GROUND2SKY LABEL TRANSFER FOR FINE-GRAINED AERIAL CAR RECOGNITION

Baochen Sun[⋆] *Xingchao Peng* [†] *Stella X. Yu*[‡] *Kate Saenko*[†]

[⋆] Microsoft AI & Research [†] Boston University [‡] UC Berkeley / ICSI

ABSTRACT

Overhead images captured by helicopters, unmanned aerial vehicles and satellites are widely available. Prior aerial target recognition methods mainly deal with generic object categories such as cars, roads, and boats. We go beyond this and aim for fine-grained recognition, e.g., distinguishing between a Toyota and a Honda sedan. This task is so challenging for human annotators that labeling images directly is no longer an option: annotators are often unable to identify the object from such an extreme viewpoint and at such a low resolution.

We propose a novel solution to collect fine-grained annotations of aerial images and develop the first ground-to-sky cross-view car dataset with instance-level correspondences. We compare the performance of human experts and deep learning approaches on fine-grained car recognition from aerial imagery. Noting that intraclass variation in aerial images is limited, we further show that with simple data augmentation, a classifier can be trained from fewer instances yet achieves comparable or even significantly better performance than human experts. Our experimental evidence demonstrates that fine-grained object recognition from overhead images is not only feasible but also well suited for deep learning methods. Our dataset is available at: http://ai.bu.edu/Ground2Sky/

Index Terms— Fine-grained Recognition, Aerial Imagery, Label Transfer

1. INTRODUCTION

Can you tell the model and make of the car in the overhead-view image shown in Fig. 1? Such fine-grained aerial classification of aerial images is challenging even for humans, yet it is increasingly important as drones and satellites are becoming widely available for surveillance, search-and-rescue, scientific research and other applications. As more images are collected from these devices, it becomes harder for human analysts to process and extract information of interest, leading to increasing demand for computer vision approaches to automatic aerial image understanding.

Existing computer vision approaches [1, 2, 3, 4] have shown great promise for the analysis of aerial imagery, but have focused on basic rather than fine-grained object categories. While [5, 6, 7, 8, 9, 10] have successfully investigated fine-grained object recognition in consumer photos, few efforts have been made to study this task in aerial images.

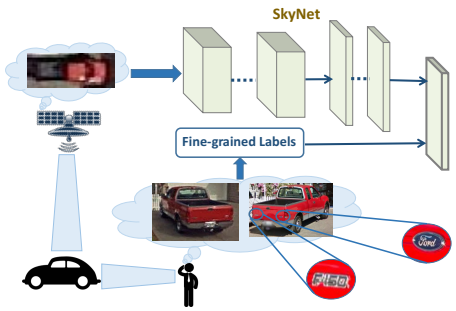

Fig. 1: We propose to train fine-grained car recognition from aerial imagery with labels obtained on corresponding ground imagery. Labelling aerial images is challenging due to extreme viewpoints and low resolutions. Instead of trying to obtain fine-grained labels from aerial images directly, we develop a novel method to accurately transfer labels from high quality and information rich street-view images. We show that a deep classifier (SkyNet) can be trained with a relatively small amount of data yet achieve comparable or even better performance than human experts due to limited intraclass variation.

Here we focus on fine-grained car recognition in aerial images (Fig. 1). To the best of our knowledge, we are the first to study fine-grained classification of mobile targets in aerial data. RegisTree [11] used both aerial images and street views to catalog trees. However, as trees are static objects, they can be more easily annotated in street view, and the labels transferred directly to the overhead images. We tackle a much more challenging task of fine-grained vehicle recognition. Cars are mobile objects, so there is no guarantee of synchronization between street-view and aerial-view. The major road-block for this line of research is the difficulty and the lack of ground-truth fine-grained annotations. Aerial images present the following unique challenges for data collection and annotation: 1) Annotator unfamiliarity; 2) Extreme viewpoints; 3) Low resolution; 4) Lack of synchronization between aerial and street views.

We propose a novel solution to transfer fine-grained labels from street-view images to aerial ones. We choose cars as our target, as vehicles are ubiquitous, and automatic vehicle model recognition is very useful for surveillance and analytics. We observe that, although vehicles are mobile, many residential addresses have the same vehicles parked in

the same driveway over a long period of time, sometimes many years. We use this guideline to select paired instances of street-view and aerial views of cars from Google Maps and obtain fine-grained labels based on the ground views. Our fine-grained labels include car body color, body type, car make, and model. We develop the first fine-grained aerial car dataset with instance-level cross-view correspondences.

Our main contributions are: **1)** the first effort to investigate fine-grained vehicle recognition from aerial images; **2)** an innovative method to collect the first fine-grained aerial car dataset with accurate labels; **3)** experiments comparing human experts and deep learning methods on car recognition from aerial imagery, and **4)** demonstrating that fine-grained object recognition from aerial images is not only feasible, but also maybe be more suitable for deep learning methods than for humans.

2. RELATED WORK

Recent advances in deep learning have shown that deep neural networks perform well for a wide variety of tasks, including image classification [12], object detection [13], semantic segmentation [14], etc. While consumer photos are the main focus, aerial images have been used in many computer vision applications, with the majority on geolocalization [2, 3, 15, 16, 17, 18, 19]. [15] introduced the cross-view image geolocalization problem, while [2, 3] used joint semantic features learned from deep convolutional neural networks for geolocalization and showed state-of-the-art performance. [19] performed ground-to-aerial image matching for robot self-localization using hand engineered features. [16] also performed ground-to-aerial image matching but to geo-register ground-level multiview stereo models. [18] performed event recognition by fusing information from ground images and co-located satellite images. [17] utilized building facades from building outlines for geolocalization. [20] introduced a well-thought-out aerial vehicle dataset, with 11 target categories including cars, boats, trucks, camping cars, (unknown) vehicles, airplanes, etc. However, the categories are not fine grained. [21] tried to leverage deep learning to detect and count cars from aerial images. [22] utilized aerial images to localize and orient ground-level query image. [23, 24, 25, 26] tried to detect vehicle from aerial images.

Fine-grained object recognition [5, 6, 7, 8, 9, 10, 27] is one of the most active areas of computer vision due to its broad range of applications. Birds [5], cars [6, 7], flowers [8], cats [9], and dogs [9, 10, 27] are the most commonly studied objects. However, most of these are small objects that are likely to end up in a few pixels in aerial images due to the much lower resolution and extreme viewpoints. We choose to investigate cars, but the same approach can be adopted to annotate similar objects such as boats and airplanes.

Fine-grained object recognition from aerial images has many important applications and could be a superior alternative to conventional approaches. For example, automatic fine-grained car recognition from aerial images can narrow down the search of a vehicle much faster than performing the same task on images from surveillance cameras. It can also enable fast surveying by comparing the distributions of car meta-data. For example, the distribution of car makes might reflect the social economical makeup of an area [28].

3. GROUND TO SKY LABEL TRANSFER

Labelling an aerial car image is more challenging for humans than labeling a street-view car image as there are much fewer useful cues. Also, we humans are not used to looking at objects from the top. For example, we can identify the type, make, and model of a car from a street-view image fairly easily. However, without additional information, it is likely to be difficult to identify them from aerial images.

In order to obtain high quality labels for aerial images, we propose to collect instance level paired aerial and street-view images. The street-view images are used by human annotators to get the fine-grained labels for aerial images. We use Google Maps to collect instance level paired data. As the Street-view and Earth-view images from Google Maps are likely to be captured at different time and cars are mobile targets, it is challenging to identify their correspondence through geographical locations.

Our idea is that the street and aerial views are likely to contain the same car if it is stationed for a long time, e.g., parked in a private driveway. We use the following heuristic criteria: residential areas, personal driveway, same location, and other information from these two views (e.g., color of the car). The intuition is that if a car is privately owned and the owner has a private driveway, it is more likely that the same car is parked at the same location. Therefore the chance of finding matched cars in the top and side views at the same location is high. After locating these possible matches, we then manually compare the cars from the two views and use other clues (e.g., color and shape) to make the final decision.

3.1. Data Collection

Figure 2 illustrates the whole process of collecting and annotating a cross-view matched car instance: 1) Start from a residential area; 2) Locate possible cars; 3) Match in street-views; 4) Get fine-grained labels from street-view images.

For each matched car, we collected 6 images, with 3 aerial images and 3 street-view ones. The reason we collected more than one image is to enable potential fine-grained pose estimation from street-view images and cross-resolution integration for aerial images. The 3 street-view images are taken from various viewpoints: left, right, and rear views. In very rare cases where a certain viewpoint is not achievable due to constraints from Google Maps, 2 or even 3 images are from the same viewpoint but with different angles. The 3 aerial images were taken at different resolutions in Google Maps, where the "resolution" in feet corresponds to the number shown at the bottom right corner of the Google Maps interface. For each car, we collected either 10 ft–20 ft–50 ft images, or 20 ft–50 ft–100 ft (if the 10 ft resolution was not available at that location).

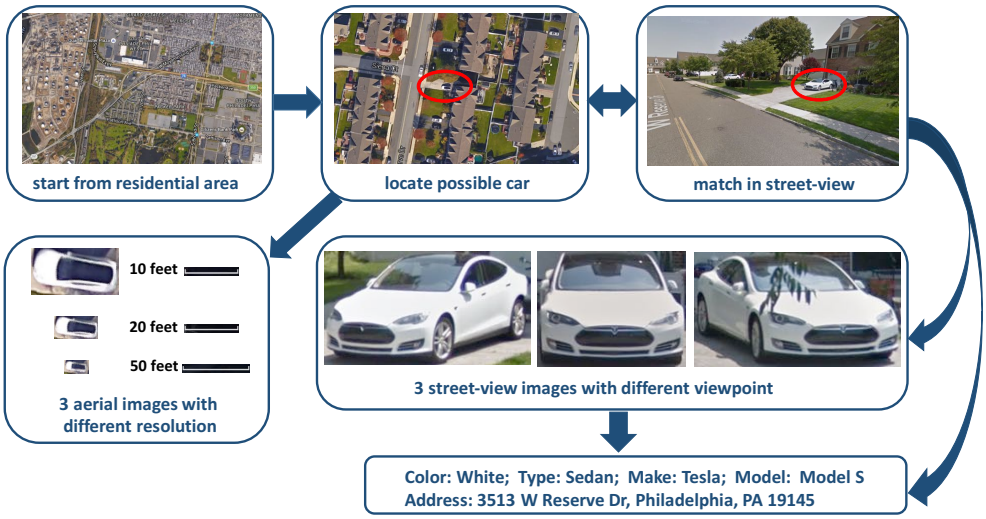

Fig. 2: Our data collection and annotation process. Starting from residential areas, we look for cars parked in personal driveways from the Earth-view of Google Maps. We then compare them with the street-view images manually. If there is a match, we extract 3 street-view images from different viewpoints and 3 aerial images with different resolutions. The labels are annotated based on street-view images.

3.2. Data Annotation

In this paper, we focus on the four most common tasks: car body color, car body type, car make, and car model. For car body color, we followed the annual surveys of PPG Industries and used these 9 fixed categories: Green, Yellow, White, Red, Black, Brown, Grey, Silver, and Blue. We followed National Appraisal Guides' division and used these 9 categories for body type: Convertible, Coupe, Hatchback, SUV, Sedan, Sports Car, Truck, Van, and Wagon. For car make and model, we started from an empty set and extended the label list over the entire data collection process. The annotation was done by undergraduate students who were trained extensively to make sure the label quality is as high as possible.

3.3. Dataset Statistics

We have so far collected 1221 cars with paired aerial and street-view images. While the collection process is still ongoing, here we report results with this initial dataset. Its size is much smaller than consumer-image datasets (e.g., ImageNet [29]). However, in contrast to consumer images, aerial images' viewpoint only differs in the in-plane rotation, and thus intraclass variation is limited.

4. EXPERIMENTS

In this paper, we conduct two sets of experiments. The first one tests the quality of labels in our dataset by calculating human agreement on meta labels and across views. The second experiment performs automatic fine-grained car recognition from aerial images. In this experiment, we compared the performance of human experts and deep learning approaches.

4.1. Human Agreement on Meta-labels

To measure the quality of our dataset, we perform a sanity check by comparing the agreement of two human annotators for each meta-label based on street view images. We randomly sampled 100 cars (street view) from the whole dataset for each meta-label and compare the agreement of two individuals without any collaboration. For each meta-label, the 100 cars might be different as some cars lack certain meta-labels. To better interpret the agreement ratio, we also performed the same experiment on the Stanford Cars dataset [6]. The Stanford Cars dataset contains 16,185 high quality images of 196 classes of cars. Our intuition is that the agreement ratio on Stanford Cars should be higher than on our data, as the images are of higher quality and have less ambiguity.

Agreement Ratio	Color	Type	Make	Model
Our Dataset	89%	92%	91%	90%
Stanford Cars [6]	89%	99%	95%	88%

Table 1: Human agreement ratios of meta-labels on our dataset (street view) and Stanford Cars dataset [6]. The comparable agreement ratios suggest that our annotations are of same high quality as those in Stanford cars.

	Color	Type	Make	Model
Training	621	625	390	73
Test	257	259	185	40

Table 2: Number of training and test data for each task.

From Table 1, we can that the agreement ratios on our dataset are comparable to those on Stanford Cars dataset, suggesting that our annotations are of similar high quality.

4.2. Matching Quality Across Views

We have shown that the quality of our labels transferred from street-view images is very high. To address further concern of mismatches between cars in aerial and street views, we conduct experiments to measure the human agreement of matching across views. We randomly sample 100 addresses from our dataset, show them to two additional human annotators and ask if the car is the same in both views. The high agreement ratio (99%) between all 3 annotators (the original annotators and the two additional ones) indicates very high matching likelihood.

Task	#Classes	Random Guess	Human	Human-Exp	SkyNet-A	SkyNet-L	SkyNet-A-A	SkyNet-L-A
Color	9	11.1%	63.5%	**68.9%**	65.6%	54.3%	64.8%	59.2%
Type	9	11.1%	45.6%	42.5%	45.9%	45.5%	**60.2%**	55.8%
Make	20	5.0%	17.8%	**20.5%**	18.9%	19.0%	19.5%	20.0%
Model	10	10.0%	20.0%	**25.0%**	20.0%	**25.0%**	22.5%	**25.0%**

Table 3: Classification accuracy for each independent meta-label task. We can see that even with relative small training, SkyNet can actually achieve comparable (color, make, and model) or even significant better (type) performance than human experts.

4.3. Fine-grained Car Recognition

We conduct experiments on fine-grained car recognition from aerial imagery and compare the performance of human annotators and deep CNN models. We train and test the model on a single aerial image per car, namely the 20 ft resolution, as it was the highest and most common resolution across the whole dataset. Due to the long-tail distribution of the dataset, some categories might contain one or two cars, especially for some rare car makes and models. This amount of data is likely not enough to train a good model. We thus picked the top 20 categories for car make and top 10 categories for car model. For color and type, as the amount of data in each category is relatively large, we used all the 9 categories. For each experiment on color, type, make, or model, we randomly sampled 70% of the cars as training data and the remaining 30% as test data. Table 2 shows the exact number of images of the training and test data for each task.

4.3.1. SkyNet: Fine-grained Car Recognition with Deep Neural Networks

We modified the commonly used AlexNet [12] deep convnet for our task. The output dimension of the last fully connected layer was changed to the number of categories for each sub-task accordingly and initialized with $\mathcal{N}(0, 0.01)$. For simplicity, each task is treated independently. We initialized the other layers from the parameters pre-trained on ImageNet [29]. In the training phase, we utilize mini-batch stochastic gradient descent (SGD) and set the base learning rate to be 10^{-3}, weight decay to be 5×10^{-4}, and momentum to be 0.9. We name the model fine-tuned on aerial imagery SkyNet. Two sets of deep CNN experiments are conducted for each task: fine-tuning All the layers (SkyNet-A) v.s. fine-tuning the Last layer only (SkyNet-L). Based on the intuition that the intraclass variation in aerial images of cars is very limited compared to street-view ones, we further Augment the data by transforming the cars to be vertically oriented and flip upside down (SkyNet-A-A and SkyNet-L-A for fine-tuning all layer and last layer). We use the Caffe deep learning framework [30] for all experiments with the default settings. During the training process, we find that the network begins to converge after 30,000 iterations and we report the accuracy at 40,000 iterations.

4.3.2. The "Daunting" Human Performance

On the contrary to most human baselines where the annotators were shown images directly without extensive training, we conducted two sets of human experiments for fair comparison. In the first experiment ('Human' column in Table 3), the human annotators were trained with street view images only. For the second one ('Human-Exp' column), they were trained with both street view images and aerial view images. Comparing Human with Human Expert, we can see that the performances are about the same. This result further confirms our conjecture that aerial imagery contains very little information and humans are not good at looking from the above.

4.3.3. Discussions

From the results in Table 3, we were surprised to find that with limited training data, SkyNet actually achieves comparable and even better performance than human experts in all four tasks. Not surprisingly, car body color is the easiest task for both humans and CNN models and their performance are comparable. Car body type is relatively easy to classify as well. One interesting thing is that there is a large gain (60.2% v.s. 45.9%) from data augmentation and the performance of SkyNet-A-A is significantly better than human or human expert (60.2% v.s. 45.6% or 42.5%). This further confirms our assumption that the intraclass variation is much smaller than in street-view images. For car make and model, neither human experts nor SkyNet achieves good performance compared to the other tasks. This might be due to the fact that different car makes usually produce similar car models (e.g., Toyota Corolla v.s. Honda Civic).

5. CONCLUSION

We investigated fine-grained car recognition from aerial imagery and proposed a novel method to collect the first fine-grained car dataset. We then compared the performance of humans and deep learning approaches. Noting the limited intraclass variation in aerial images, we further showed that deep CNN models achieve comparable or even significantly better performance than humans with limited data and simple data augmentation, compared to recognition from street-view images. Our initial evidence suggests that fine-grained car recognition in aerial images is not only feasible but also well suited for deep learning methods. We hope that the dataset will be used by the community as a standard benchmark.

Acknowledgments: This research is supported in part by a grant from the National Geospatial-Intelligence Agency; The authors would like to thank Anthony DiVirgilio, Anthony Vaccaro, Luke Beaulieu, and Raysa Rivera-Bergollo's help in collecting the data.

6. REFERENCES

[1] J. Yuan, S. S. Gleason, and A. M. Cheriyadat, "Systematic benchmarking of aerial image segmentation," *IEEE Geoscience and Remote Sensing Letters*, 2013. 1

[2] Scott Workman, Richard Souvenir, and Nathan Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *ICCV*, 2015. 1, 2

[3] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays, "Learning deep representations for ground-to-aerial geolocalization," in *CVPR*, 2015. 1, 2

[4] Joshua Gleason, Ara V Nefian, Xavier Bouyssounousse, Terry Fong, and George Bebis, "Vehicle detection from aerial imagery," in *ICRA*, 2011. 1

[5] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Tech. Rep., 2011. 1, 2

[6] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, "3d object representations for fine-grained categorization," in *ICCV Workshops*, 2013. 1, 2, 3

[7] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "A large-scale car dataset for fine-grained categorization and verification," in *CVPR*, 2015. 1, 2

[8] M-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 1, 2

[9] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *CVPR*, 2012. 1, 2

[10] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei, "Novel dataset for fine-grained image categorization," in *CVPR Workshops*, 2011. 1, 2

[11] Branson S. Hall D. Schindler K. Perona P. Wegner, J.D., "Cataloging public objects using aerial and street-level images urban trees," in *CVPR*, 2016. 1

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012. 2, 4

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014. 2

[14] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015. 2

[15] T. Y. Lin, S. Belongie, and J. Hays, "Cross-view image geolocalization," in *CVPR*, 2013. 2

[16] Qi Shan, Changchang Wu, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M Seitz, "Accurate geo-registration by ground-to-aerial image matching," in *3DV*, 2014. 2

[17] Mayank Bansal, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis, "Geo-localization of street views with aerial image databases," in *ACM Multimedia*, 2011. 2

[18] Jiebo Luo, Jie Yu, Dhiraj Joshi, and Wei Hao, "Event recognition: viewing the world with a third eye," in *ACM Multimedia*, 2008. 2

[19] Anirudh Viswanathan, Bernardo R Pires, and Daniel Huber, "Vision based robot localization by ground to satellite matching in gps-denied situations," in *IROS*, 2014. 2

[20] Sebastien Razakarivony and Frederic Jurie, "Vehicle detection in aerial imagery (vedai) : a benchmark," Tech. Rep., 2015. 2

[21] T. Nathan Mundhenk, Goran Konjevod, Wesam A. Sakla, editor="Leibe Bastian Boakye, Kofi", Jiri Matas, Nicu Sebe, and Max Welling, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *ECCV*. 2

[22] Nam N. Vo and James Hays, "Localizing and orienting street views using overhead imagery," in *ECCV*, 2016. 2

[23] Liujuan Cao, Qilin Jiang, Ming Cheng, and Cheng Wang, "Robust vehicle detection by combining deep features with exemplar classification," in *Neurocomputing*, 2016. 2

[24] Zhaohui H. Sun, Mathew Leotta, Anthony Hoogs, Rusty Blue, Robert Neuroth, Juan Vasquez, Amitha Perera, Matthew Turek, and Erik Blasch, "Vehicle change detection from aerial imagery using detection response maps," in *SPIE*, 2014. 2

[25] Cheng-Lin Liu Xueyun Chen, Shiming Xiang and Chun-Hong Pan, "Vehicle detection in satellite images by parallel deep convolutional neural networks," in *ACPR*, 2013. 2

[26] Cheng-Lin Liu Xueyun Chen, Shiming Xiang and Chun-Hong Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," in *IEEE GEOSCIENCE AND REMOTE SENSING LETTERS*, 2014. 2

[27] Jiongxin Liu, Angjoo Kanazawa, David Jacobs, and Peter Belhumeur, "Dog breed classification using part localization," in *ECCV*, 2012. 2

[28] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei, "Visual census: Using cars to study people and society," in *Bigvision*, 2015. 2

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009. 3, 4

[30] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014. 4