extra data. However, rather than relying on false positives to improve the training of our model, we instead rely on the distinctive images since our labels are noisy.

Our training procedure has two stages. We first perform conventional end-to-end learning using our training data and then fine tune the model using the proposed online adaptive training framework. We show in the experiments below that

TABLE I
LAND USE CLASSIFICATION PERFORMANCE: BOTH IMAGE-LEVEL CLASSIFICATION AND SHAPEFILE-LEVEL MAPPING ACCURACY.

Method	Classification	Mapping		
	Accuracy	Precision	Recall	F1 Score
SIFT	29.16	4.56	12.85	3.37
SIFT + Fisher Vector Encoding	31.20	5.01	13.67	3.67
ResNet101 fc Layer (Pre-trained)	37.87	7.92	18.98	5.59
ResNet101 (Fine-Tuned)	43.90	10.57	21.67	7.10
ResNet101 (Adaptive, Object)	46.73	12.30	25.41	8.29
ResNet101 (Adaptive, Scene)	42.93	10.11	20.09	6.89
ResNet101 (Two-Stream)	49.54	14.21	29.06	9.54

the online adaptive training improves the performance of our classifiers.

D. Two Stream Network: Object and Scene

Multiple aspects of ground level images can be informative of land use. Clearly, objects and their interactions can provide clues about land use. But, so can the overall scenes or environment depicted in the images. Therefore, inspired by other work on multi-model learning [75], [76], we propose a two-stream architecture with one stream focused on objects and another on scenes.

Our object stream is a CNN model pretrained on the ImageNet dataset [77]. We complement this with a scene-centric model pretrained on the Places365 dataset [78]. We hope our object stream can capture features about object shape, color, texture, etc. Our scene stream can capture features about scene layout, object interactions, etc. These two streams are able to complement each other and perform better land use classification. Both streams are fine tuned in two steps, first using our training data and then using the adaptive online training strategy described above.

In order to prevent the object and scene models from collapsing into the same generic model during fine-tuning, we preserve the models' complementarity by fixing some of their parameters. Specifically, we fix the weights of the first convolutional groups of both the object and scene ResNet101 models during fine tuning. This helps ensure that the low- and mid-level features remain object- and scene- oriented.

Our two-stream land use classification framework can be seen in Figure 4.

IV. EXPERIMENTS

We first describe our dataset in Section IV-A and the implementation details in Section IV-B. Then we report the performance of our proposed approach in Section IV-C.

A. Dataset

As described above, our ground level image dataset has two components, the training/validation set and the mapping set.

The training/validation is used to train our classifier and evaluate it at the image level. It is constructed through keyword search at Google Images and Flickr and contains 2,159,460 images spread over 45 land use class. The dataset is generally balanced and has around 45,000 images per class. We split the dataset with a ratio of 0.8 to 0.2 for training and validation.

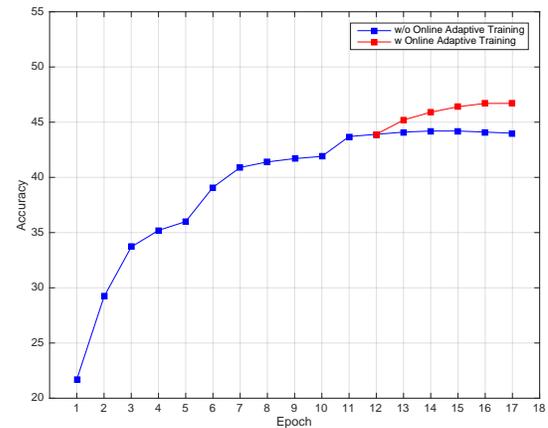


Fig. 6. Image-level land use classification accuracy versus training epochs. Without online adaptive training, the accuracy plateaus at epoch 12 (blue curve). With adaptive training, the accuracy continues to increase (red curve).

The image-level classification accuracy of our CNN model is evaluated on the validation set.

The mapping set consists of geotagged Flickr images taken in the City of San Francisco in 2016. We download a total of 96,382 images using the Flickr API. However, after geo-filtering using the land parcel shapefiles, the final mapping set has 58,418 images.

The mapping images are used to derive a land use map through simple label propagation. That is, if an image is associated with a particular parcel because it either falls inside the parcel or is within five meters of it, and our classifier predicts a particular land use class, then that class is assigned to the parcel. Note that a single parcel can be assigned multiple land uses which makes sense.

We evaluate the predicted land use map using the ground truth map derived from Google Places (see Section III-A above). We compute precision at the parcel level as the number of correct predictions divided by the total number of predictions. A prediction is considered correct if the class is in the ground truth for that parcel. We also compute recall as the number of ground truth classes we predict for a parcel divided by the number of ground truth classes for that parcel. Finally, we also compute the F1 score as the harmonic average of precision and recall.

B. Implementation Details

For the CNNs, we use the PyTorch toolbox. For all the experiments and speed evaluation, we use a workstation with an Intel Core I7 (4.00GHz) and 4 NVIDIA Titan X GPUs.

End-to-End learning: We use ResNet101 as our network architecture for both the object and scene streams. The object stream is pre-trained on the ImageNet dataset [77] and the scene stream is pre-trained on the Places365 dataset [78]. We change the final layers of each stream to 45-way classifiers for fine tuning. The model is trained using stochastic gradient descent with the default parameter values. The batch size is set to 256. The initial learning rate is set to 0.01 and is divided by 10 every 5 epochs. We end our training at epoch 12.

Online Adaptive Training Given the fine-tuned model, we perform online adaptive training. We feed a batch of images

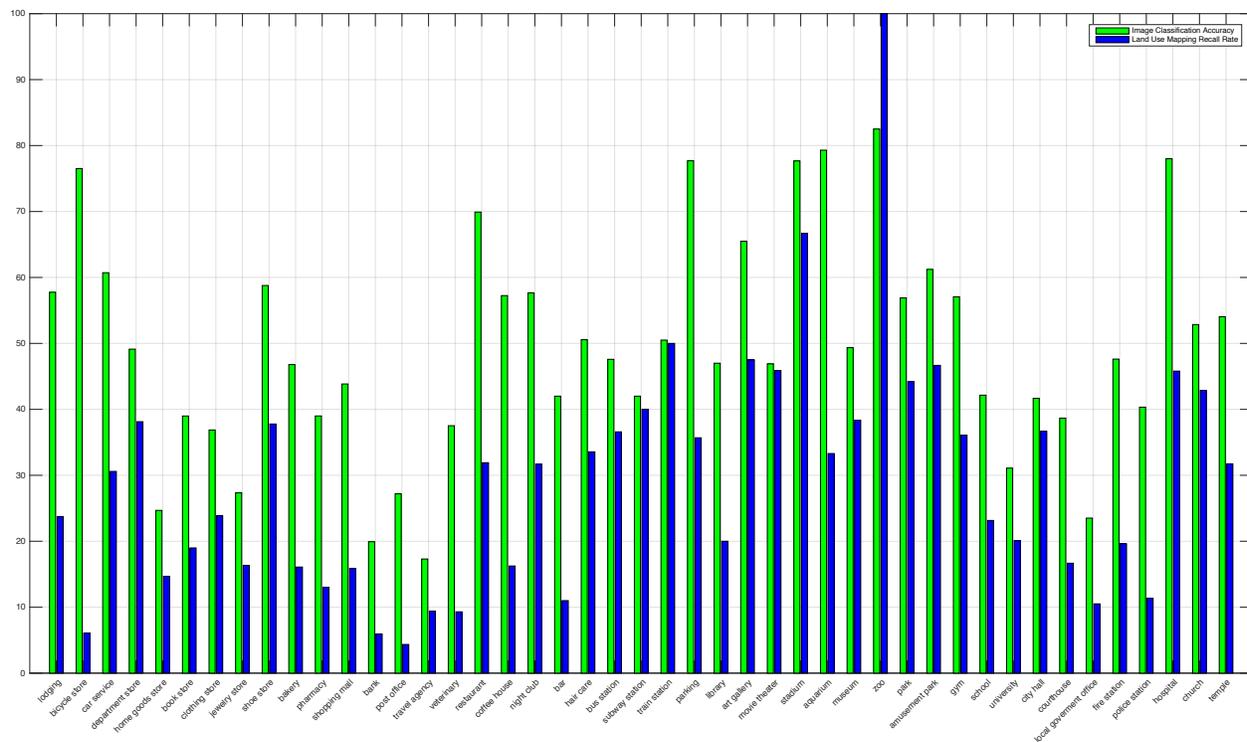


Fig. 7. Per-class image-level classification accuracy (green) versus shapefile-level mapping recall (blue). The figure is best viewed in color.

to the network and forward compute the land use prediction scores. We implement a custom loss layer to compute the cross entropy loss according to (1). We favor the samples with distinct prediction scores to perform back propagation, and discard those samples with more uniformly distributed prediction scores. Since the model is already fine-tuned, the initial learning rate for the adaptive online training is set to 10^{-5} , and divided by 10 every epoch. We stop the training at epoch 4.

Two Stream Network As mentioned above, we fix parts of the object and scene stream models during fine tuning to preserve their complementarity. During inference, the results of the two streams are combined using late fusion by an equal averaging of the individual softmax scores.

C. Results

This section presents our land use classification results. We compare end-to-end deep learning approaches to traditional two-stage approaches which perform feature extraction and classification separately; demonstrate the benefits of fine tuning the networks on our image dataset; show the effectiveness of our online adaptive learning strategy; and establish that combining object and scene streams improves performance. Further ablative studies are presented in Section V Discussion.

We present classification results at both the image and parcel level. Image-level accuracy is determined using the validation set downloaded from Google Images and Flickr (see Sections III-A and IV-A). An image is considered as being labeled correctly if the predicted land use class matches the class the image is assigned to during the download process. Parcel-level

accuracy is determined by comparing the land use class(es) assigned to the parcel, based on the image level classification, to the ground truth label(s) of the parcel. Precision, recall, and F1 scores are computed for the parcel-level results (see Section IV-A).

Image-level Classification Results:

Top section of Table I: We first present the performance of traditional two-stage approaches which perform feature extraction and classification separately. We consider both hand-crafted and deep-learning features. For hand-crafted features, we extract Scale Invariant Feature Transform (SIFT) features [79] which are then encoded as single, global bag of visual words feature vectors through k-means clustering or Fisher vector encoding. For deep-learning features, we extract a 1000-dimension feature vector per image using the last fully-connected layer of a ResNet101 CNN trained on the ImageNet challenge. Classification for the hand-crafted or deep-learning features is performed using support vector machines (SVMs) trained on the training image dataset and evaluated on the validation image dataset. As shown in Table I, the deep-learning features result in much higher image-level classification performance than the hand-crafted features. This result is in line with many, similar findings in other problems that the deep-learning features capture higher level semantics and generalize better than hand-crafted features. Fisher vector encoding results in a marginal improvement over k-means clustering but is still not competitive with the deep-learning features.

Bottom section of Table I: Here we list the performance of

our proposed approach and detail the improvements brought about by our novel strategies. First, we fine tune the deep networks (object stream) on our data. Despite the fact that our training dataset is quite noisy, our end-to-end trained model outperforms the pre-trained model used as a generic feature extractor. This indicates that end-to-end training is better than a two-stage method, especially for domain transfer learning. Second, we show the effectiveness of our proposed online adaptive training strategy. It improves 3% over the end-to-end trained model by discarding hard examples during fine-tuning. As shown in Figure 6, the image-level classification accuracy plateaus at epoch 12 (blue curve) without the online adaptive training but, with our proposed method, the accuracy continues to increase (red curve). Finally, combining the object- and scene-streams results in an accuracy of 49.54% on the image classification task with 45 classes. This result is promising given how noisy our crowd-sourced dataset is (as shown in Figure 3).

Parcel-level Mapping Results:

We now evaluate our predicted land use maps at the parcel level. We first observe that the parcel-level performance of the various approaches is correlated with the image-level performance. This indicates that future work on improving the image-level accuracy will result in better land use maps.

In general, we observe that precision at the parcel level is quite low, underscoring the difficulty of our problem. This low precision is due to the large number of false positives that result from the noisy collections of images often associated with the parcels. This suggests an improved method for propagating the image-level labels to the parcels, such as applying a threshold. (Note that taking the majority vote is not appropriate since a parcel can have multiple land use classes). This is a topic for future work.

Recall at the parcel level is much better than precision. Our two-stream network with adaptive online training achieves a recall rate of 29.06% on the challenging 45-class land use mapping problem, which is almost a 17% improvement over the baseline two-stage approach using SIFT features. In the following, we only discuss recall rates.

Sample geo-visualizations of the predicted land use maps are shown in Figure 5. These regions were randomly picked from downtown San Francisco. We show the majority vote for each parcel to make the visualization simpler. The slashed parcels indicate that the majority vote is one of the ground truth classes. Taking the leftmost image in Figure 5 as an example, we correctly predict a coffee house (Starbucks), a clothing store (Ross Dress For Less), a school (music training school) and a shopping mall (Macy’s). However, the museum prediction is wrong because the building is a library. After a sanity check, we actually have library prediction for that shapefile, it is just museum has more photos and wins the majority vote. Overall, we see our approach does well on producing land use maps that are fine grained at both the class and spatial levels.

Our results establish a starting baseline on this difficult problem as represented in our dataset, and leave plenty of room for other researchers to improve upon. New problems

TABLE II
CNN ARCHITECTURE SEARCH.

Method	Accuracy (%)	Speed (fps)
AlexNet	36.45	68.9
VGG16	43.88	25.6
GoogleNet	42.03	7.8
ResNet34	43.12	19.4
ResNet101	46.73	6.4
DenseNet121	47.29	3.8

and datasets with room for improvement are important for continued progress in deep learning. For example, state-of-the-art accuracy on the new action recognition dataset Charades [80], which is also crowd sourced and fine grained, is only around 15%, which provides significant opportunity for improvement versus older datasets, such as ActivityNet [81], on which accuracy is saturating. Our problem and dataset provide similar opportunities for making progress on the difficult problem of fine-grained, large-scale land use classification.

V. DISCUSSION

In this discussion section, we first compare different CNN architectures in terms of their accuracy and efficiency on our problem. Section V-B presents the per-class image- and parcel-level performances, and Section V-C investigates why and how the object and scene models are complementary. In Section V-D, we explore how performance varies with class granularity. Finally, in Section V-E, we investigate the domain adaptability of our framework by applying it to other data sources, in particular Instagram images.

A. CNN Architecture Search

As is well known, the performance of a CNN for a particular task can depend greatly on its architecture, particularly its depth, width, and number of internal connections. We therefore perform a CNN architecture search to identify the optimal network for our problem of classifying land use using noisy web images, in terms of the trade-off between accuracy and efficiency. We compare several architectures including AlexNet, VGG16, GoogleNet, ResNet34, ResNet101 and DenseNet121. These architectures are the result of careful design and have been widely used in different areas.

The results of our architecture search can be seen in Table II. Here, we only report the image classification accuracy of the object-centric model. The speed is reported as frames (images) per second (fps). The higher the fps, the faster the model runs during inference.

In general, deeper networks result in better performance. One interesting observation is that VGG16 performs better than ResNet34 despite the fact that VGG16 has 16 layers while ResNet34 has 34 layers. ResNet34 is shown to perform better than VGG16 on the object recognition task in the ImageNet challenges [77]. This demonstrates that VGG16 is a more robust model which has good generalization to noisy datasets.

DenseNet121 performs the best due to its deeper network, implicit supervision and being less prone to overfitting. However, it is both memory and time expensive. We therefore

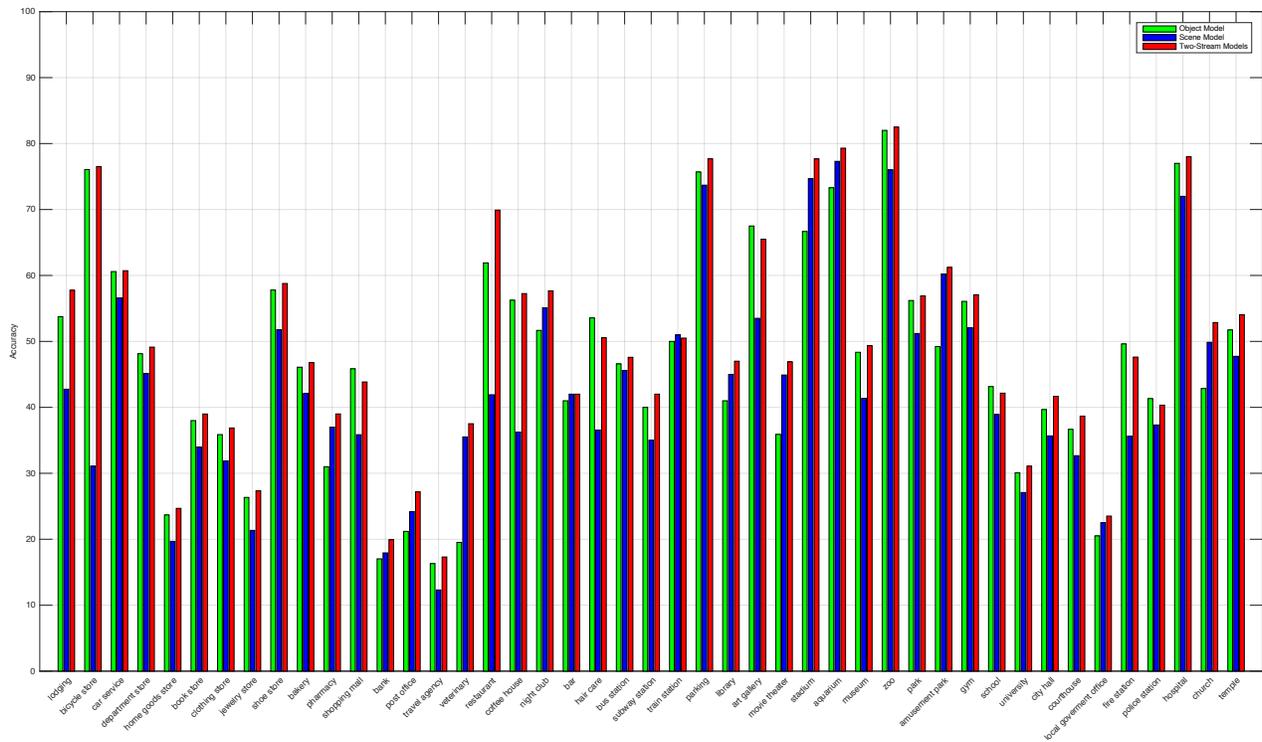


Fig. 8. Per-class image-level classification accuracy of the object stream (green), scene stream (blue) and two-stream (red). We observe that the object and scene information is complementary for recognizing most land use types. The figure is best viewed in color.

choose ResNet101 as our base CNN architecture due to its trade-off between accuracy and efficiency. ResNet101 can perform inference almost twice as fast as DenseNet121 (6.4 fps vs 3.8 fps), with a minor 0.5 performance drop.

B. Image-Level Classification versus Land Use Mapping

In Section IV-C, we found that land use mapping performance is closely related to the image-level classification accuracy. Intuitively, this makes sense because better image-level classification should lead to better performance for subsequent tasks that build upon it. We here explore that relationship in greater detail by examining the per-class results as shown in Figure 7.

We observe that there are some classes for which image-level classification accuracy and land use mapping recall are not correlated. For example, the image-level accuracy for the “bicycle store” class is almost 80% since our object model is good at detecting bicycles. However, the mapping recall rate is less than 10%. This is likely due to the fact that bicycle stores are usually quite small and so the noisiness of the web images and the sensitivity of the geotags leads to poor mapping performance. Also, for other land use types, such as “bank”, “local government office”, “courthouse” and “library”, the number of photos used for mapping is quite small due to privacy issues or lack of photographer interest in photographing these locations. Thus mapping recall rate of these classes is thus also very low.

C. Object and Scene Complementarity

As shown in Table I, two-stream networks outperform the single-stream ones, indicating that the object and scene streams are complementary to each other. We now investigate this in more detail.

Figure 8 compares the per-class accuracy of the object, scene and two-stream models. Based on this, we make the following two observations. (i) The object model performs better than the scene model on most of the land use classes. This could be because the presence of specific objects can be more indicative of land use than the general scene. For example, the object model outperforms the scene model on the class “bicycle stores” by a large margin likely because the presence of a bicycle is key for identifying this class. The overall scene of a bicycle store can be similar to other land use classes. (ii) The object and scene models are complementary to each other. 39 out of 45 classes obtain better performance when the outputs of the two models are fused. For the remaining six classes, the performance decrease is marginal.

The five classes with the most improvement (amount in parenthesis) after incorporating the scene model are “veterinary” (18.12%), “amusement park” (12.30%), “movie theater” (11.04%), “stadium” (10.97%) and “church” (9.86%). We believe that the scene cue is important for the recognition of these land use classes because there are few objects specifically related to these classes.

The six classes that have decreased performance are “hair care” (−3.77%), “fire station” (−2.49%), “shopping mall” (−2.31%), “art gallery” (−2.04%), “school” (−1.68%) and

TABLE III
IMAGE-LEVEL CLASSIFICATION ACCURACY FOR DIFFERENT DATASET GRANULARITIES.

Method	45-way	16-way	5-way
Fine Granularity	46.7	61.8	75.6
Middle Granularity	—	60.2	68.4
Coarse Granularity	—	—	49.3

“police station” (−1.20%). The reason for the decrease in performance is that the scene model simply performs poorly on these classes.

D. Fine to Coarse

In some scenarios, fine-grained mapping might not be necessary and so using the 16 middle or the 5 top level classes is sufficient. This raises the question of whether is it better to 1) still use a fine-grained classifier and aggregate the classification results to derive the coarser classes, or 2) train models specifically targeted at the coarser granularities. We investigate this here by training 16-way and 5-way classifiers on our dataset.

We call our original 45-way classifier the fine granularity model, the 16-way classifier the middle granularity model and the 5-way classifier the coarse granularity model. The performances of these three classifiers on their respective problems are shown on the diagonal in Table III (this corresponds to method 2 above). The other entries in this table show the results of applying the fine granularity model and then performing aggregation (method 1 above). We see that the fine granularity model is beneficial even if the target is a coarser set of classes. The fine granularity model achieves 75.6% for the 5 class problem compared to 68.4% for the middle and just 49.3% for the coarse granularity models. The image-level models are not able to discriminate between the coarser classes due the large intra-class image variation. For example, the “General sales or services” top-level class includes many concepts that are visually quite different, like bank and bakery, hair care and restaurant.

E. Generalization to other Image Data Sources

We here explore whether our framework, in particular the trained image classifiers, generalizes to mapping land use given another source of images. We use Instagram as this other source. This will allow us to observe the robustness of our method and its transfer learning capability. It will also demonstrate that our ground truth map can be used to evaluate other approaches or data sources.

We download a total of 121,567 Instagram images within the city of San Francisco for the year of 2014 using the Instagram API. We used our existing, trained models to classify each Instagram image with its predicted land use class and then mapped the results. The recall rate of this mapping performance is 17.3%. Although it is lower than the 29.03% recall achieved on the Flickr dataset, it nonetheless demonstrates our model has decent domain adaptability. We note that the Instagram images often differ substantially from

the Flickr images in style. Most Instagram images are selfie or selfie-like, which portrays only close-by scenes and therefore is not optimal for the recognition of land use classes.

VI. CONCLUSION

We presented a novel framework for fine-grained land use classification at the city scale using ground-level images. We established a three-level land use class taxonomy with 45 fine-grained classes and created a corresponding ground truth map for San Francisco. Our algorithmic contributions include online adaptive learning and a two-stream image-level classifier that is trained in an end-to-end fashion. Our results set a strong baseline for our problem and dataset.

In the future, we would like to improve our two-stream model in the following directions. First, we plan to explore multi-modal information, such as text, audio, video or other input signals, to investigate their complementarity. Second, we would like to modify our framework to incorporate a human-in-the-loop since accurate fine-grained land use mapping likely requires human knowledge or guidance. Third, based on our observation that the object stream achieves better performance than scene stream, we plan to further improve the object stream by applying off-the-shelf object detectors that which specific objects appear as well where they appear. This additional knowledge can help identify which objects, including, possibly their spatial co-occurrences, are key to determining various land use classes.

REFERENCES

- [1] J. Liu, Z. Li, J. Tang, Y. Jiang, and H. Lu, “Personalized Geo-Specific Tag Recommendation for Photos on Social Websites,” *IEEE Transactions on Multimedia*, 2014.
- [2] Y. Yin, Y. Yu, and R. Zimmermann, “On Generating Content-Oriented Geo Features for Sensor-Rich Outdoor Video Search,” *IEEE Transactions on Multimedia*, 2015.
- [3] J. Yuan, J. Luo, and Y. Wu, “Mining Compositional Features From GPS and Visual Cues for Event Recognition in Photo Collections,” *IEEE Transactions on Multimedia*, 2010.
- [4] X. Li, M. Larson, and A. Hanjalic, “Geo-Distinctive Visual Element Matching for Location Estimation of Images,” *IEEE Transactions on Multimedia*, 2018.
- [5] X. Zhang, X. Hu, S. Wang, Y. Yang, Z. Li, and J. Zhou, “Learning Geographical Hierarchy Features via a Compositional Model,” *IEEE Transactions on Multimedia*, 2016.
- [6] X. Qian, H. Wang, Y. Zhao, X. Hou, R. Hong, M. Wang, and Y. Y. Tang, “Image Location Inference by Multisaliency Enhancement,” *IEEE Transactions on Multimedia*, 2017.
- [7] Y. Song, X. Chen, X. Wang, Y. Zhang, and J. Li, “6-DOF Image Localization From Massive Geo-Tagged Reference Images,” *IEEE Transaction on Multimedia*, 2016.
- [8] Y. Wu, X. Shen, T. Mei, X. Tian, N. Yu, and Y. Rui, “Monet: A System for Reliving Your Memories by Theme-Based Photo Storytelling,” *IEEE Transactions on Multimedia*, 2016.
- [9] W. Yin, T. Mei, C. W. Chen, and S. Li, “Socialized Mobile Photography: Learning to Photograph With Social Context via Mobile Devices,” *IEEE Transactions on Multimedia*, 2014.
- [10] Y. Zhang and R. Zimmermann, “Efficient Summarization From Multiple Georeferenced User-Generated Videos,” *IEEE Transactions on Multimedia*, 2016.
- [11] Z. Xu, L. Chen, Y. Dai, and G. Chen, “A Dynamic Topic Model and Matrix Factorization-Based Travel Recommendation Method Exploiting Ubiquitous Data,” *IEEE Transactions on Multimedia*, 2017.
- [12] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, “Author Topic Model-Based Collaborative Filtering for Personalized POI Recommendations,” *IEEE Transactions on Multimedia*, 2015.

- [13] X. Wang, Y. L. Zhao, L. Nie, Y. Gao, W. Nie, Z. J. Zha, and T. S. Chua, "Semantic-Based Location Recommendation With Multimodal Venue Semantics," *IEEE Transactions on Multimedia*, 2015.
- [14] L. Herranz, S. Jiang, and R. Xu, "Modeling Restaurant Context for Food Recognition," *IEEE Transactions on Multimedia*, 2017.
- [15] X. Zhang, J. Jia, K. Gao, Y. Zhang, D. Zhang, J. Li, and Q. Tian, "Trip Outfits Advisor: Location-Oriented Clothing Recommendation," *IEEE Transactions on Multimedia*, 2017.
- [16] Y. Y. Chen, W. H. Hsu, and H. Y. M. Liao, "Automatic Training Image Acquisition and Effective Feature Selection From Community-Contributed Photos for Facial Attribute Detection," *IEEE Transactions on Multimedia*, 2013.
- [17] P. Fisher, A. J. Comber, and R. Wadsworth, "Land Use and Land Cover: Contradiction or Complement," in *Re-presenting GIS*, P. Fisher and D. J. Unwin, Eds. Wiley, 2005, pp. 85–98.
- [18] Y. Yamagata and H. Seya, "Simulating a Future Smart City: An Integrated Land Use-Energy Model," *Applied Energy*, 2013.
- [19] M.-L. Marsal-Llacuna and M.-B. Lpez-Ibez, "Smart Urban Planning: Designing Urban Land Use from Urban Time Use," *Journal of Urban Technology*, 2014.
- [20] J. Rawat and M. Kumar, "Monitoring Land Use/Cover Change using Remote Sensing and GIS Techniques: A Case Study of Hawalbagh Block, District Almora, Uttarakhand, India," *The Egyptian Journal of Remote Sensing and Space Science*, 2015.
- [21] Y. Liu, J. Peng, L. Jiao, and Y. Liu, "PSOLA: A Heuristic Land-Use Allocation Model Using Patch-Level Operations and Knowledge-Informed Rules," *PLOS ONE*, 2016.
- [22] J. Tang, X. Tang, and J. Yuan, "Traffic-Optimized Data Placement for Social Media," *IEEE Transaction on Multimedia*, 2017.
- [23] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land Use Classification in Remote Sensing Images by Convolutional Neural Networks," *arXiv preprint arXiv:1508.00092*, 2015.
- [24] B. Zhao, B. Huang, and Y. Zhong, "Transfer Learning With Fully Pretrained Deep Convolution Networks for Land-Use Classification," *IEEE Geoscience and Remote Sensing Letters*, 2017.
- [25] M. Li, K. M. de Beurs, A. Stein, and W. Bijker, "Incorporating Open Source Data for Bayesian Classification of Urban Land Use From VHR Stereo Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2017.
- [26] Y. Yang and S. Newsam, "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.
- [27] T. Hu, J. Yang, X. Li, and P. Gong, "Mapping Urban Land Use by Using Landsat Images and Open Social Data," *Remote Sensing*, 2016.
- [28] X. Deng and S. Newsam, "Quantitative Comparison of Open-Source Data for Fine-Grain Mapping of Land Use," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2017.
- [29] S. Jiang, A. Alves, F. Rodrigues, J. Ferreira, and F. C. Pereira, "Mining Point-of-Interest Data from Social Networks for Urban Land Use Classification and Disaggregation," *Computers, Environment and Urban Systems*, 2015.
- [30] J. Kang, M. Krner, Y. Wang, H. Taubenbck, and X. X. Zhu, "Building Instance Classification using Street View Images," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018.
- [31] S. Workman, M. Zhai, D. Crandall, and N. Jacobs, "A Unified Model for Near/Remote Sensing," in *ICCV*, 2017.
- [32] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, "A New Insight into Land Use Classification Based on Aggregated Mobile Phone Data," *International Journal of Geographical Information Science*, 2014.
- [33] X. Liu, J. He, Y. Yao, J. Zhang, H. Liang, H. Wang, and Y. Hong, "Classifying Urban Land Use by Integrating Remote Sensing and Social Media Data," *International Journal of Geographical Information Science*, 2017.
- [34] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the World's Photos," in *WWW*, 2009.
- [35] J. Hays and A. A. Efros, "IM2GPS: Estimating Geographic Information from a Single Image," in *CVPR*, 2008.
- [36] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the World from Internet Photo Collections," *International Journal of Computer Vision*, 2008.
- [37] S. Paldino, I. Bojic, S. Sobolevsky, C. Ratti, and M. C. Gonzalez, "Urban Magnetism Through The Lens of Geo-tagged Photography," *arXiv preprint arXiv:1503.05502*, 2015.
- [38] Y. Zhu and S. Newsam, "Spatio-Temporal Sentiment Hotspot Detection using Geotagged Photos," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2016.
- [39] Y. Zhu, S. Liu, and S. Newsam, "Large-Scale Mapping of Human Activity using Geo-Tagged Videos," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2017.
- [40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *NIPS*, 2012.
- [41] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *CVPR*, 2014.
- [42] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," in *CVPR*, 2014.
- [43] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *ECCV*, 2014.
- [44] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks," in *CVPR*, 2014.
- [45] Y. Zhong, F. Fei, and L. Zhang, "Large Patch Convolutional Neural Networks for the Scene Classification of High Spatial Resolution Imagery," *Journal of Applied Remote Sensing*, 2014.
- [46] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains?" in *CVPR*, 2015.
- [47] Q. Weng, Z. Mao, J. Lin, and W. Guo, "Land-Use Classification via Extreme Learning Classifier Based on Deep Convolutional Features," *IEEE Geoscience and Remote Sensing Letters*, 2017.
- [48] F. P. S. Luus, B. P. Salmon, F. van den Bergh, and B. T. J. Maharaj, "Multitasking Deep Learning for Land-Use Classification," *IEEE Geoscience and Remote Sensing Letters*, 2015.
- [49] L. Tracewski, L. Bastin, and C. C. Fonte, "Repurposing a Deep Learning Network to Filter and Classify Volunteered Photographs for Land Cover and Land Use Characterization," *Geo-spatial Information Science*, 2017.
- [50] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks," *IEEE Geoscience and Remote Sensing Letters*, 2016.
- [51] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," *arXiv preprint arXiv:1709.00029*, 2017.
- [52] Q. Liu, R. Hang, F. Z. Huihui Song, J. Plaza, and A. Plaza, "Adaptive Deep Pyramid Matching for Remote Sensing Scene Classification," *arXiv preprint arXiv:1611.03589*, 2016.
- [53] Q. Liu, R. Hang, H. Song, and Z. Li, "Learning Multiscale Deep Features for High-Resolution Satellite Image Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, pp. 117–126, 2018.
- [54] R. Hang, Q. Liu, H. Song, and Y. Sun, "Matrix-Based Discriminant Subspace Ensemble for Hyperspectral Image SpatialSpectral Feature Fusion," *IEEE Transactions on Geoscience and Remote Sensing*, 2015.
- [55] R. Hang, Q. Liu, Y. Sun, X. Yuan, H. Pei, J. Plaza, and A. Plaza, "Robust matrix discriminative analysis for feature extraction from hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2017.
- [56] H. Oba, M. Hirota, R. Chbeir, H. Ishikawa, and S. Yokoyama, "Towards Better Land Cover Classification Using Geo-tagged Photographs," *IEEE International Symposium on Multimedia*, 2014.
- [57] D. M. Theobald, "Development and Applications of a Comprehensive Land Use Classification and Map for the US," *PLOS ONE*, 2014.
- [58] S. Shekhar, P. R. Schrater, R. R. Vatsavai, W. Wu, and S. Chawla, "Spatial Contextual Classification and Prediction Models for Mining Geospatial Data," *IEEE Transactions on Multimedia*, 2002.
- [59] D. Leung and S. Newsam, "Proximate Sensing: Inferring What-Is-Where From Georeferenced Photo Collections," in *CVPR*, 2010.
- [60] Y. Zhu and S. Newsam, "Land Use Classification Using Convolutional Neural Networks Applied to Ground-Level Images," in *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015.
- [61] J. Untenecker, B. Tiemeyer, A. Freibauer, A. Laggner, F. Braumann, and J. Luterbacher, "Fine-Grained Detection of Land Use and Water Table Changes on Organic Soils over the Period 1992-2012 using Multiple Data Sources in the Drmling Nature Park, Germany," *Land Use Policy*, 2016.
- [62] N. Attari, F. Ofii, M. Awad, J. Lucas, and S. Chawla, "Nazr-CNN: Fine-Grained Classification of UAV Imagery for Damage Assessment," *arXiv preprint arXiv:1611.06474*, 2016.
- [63] Y. Zhang, Q. Li, H. Huang, W. Wu, X. Du, and H. Wang, "The Combined Use of Remote Sensing and Social Sensing Data in Fine-Grained Urban Land Use Mapping: A Case Study in Beijing, China," *Remote Sensing*, 2017.

[64] L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian, "DisturbLabel: Regularizing CNN on the Loss Layer," in *CVPR*, 2016.

[65] "DataSF: San Francisco Open Data." [Online]. Available: <https://datasf.org/opendata/>

[66] D. Leung and S. Newsam, "Exploring Geotagged Images for Land-Use Classification," in *ACM workshop on Geotagging and its applications in multimedia*, 2012.

[67] "Geograph Britain and Ireland - photograph every grid square!" [Online]. Available: <http://www.geograph.org.uk/>

[68] J. D. Wickhama, S. V. Stehmanb, L. Gassc, J. Dewitzd, J. A. Fryd, and T. G. Wadea, "Accuracy Assessment of NLCD 2006 Land Cover and Impervious Surface," *Remote Sensing of Environment*, 2013.

[69] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.

[70] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *ICLR*, 2015.

[71] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *CVPR*, 2015.

[72] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[73] Q. You, J. Luo, H. Jin, and J. Yang, "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks," in *AAAI*, 2015.

[74] A. Shrivastava, A. Gupta, and R. Girshick, "Training Region-based Object Detectors with Online Hard Example Mining," in *CVPR*, 2016.

[75] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *NIPS*, 2014.

[76] J. Hu, L. Mou, A. Schmitt, and X. X. Zhu, "FusioNet: A Two-Stream Convolutional Neural Network for Urban Scene Classification using PolSAR and Hyperspectral Data," in *Joint Urban Remote Sensing Event (JURSE)*, 2017.

[77] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[78] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database," in *NIPS*, 2014.

[79] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Key-points," *International Journal of Computer Vision*, 2004.

[80] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *ECCV*, 2016. [Online]. Available: <http://allenai.org/plato/charades/>

[81] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding," in *CVPR*, 2015.

- bank
- c) Business, professional, scientific, and technical services
 - post_office
 - travel_agency
 - veterinary_care
- d) Food services
 - restaurant
 - coffee_house
 - night_club
 - bar
- e) Personal services
 - hair_care
- 3) Transportation, communication, information, and utilities
 - a) Transportation service
 - bus_station
 - subway_station
 - train_station
 - parking
 - b) Communications and information
 - library
- 4) Arts, entertainment and recreation
 - a) Performing arts or supporting establishment
 - art_gallery
 - movie_theater
 - stadium
 - b) Museums and other special purpose recreational institutions
 - aquarium
 - museum
 - zoo
 - c) Amusement, sports, or recreation establishment
 - park
 - amusement_park
 - gym
- 5) Education, public admin, health care and other institution
 - a) Educational services
 - school
 - university
 - b) Public administration
 - city_hall
 - courthouse
 - local_government_office
 - c) Public safety
 - fire_station
 - police_station
 - d) Health and human services
 - hospital
 - e) Religious institutions
 - church
 - temple

APPENDIX A FULL DATASET HIERARCHY

We list the full dataset hierarchy in 3-level as below. There are 5 top classes, 16 middle classes and 45 bottom classes.

- 1) Residence or accommodation functions
 - a) Hotels, motels, or other accommodation services
 - lodging
- 2) General sales or services
 - a) Retail sales or service
 - bicycle_store
 - car_service
 - department_store
 - home_goods_store
 - book_store
 - clothing_store
 - jewelry_store
 - shoe_store
 - bakery
 - pharmacy
 - shopping_mall
 - b) Finance and Insurance



Yi Zhu received the B.S. degree in Electrical Engineering from the Northwestern Polytechnical University in 2011 and the M.S. degree in Electrical Engineering from the University of Kansas in 2014. He is currently a Computer Science PhD candidate at the University of California, Merced. His research interests are in computer vision and deep learning. He has been working on problems such as video classification, semantic segmentation, optical flow estimation, surveillance anomaly detection and geo-event discovery.



Xueqing Deng received the B.Sc. degree in Geographic Information System and Remote Sensing from Sun Yat-Sen University, Guangzhou, China, in 2016. She has been working toward in land use classification using geotagged social media with deep learning, domain adaptation with generative adversarial learning at Computer Vision Lab from the University of California, Merced, United States. Her research interests include machine learning as well as deep learning in the context of geospatial problems and remote sensing.



Shawn Newsam Dr. Shawn Newsam received the B.S. degree in Electrical Engineering and Computer Science from the University of California, Berkeley, in 1990, the M.S. degree in Electrical and Computer Engineering from the University of California, Davis, in 1996, and the Ph.D. degree in Electrical and Computer Engineering from the University of California, Santa Barbara, in 2004.

He is currently an Associate Professor and Founding Faculty of Electrical Engineering and Computer Science at the University of California, Merced.

Prior to joining U.C. Merced, he was a Postdoctoral Scholar in the Center for Applied Scientific Computation at the Lawrence Livermore National Laboratory. His research interests are in image processing, computer vision, and applied machine learning particularly as applied to geospatial data. Dr. Newsam is the recipient of a U.S. Department of Energy Early Career Scientist and Engineer Award, a U.S. National Science Foundation Faculty Early Career Development (CAREER) Award, and a U.S. Office of Science and Technology Policy Presidential Early Career Award for Scientists and Engineers (PECASE).