Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Geolocating Images with Crowdsourcing and Diagramming*

Rachel Kohler, John Purviance, Kurt Luther
Department of Computer Science and Center for HCI, Virginia Tech
{rkohler1, ffdd4846, kluther}@vt.edu

Abstract

Many types of investigative work involve verifying the legitimacy of visual evidence by identifying the precise geographic location where a photo or video was taken. Professional geolocation is often a manual, time-consuming process that can involve searching large areas of satellite imagery for potential matches. In this paper, we explore how crowdsourcing can be used to support expert image geolocation. We adapt an expert diagramming technique to overcome spatial reasoning limitations of novice crowds so that they can support an expert's search. In an experiment (n=540), we found that diagrams work significantly better than ground-level photos and allow crowds to reduce a search area by half before any expert intervention. We also discuss hybrid approaches to complex image analysis combining crowds, experts, and computer vision.

1 Introduction

Photos, videos, and other visual evidence documents the world we live in and provides a foundation for modern investigations in journalism, law enforcement, and human rights advocacy. This imagery is increasingly distributed online through social media. Governments post photos of political events, terrorist organizations share propaganda, and everyday people use smartphones to document crimes, natural disasters, and other important events.

Because visual evidence can be so compelling, it must be treated with skepticism. Photos and videos can be edited or shared with misleading contextual information, intentionally or by accident. *Image verification* is the challenging process of determining if imagery is what its surrounding context claims it to be; or if not, what it actually depicts [Barot, 2014]. One of the key subtasks of image verification is *geolocation*, which involves mapping the precise location in the world where a photo or video was made. Geolocation allows the investigator to determine where the image was actually made, and compare that with contextual claims about its meaning and purpose.

*This paper is an abridged version of a paper titled "Supporting Image Geolocation with Diagramming and Crowdsourcing" that won the Notable Paper Award at AAAI HCOMP 2017.

Expert geolocators draw on many skills and resources to make these determinations [Higgins, 2014; Kohler and Luther, 2017]. The process is often manual, and sometimes tedious. Experts inspect the image for clues, such as familiar landmarks, weather, architecture, and landscapes. Text and graphics, such as logos and road signs, can often be researched online to narrow down possibilities. When these clues are not definitive, expert geolocators often turn to diagramming and satellite image analysis. They first draw an aerial diagram of the ground-level image under investigation, a spatial reasoning skill requiring substantial practice. Then, they use commercial GIS services like Google Maps to systematically search the area for distinctive buildings, roads, or other structures matching their diagram. Depending on the size and density of the search area, this process can require hours or days even for experts, and may still prove fruitless. If the image cannot be geolocated, it may not be verifiable.

In this paper, we explore how crowdsourcing can support this geolocation process, with the goal of helping an expert locate an image faster and more accurately. Crowds have proven to be effective at analyzing satellite imagery, but novice crowds lack an expert's spatial reasoning skill in recognizing ground-level features from aerial imagery. We close the gap by leveraging the diagramming technique from expert practice and adapting it for novice crowds to improve their satellite image analysis.

We demonstrate the value of our approach in a large-scale experiment (n=540). We find that giving crowds a ground-level photo results in unacceptably poor performance, but an aerial diagram significantly improves their performance to near-perfect levels. Our crowdsourcing technique can reduce a geolocation search area by half in about 10 minutes while finding the target area 98.3% of the time. We also discuss the real-world applications and next steps for this work, including new opportunities to leverage the complementary strengths of crowds, experts, and computer vision, for complex image analysis tasks like geolocation.

2 Related Work

2.1 Computer Vision Approaches to Geolocation

Image geolocation is a longstanding problem of interest for computer vision researchers. IM2GPS [Hays and Efros, 2008] compares features in a ground photo to a reference

dataset of 6.4 million geolocated Flickr images and outputs a distribution of the most probable regions of the earth. More recently, PlaNet [Weyand *et al.*, 2016] takes a photo as input and, using a convolutional neural network trained on 126 million geotagged photos from the web, generates a probability for 26,000 cells in a grid covering the earth. Computer vision approaches like IM2GPS and PlaNet cannot yet consistently achieve the point-level specificity typically required for verification work, but may provide excellent starting points for expert geolocators.

Other work in computer vision seeks to bridge the gap between satellite and ground level imagery. [Ghouaiel and Lefèvre, 2016] developed a technique to automatically translate ground photos into aerial perspectives, but the approach requires panoramic photos and overall translation accuracy was 54%. [Zhai *et al.*, 2016] trained a neural network to generate ground-level panoramas from satellite imagery. Their approach shows promise, but had limited effectiveness in handling high variability features like buildings. Unlike these approaches, we bootstrap an expert diagramming technique to translate between ground and satellite images.

Combining elements from the above categories, Where-CNN [Lin *et al.*, 2015] used cross-view pairs of ground-level and 45° aerial imagery to train a neural network to localize ground-level photos. Their approach narrowed the location estimate to 1% of the search area for 7–22% of query images (depending on the city). While 45° imagery is not yet available in many areas, these automated results provide a point of comparison to our crowdsourced results.

2.2 Crowdsourced Image Analysis

Due to the impressive capabilities of the human vision system, crowds have been used to perform a variety of visual recognition tasks [Bigham *et al.*, 2010; Noronha *et al.*, 2011]. Many of these applications rely on crowds to identify everyday objects, scenes, or locations that do not require specialized knowledge. However, tools like scaffolding and computer vision have been used to help novice crowds analyze less familiar content, like graphic designs [Greenberg *et al.*, 2015] or accessibility issues [Hara *et al.*, 2015]. A rich source of examples comes from citizen science, where novice crowds recognize and categorize diverse natural phenomena [Wiggins and Crowston, 2014].

Satellite image analysis often leverages crowdsourcing, especially for humanitarian efforts like locating missing persons or assessing damage from natural disasters [Meier, 2015]. Studies of these projects emphasize the challenges novices face in translating their own observations into abstract representations [Zacks *et al.*, 2000]. To overcome these challenges, researchers recommend partnerships between experts and novices [Kerle and Hoffman, 2013; Bianchetti and MacEachren, 2015], an idea we explore in this paper.

2.3 Expert Image Geolocation and Diagramming

Many types of professionals perform image geolocation, including journalists, intelligence analysts, human rights advocates, and private investigators [Barot, 2014; Brandtzaeg *et al.*, 2016]. [Kohler and Luther, 2017] conducted an interview study with geolocation experts in diverse fields, focusing on

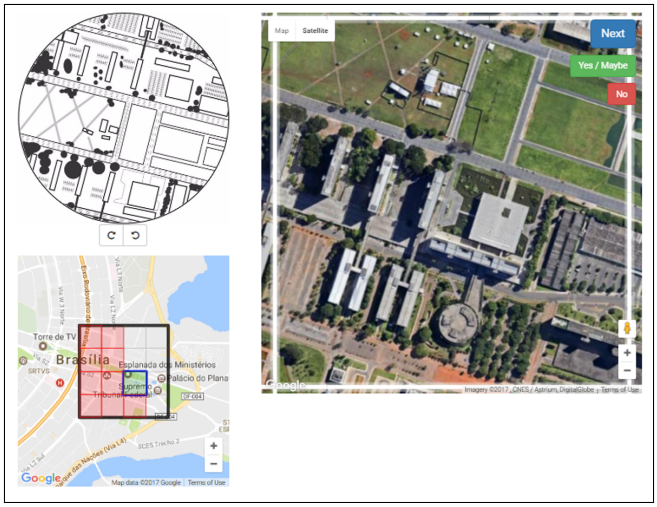

Figure 1: The crowd interface.

their motivations, process, and use of crowdsourcing. Interviewees emphasized the importance of drawing diagrams as a tool for converting a ground-level photo into a more effective abstraction. One expert said he would *"draw a bird's eye perspective, or a satellite image perspective, of how I think it may look like from the air. So I can then compare it with satellite imagery just to get a better impression."* Another expert emphasized the difficulty of this mental translation: *"Perspective distortion can throw off a novice or a beginner really easily because things that you see from the air tend not to look how you would think they would from the ground."* These observations align with psychological research showing that people with high spatial ability use different cognitive strategies for mental rotation tasks [Just and Carpenter, 1985].

Building on these findings, we consider how diagramming can be adapted for crowds who lack an expert geolocator's spatial thinking skills. In the following experiment, we investigate whether giving a crowd an aerial diagram or a ground-level photo leads to better geolocation results. We hypothesize that the diagram will yield higher true positive rates because it distills the most important features, but it will also yield higher false positives because an abstraction can potentially match more areas due to lack of discriminating details.

3 Study

3.1 System Design

We built a web-based system using a Python/Django framework, a PostgreSQL database, and the Google Maps API for satellite imagery and GIS functions.

The main component of the system is the crowd interface (Fig. 1). The top left of the interface shows a type of reference material, depending on the study condition. In the photo-only condition, it shows a ground-level photo. In the diagram-only condition, it shows an aerial diagram. In the both condition, there is a toggle that allowed the user to switch between the diagram and the ground photo.

The bottom left shows a small Google Map (in Map mode) of a region with a 4×4 grid of 16 equal-sized subregions over-

laid in black lines. After experimenting with different-sized regions, we found that a 4×4 grid struck an effective balance between context and effort.

The right side of the interface shows a Google Map (in Satellite mode) of the region, divided by translucent white lines into the same 4×4 grid. The user is confined to one subregion at a time, but can zoom in and out and toggle Map / Satellite mode. The user clicks a green `Yes/Maybe` button if it looks like a potential match, or a red `No` button if it does not, and then clicks `Next`. This advances the user to the next subregion, and marks in either red or green the corresponding subregion in the small map. The system advanced through the subregions in a Creeping Line search pattern, following best practices used in search and rescue [Wollan, 2004].

3.2 Locations

We used three locations for the study. BSB showed a crowded area near the Monumental Axis in Brasília, Brazil. CLT was a highway near an overpass in Charlotte, NC, USA. LAX showed an intersection with crosswalks in downtown Los Angeles, CA, USA. We selected these locations and corresponding ground photos from a set of geolocation training materials prepared by an expert. Our selection criteria included similarly moderate difficulty and geographic and visual diversity.

3.3 Diagrams

We also needed a set of aerial diagrams corresponding to the above locations to compare to the ground-level photos. We considered using existing expert-drawn diagrams, but differences across experts and locations would be difficult to control in our experimental setting. Instead, we designed a set of diagram-drawing guidelines, informed by expert practice [Kohler and Luther, 2017; Higgins, 2014] and relevant standards [Painho *et al.*, 2010; Kolbe *et al.*, 2005], and used these guidelines to draw a low-, medium-, and high-detail diagram for all three locations using Adobe Photoshop. Low-detail diagrams showed streets, roadways, and pathways. Medium-detail diagrams showed road markings and building outlines plus low-detail features. High-detail diagrams showed vegetation and street-level details (e.g., parking) plus medium- and low-detail features. In the expanded version of this paper, we report on a second experiment comparing these levels of detail, finding that medium-detail diagrams produced slightly better results than the others [Kohler *et al.*, 2017]. Therefore, we used medium-detail diagrams in this experiment comparing diagrams to ground-level photos.

3.4 Experiment Design

The study was a between-subjects experiment. The independent variable was reference material with three levels: diagram only, photo only, or both. Location depicted in the diagram was a covariate with three levels: BSB, LAX, or CLT. Therefore, there were nine possible conditions. The dependent variables were the participants' binary judgements on each of the 16 subregions.

We recruited participants from Amazon Mechanical Turk (MTurk). We randomly assigned each worker to one of the nine conditions, and we assigned 60 workers to each condition, for a total of 540 workers. Pilots showed that workers

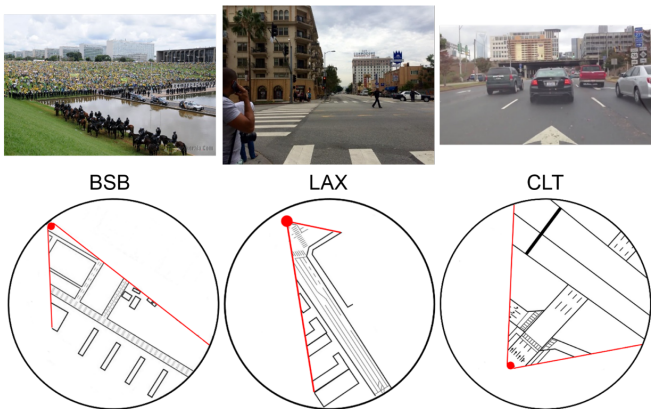

Figure 2: Ground-level photos and aerial diagrams.

took an average of five minutes to complete the task, so we paid $1.21 per task, reflecting minimum wage in our location for 10 minutes of work. We restricted the task to US-based workers but used no other qualifications.

Task and Procedure

After accepting the task, each participant completed an online IRB consent form and a short, self-paced tutorial. The participant then proceeded to examine each of 16 subregions in the grid and mark it as `Yes/Maybe` or `No`. Based on average completion times in pilot studies, we set the time limit at 10 minutes to encourage fast responses. All tasks had exactly one correct subregion and 15 distractors.

We took care to design the crowd task and interface to be as realistic as possible from the worker's perspective. Workers did not know whether their region contained a correct subregion, and received no feedback on their judgements. Therefore, the worker experience would be the same for real-world scenarios where it was unknown whether the region contained a correct subregion.

Data Cleaning and Analysis

In our pilot studies, individual workers showed high variance in task performance. We experimented with different aggregation strategies and found that forming triads (groups of three workers) with a one-yes rule yielded the best results. The one-yes rule means that if at least one of the three workers judged a subregion to be a `Yes/Maybe`, then it would be categorized as a yes, while only a unanimous judgement of `No` across all three workers would be categorized as a no. We randomly grouped the 60 workers for each condition into 20 triads per condition in the results that follow.

Next, we compared each triad's judgement to our gold standard to calculate true positives and false positives. We used these measures rather than precision and recall because geolocation is a needle-in-the-haystack problem, where false negatives are much worse than false positives.

We performed statistical analyses in R. Shapiro–Wilk tests showed that the dependent variables failed a normality assumption, so we used Kruskal–Wallis tests as a nonparametric alternative to ANOVAs. We used Dunn's tests to perform post-hoc analyses, with Bonferroni correction to adjust p-values for multiple comparisons.

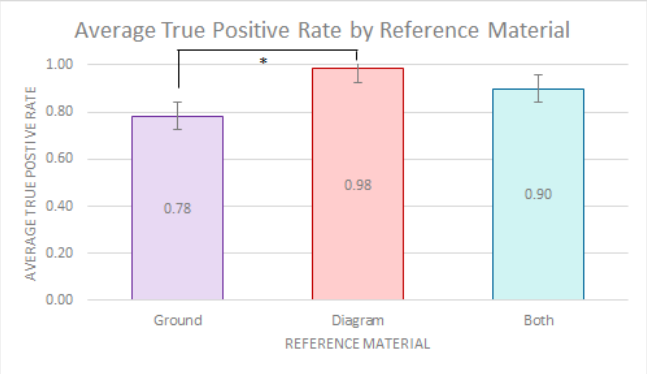

Figure 3: True positives by reference material. Crowd workers using only an aerial diagram performed significantly better than workers who could see only the ground-level photo.

4 Results

For true positives (Fig. 3), the diagram-only condition performed best, with 98.3% of triads marking the correct subregion. The both condition performed slightly less well with a 90% success rate. Ground photo-only trailed behind with 78.3% of triads finding the correct subregion. Reference material had a significant effect on true positives, $\chi^2(1) = 4.111$, $p < 0.05$. Post-hoc analysis showed that true positives were significantly higher for diagram-only compared to ground photo-only, $z = 3.476$, $p < 0.01$. Also, the both condition performed marginally significantly better than ground photo-only, $z = 2.028$, $p = 0.128$. There was no significant difference in true positives for the both condition vs. diagram-only, $z = -1.448$, $p = 0.443$.

For false positives, the crowd generally reduced the search area by about half, regardless of reference material. Diagram-only workers produced slightly more false positives (M = 51.3%), followed by both (M = 48%), and then photo-only (47%), but the differences were not significant, $\chi^2(15) = 12.691$, $p = 0.626$.

5 Discussion

This study investigated how the type of reference material affected the crowd's geolocation performance. We found that the diagram by itself results in significantly higher true positives compared to the ground photo by itself. The diagram allows crowds to achieve near-perfect performance (98.3% of triads found the correct subregion), whereas only 78.3% of triads found it in the ground photo-only condition. Our intuition is that experts would be unlikely to trust a crowd that misses the target one out of every five times, so the ground photo by itself may not be a viable approach.

False positives were around 50% for all conditions and reference material did not have a significant effect. This means that in all cases, the crowd reduced the search area by about half. More importantly, for the diagram-only condition, the search area was cut in half while still including the correct subregion 98.3% of the time. Further, the significance test showed that quality is not a zero-sum game: the diagram condition's excellent true positives do not come at a cost of more false positives.

The above results indicate that the aerial diagram yields a significant improvement in quality for crowdsourced image geolocation. When a triad of workers is shown just the ground-level photo, they miss the correct subregion one out of every five times. When the photo is replaced with an aerial diagram, crowds found the target 98.3% of the time, with no increase in false positives. Crowds also reduce the search area by half. Therefore, the evidence suggests that crowds provided with a diagram could substantially augment an expert's image geolocation process.

6 Conclusion and Future Work

This paper explored how crowdsourcing could support experts in a geolocation task. We contributed a new diagramming technique, adapted from expert practice, that can be used to help novice crowds more effectively analyze satellite imagery. We also contributed a large-scale crowdsourcing experiment demonstrating the value of our technique. We found that aerial diagrams are significantly better than ground-level photos in supporting crowdsourced satellite image analysis.

As our study locations were all city-based, our results primarily speak to urban geolocation tasks. Our approach may also extend to rural areas, which share many task characteristics with geolocation of urban imagery, but also face some distinct challenges, such as scarcity vs. overabundance of image clues [Mehta *et al.*, 2016].

The approach presented here seeks to minimize expert intervention, but future work is needed to understand how best to integrate the crowd's judgements into an expert's workflow. Other opportunities involve leveraging computer vision tools to support crowds and experts, such as:

- **Context identification**: Many images on social media do not have surrounding context suggesting a general location. Systems like PlaNet could suggest high-probability sectors to narrow the search space for crowds.
- **Diagram generation**: Ground-to-aerial systems like Where-CNN could not only suggest potential location matches, but also extract distinctive features to help experts build a diagram more quickly and accurately.
- **Image comparison**: Sketch recognition systems like Google's Quick, Draw! could compare an expert diagram to satellite imagery and return potential matches.

We propose that hybrid pipelines or mixed-initiative systems composed of crowds, experts, and algorithms, each complementing the others with its unique strengths, offer the greatest potential to support complex image analysis and sensemaking. This paper offers a glimpse of these possibilities in demonstrating how novice crowds can augment the work of experts in geolocation and verification tasks.

Acknowledgments

We thank the study participants, the anonymous reviewers, and members of the Crowd Intelligence Lab. This research was supported by NSF IIS-1527453 and IIS-1651969.

References

[Barot, 2014] Trushar Barot. Verifying Images. In *Verification Handbook: A Definitive Guide to Verifying Digital Content for Emergency Coverage*. January 2014.

[Bianchetti and MacEachren, 2015] Raechel A. Bianchetti and Alan M. MacEachren. Cognitive Themes Emerging from Air Photo Interpretation Texts Published to 1960. *ISPRS International Journal of Geo-Information*, 4(2):551–571, April 2015.

[Bigham *et al.*, 2010] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. VizWiz: Nearly Real-time Answers to Visual Questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 333–342, New York, NY, USA, 2010. ACM.

[Brandtzaeg *et al.*, 2016] Petter Bae Brandtzaeg, Marika Lüders, Jochen Spangenberg, Linda Rath-Wiggins, and Asbjørn Følstad. Emerging Journalistic Verification Practices Concerning Social Media. *Journalism Practice*, 10(3):323–342, April 2016.

[Ghouaiel and Lefèvre, 2016] Nehla Ghouaiel and Sébastien Lefèvre. Coupling ground-level panoramas and aerial imagery for change detection. *Geo-spatial Information Science*, 19(3):222–232, July 2016.

[Greenberg *et al.*, 2015] Michael D. Greenberg, Matthew W. Easterday, and Elizabeth M. Gerber. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In *Proceedings of ACM Creativity & Cognition 2015*, Glasgow, Scotland, 2015. ACM.

[Hara *et al.*, 2015] Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L. Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H. Ng, and Jon E. Froehlich. Improving Public Transit Accessibility for Blind Riders by Crowdsourcing Bus Stop Landmark Locations with Google Street View: An Extended Analysis. *ACM Trans. Access. Comput.*, 6(2):5:1–5:23, March 2015.

[Hays and Efros, 2008] J. Hays and A. A. Efros. IM2gps: estimating geographic information from a single image. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[Higgins, 2014] Eliot Higgins. A Beginner's Guide to Geolocating Videos, July 2014.

[Just and Carpenter, 1985] Marcel Adam Just and Patricia A. Carpenter. Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, 92(2):137–172, April 1985.

[Kerle and Hoffman, 2013] N. Kerle and R. R. Hoffman. Collaborative damage mapping for emergency response: the role of Cognitive Systems Engineering. *Nat. Hazards Earth Syst. Sci.*, 13(1):97–113, January 2013.

[Kohler and Luther, 2017] Rachel Kohler and Kurt Luther. Crowdsourced Image Geolocation as Collective Intelligence. In *Collective Intelligence 2017*, New York, NY, USA, 2017.

[Kohler *et al.*, 2017] Rachel Kohler, John Purviance, and Kurt Luther. Supporting Image Geolocation with Diagramming and Crowdsourcing. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*, September 2017.

[Kolbe *et al.*, 2005] Thomas H. Kolbe, Gerhard Gröger, and Lutz Plümer. CityGML: Interoperable Access to 3d City Models. In Professor Dr Peter van Oosterom, Dr Siyka Zlatanova, and Elfriede M. Fendel, editors, *Geoinformation for Disaster Management*, pages 883–899. Springer Berlin Heidelberg, 2005.

[Lin *et al.*, 2015] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning Deep Representations for Ground-to-Aerial Geolocalization. In *Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015.

[Mehta *et al.*, 2016] Sneha Mehta, Chris North, and Kurt Luther. An Exploratory Study of Human Performance in Image Geolocation Tasks. In *HCOMP 2016 GroupSight Workshop on Human Computation for Image and Video Analysis*, Austin, TX, USA, 2016. AAAI.

[Meier, 2015] Patrick Meier. *Digital Humanitarians: How Big Data Is Changing the Face of Humanitarian Response*. Routledge, Boca Raton, FL, null edition edition, January 2015.

[Noronha *et al.*, 2011] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 1–12, New York, NY, USA, 2011. ACM.

[Painho *et al.*, 2010] Marcos Painho, Maribel Yasmina Santos, and Hardy Pundt, editors. *Geospatial Thinking*. Springer, 2010.

[Weyand *et al.*, 2016] Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet - Photo Geolocation with Convolutional Neural Networks. *arXiv:1602.05314 [cs]*, February 2016. arXiv: 1602.05314.

[Wiggins and Crowston, 2014] Andrea Wiggins and Kevin Crowston. Surveying the citizen science landscape. *First Monday*, 20(1), December 2014.

[Wollan, 2004] Helen Wollan. Incorporating Heuristically Generated Search Patterns in Search and Rescue. Master's thesis, University of Edinburgh, 2004.

[Zacks *et al.*, 2000] Jeffrey M. Zacks, Jon Mires, Barbara Tversky, and Eliot Hazeltine. Mental spatial transformations of objects and perspective. *Spatial Cognition and Computation*, 2(4):315–332, December 2000.

[Zhai *et al.*, 2016] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting Ground-Level Scene Layout from Aerial Imagery. *arXiv:1612.02709 [cs]*, December 2016. arXiv: 1612.02709.