Sky Segmentation in the Wild: An Empirical Study

Radu P. Mihail[1] rpmihail@valdosta.edu
Scott Workman[2] scott@cs.uky.edu
Zach Bessinger[2] zach@cs.uky.edu
Nathan Jacobs[2] jacobs@cs.uky.edu
[1]Valdosta State University
[2]University of Kentucky

Abstract

Automatically determining which pixels in an image view the sky, the problem of sky segmentation, is a critical pre-processing step for a wide variety of outdoor image interpretation problems, including horizon estimation, robot navigation and image geolocalization. Many methods for this problem have been proposed with recent work achieving significant improvements on benchmark datasets. However, such datasets are often constructed to contain images captured in favorable conditions and, therefore, do not reflect the broad range of conditions with which a real-world vision system must cope. This paper presents the results of a large-scale empirical evaluation of the performance of three state-of-the-art approaches on a new dataset, which consists of roughly 100k images captured "in the wild". The results show that the performance of these methods can be dramatically degraded by the local lighting and weather conditions. We propose a deep learning based variant of an ensemble solution that outperforms the methods we tested, in some cases achieving above 50% relative reduction in misclassified pixels. While our results show there is room for improvement, our hope is that this dataset will encourage others to improve the real-world performance of their algorithms.

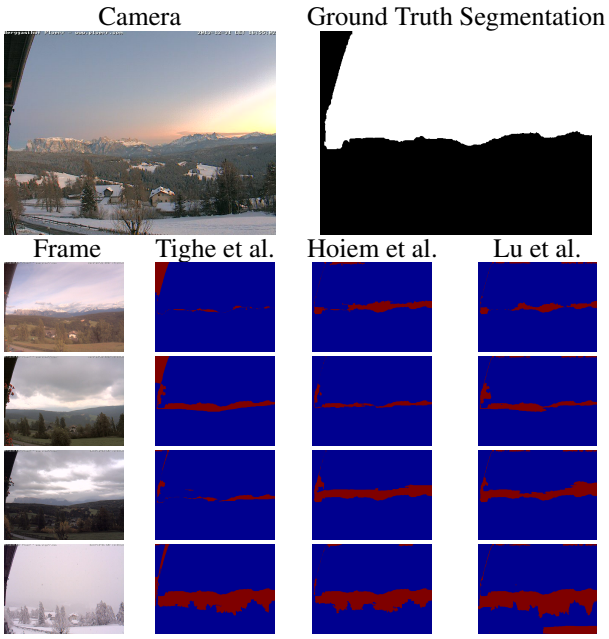

Figure 1. One scene, observed over a long period of time, can change drastically in appearance due to time of day, day of year and weather. Above, we show six sample scenes from one static outdoor webcam and the results of three sky labeling methods (blue represents correct classification, red represents incorrect classification).

1. Introduction

Image labeling algorithms assign a label (e.g., car, ground, sky, building) to every pixel in an image. Outdoor imagery captured "in the wild" poses challenges to these algorithms due to the variety of possible lighting and weather conditions. We focus on sky segmentation in single images which, while seemingly simple, is actually a very challenging and unsolved problem. Outdoor scene labeling has received much attention from vision research in the past few years, since it is an important pre-processing step for many high-level vision algorithms. Existing approaches perform well in favorable conditions (e.g., clear blue sky), however the effects of weather, season, and time drastically alter the appearance of an outdoor scene (Figure 1). This paper presents an extensive evaluation of three existing methods [3, 12, 17] on a challenging real-world dataset.

We selected these methods for both their contributions to the community and their performance on benchmark datasets. However, benchmark datasets do not completely capture the variational appearance of the sky. Camera optics combined with the dynamics of sun position, cloud formations, and more, results in tremendous variability of sky appearance. The combination of these factors motivates the construction of a dataset representative of these real-world scenarios.

To build such a dataset, we take advantage of imagery collected from static outdoor cameras over long periods of

time, thus increasing the probability of exposure to different atmospheric conditions. Since the cameras we use are static, one ground truth mask can be used to generate a large training set with significant sky appearance variability. In order to highlight various conditions in which labeling performance is suboptimal, we augment each image with weather observations collected from nearby weather stations. Often in real-world applications, it may not be possible to obtain weather data, so we also compute high-level transient scene attributes, related to weather and appearance, for each image using the method of Laffont et al. [10].

Using this dataset, we perform several experiments to examine the impact that weather and time has on each methods performance. The key contributions of this work are: 1) introducing a large labeled image dataset; 2) presenting the results of a large-scale empirical evaluation of three state-of-the-art techniques for sky segmentation; 3) suggesting how the observations we make from our evaluation should guide future work on sky segmentation, and more generally pixel labeling; 4) a deep ensemble method that combines raw image data and existing methods' output to make better predictions.

2. Related work

Recent interest in outdoor imagery has led to various algorithms for calibration [8, 20], labeling [3, 18], geolocation [7, 22], geometry recovery [2, 8] and others. For example, Jacobs et al. [5] exploit pixel time-series from static webcams to recover a cloudmap and an estimate of scene geometry. Cues from multiple images (e.g., moving clouds, sun position) can be used as input to higher level vision algorithms (e.g., calibration [20], scene shape [4, 21], horizon estimation, geolocation [7]), but automatic detection of sky regions is difficult from single images. Weather detection and understanding has been successfully used in robot vision for navigation [9], driver assistance systems [14, 23] and image searching [16].

Scene labeling methods attempt to assign each pixel in an image to one of several categories of objects (e.g., sky, ground, road, tree, building). These methods [3, 11, 17, 18, 19] rely on the local appearance of the objects learned from a training set of instances. The scene parsing problem is most commonly addressed by a local classifier (using engineered features, or more recently, learning features using deep learning architectures) constrained by a graphical probability model (e.g., CRF or MRF) where global decisions are made to include high-level reasoning about spatial relationships.

In this work, we present an empirical assessment of the performance of sky segmentation in the wild. Our work is most similar in conception to that of Stylianou et al. [15] who analyze feature matching performance over long time periods. We evaluate three methods that output (either directly or as a by-product) a sky segmentation: Hoiem et al. [3], Tighe et al. [17] and Lu et al. [12]. The choice for these methods was motivated by their impact in the vision community and publicly available code. The contribution relevant to sky segmentation in the work of Hoiem et al. [3] is their use of geometric context (e.g.: a rough estimate of scene geometry) for three classes (sky, ground and vertical) inferred using statistical appearance learning. We only use the geometric label outputs from their algorithm. Tighe et al. [17] introduce an image parsing method based on combining region-level features with exemplar-SVM sliding window detectors. We only use the "sky" label from the final results of their algorithm (i.e., other labels are considered "not sky"). Lu et al. [12] explore single image weather classification into two classes: sunny or cloudy. We use their sky segmentation output in our evaluation.

3. Dataset

We introduce a new dataset [1] of labeled outdoor images captured in a wide range of weather and illumination conditions. To the best of our knowledge, this dataset is the largest in existence with annotated sky pixels and associated weather data. Motivated by difficulties of existing methods to handle extreme appearance variations of sky regions, we take advantage of the long-term webcam imagery from the Archive of Many Outdoor Scenes (AMOS) [6]. We selected 53 cameras from AMOS that were static (i.e., no camera movement throughout one or more calendar years) and downloaded all the available images for that period, (an average of one year, with around ten thousand images per camera). To keep the dataset size reasonable, we keep five randomly selected frames for each day. For each camera we manually created a binary mask segmenting sky and ground. The average coverage of sky pixels for all the webcams is 41.19%, with standard deviation 15.71%.

To quantify the effect of weather condition on labelers, we retrieved weather data from `Wunderground.com` for all of our cameras. Weather data was retrieved at every camera location for the entire period of the downloaded imagery, and the closest observation was associated with each image. For each frame we have indicators for several weather conditions: light rain, partly cloudy, clear, haze, scattered clouds, mostly cloudy, overcast, fog and light freezing fog.

In addition, we augment each image in the dataset with Laffont et al.'s [10] "transient attributes", which are high-level properties describing the appearance of a scene, identified via crowdsourcing on thousands of images from 101 webcams. We take advantage of their learned regression models to extract transient attributes for all of our images.

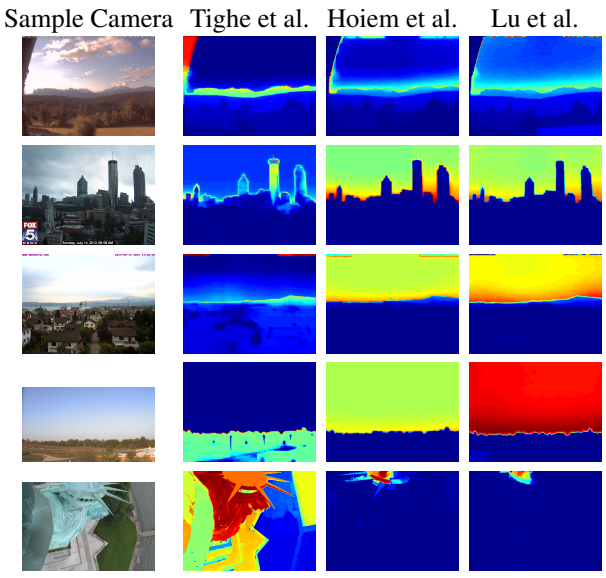

Figure 2. Sample images from five cameras (left-most column). In the right columns, we show cumulative per-pixel MCR for all images, for all three methods evaluated. Red and blue indicate high and low, respectively, misclassification.

4. Experiments

We processed every image in our dataset through the previously mentioned methods using the source code provided by their respective authors. While each method was pre-trained, running approximately 100,000 images through each method is computationally challenging. Our attempt to compensate for the large-scale processing of each method involved distributing the computation over a 16-node cluster for several weeks.

Running each method on tens of thousands of images simply led to software crashes. To minimize this, we optimized various components of the methods, which allowed us to successfully run this large scale evaluation. There were some images that led to failure. These images were either completely dark, saturated, or with reduced visibility due to fog, etc. If a method failed for a given image, we excluded the image from the evaluation set. The tunable parameters used in this evaluation were the same as the original authors in their own evaluations.

4.1. Overall Accuracy

To evaluate the methods we use a per-pixel performance metric, the misclassification rate (MCR), computed per frame as follows: $MCR = \frac{\text{\# of incorrectly classified pixels}}{\text{total \# of pixels}}$. Each of the 53 webcams has roughly 1500 images. In Figure 2 we show five sample cameras (left column) with average MCR for every method (right three columns). Overall, Tighe et al. achieved the best performance, with lowest average MCR of 16.41%, $\sigma = 18.98\%$. Hoiem et al. came in second place, with average MCR of 20.69%, $\sigma = 22.13\%$. Finally, Lu et al. with average MCR of 27.69%, $\sigma = 23.50\%$.

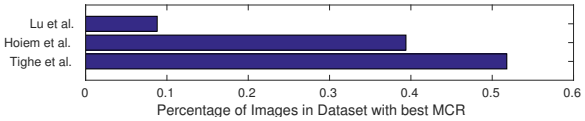

Figure 3. Distribution of best MCR per image with respect to method for the entire dataset.

While MCR is a good overall performance indicator, we also report type I errors (false positive, pixel is labeled as sky when it is not) and type II (false negative, a pixel is labeled as ground when it is not) for each method. As seen in Table 1, Lu et al. and Hoiem et al. have significantly less false positive errors than Tighe et al. who achieved lower overall MCR. To gain further insight into the methods, we counted the number of images in the dataset for each method with the best MCR. While the overall lowest MCR was achieved by Tighe et al., their method was outperformed on slightly less than half of the images in our dataset. The lowest MCR was achieved on roughly 40% of the images by Hoiem et al., and 10% by Lu et al., as seen in Figure 3.

We now consider individual pixels. It is often the case pixels are independently labeled correctly by at least one method. We compute the per-pixel MCR, by counting a pixel as correctly classified when either one of the three methods labeled it correctly, and incorrect otherwise. Overall, the per-pixel MCR is 1.9%. This suggests that improvements can be achieved using the methods results as a strong prior, combined with other factors that were found to have an impact on accuracy.

4.2. Impact of Lighting on Accuracy

We observe that lighting conditions have a significant impact on the accuracy of scene labeling methods. When the sun is at its highest point in the sky, the scene is most equally illuminated. As the day progresses, and the sun lowers, the possibility of shadows or the appearance of the sun in the view increases. To investigate the effects of lighting on accuracy, we visualize the average MCR with respect to the time of day in Figure 4.

When the sun is at its highest point in the sky, around noon, all labelers are at peak performance. A similar result is obtained when visualizing average MCR with respect to

Table 1. Average type I & II errors.

Method	Avg. type I error	Avg. type II error
Lu et al.	0.87%	26.81%
Hoiem et al.	0.91%	19.7%
Tighe et al.	14.68%	1.73%

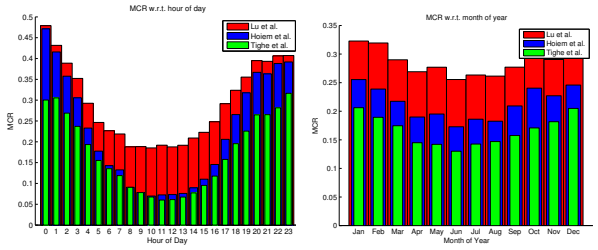

Figure 4. Time of day has a significant impact on labeler performance. All methods achieve their best performance when the sun is at its highest point, around noon. Less dramatic, month of year also has an impact on performance, with the best results seen during the spring and summer months.

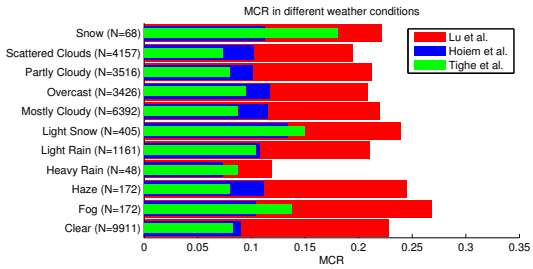

Figure 5. Average MCR for each method given a subset of weather conditions. To reduce the effect of lighting, we only evaluated images taken between 8:00 a.m. and 6:00 p.m. with respect to the local time of each image.

month of year. We believe the combined higher rates of failure in the winter months of the year are likely due to shorter days as a result of the Earth's tilt.

4.3. Impact of Weather on Accuracy

We now explore the effects of known weather conditions on MCR for the three methods. In Figure 5 we plot the MCRs of each method given a subset of weather conditions. We highlight four weather indicators (fog, heavy rain, light snow and snow) where the most accurate labeler, Tighe et al., is outperformed by Hoiem et al. We believe this effect is attributed to labeler confidence, i.e., low type I error methods have an advantage for images with general occlusion. Contrary to our expectations, two of the sky labeling approaches (Tighe et al. and Hoiem et al.) are mostly robust to cloud coverage.

4.4. Impact of Other Attributes on Accuracy

We select three transient attributes that are related to a scene appearance as a function of time of day: bright, night, midday, and three weather-related attributes: dry, winter, summer. We threshold the regressor responses (a real-valued score on the interval $[0, 1]$ indicating the presence of the attribute) to $> .6$ and plot the distribution of the resulting images in Figure 7. We observed that the transient features, when above threshold, are indeed related to time

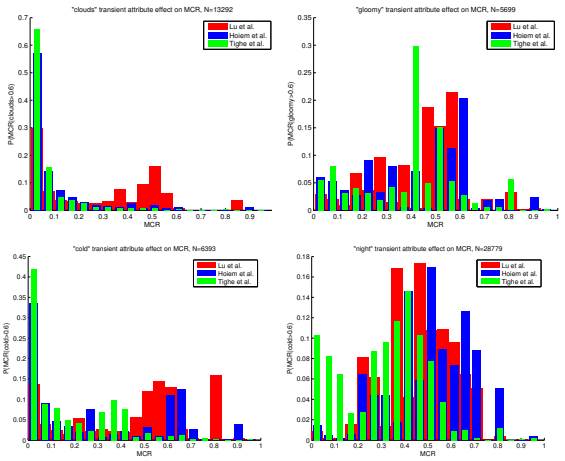

Figure 6. Probability distribution of MCR given transient features above a threshold. Top left: Tighe et al. and Hoiem et al. more robust to cloudiness than Lu et al. Top right: high failure rates for all three methods when "gloomy" is detected. Bottom left: all three methods are not robust to "cold" images. Bottom right: as expected, high failure rates occur for poorly lit images. The threshold for the transient attribute regressors was 0.6.

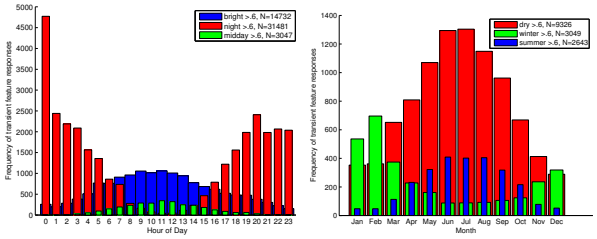

Figure 7. Left: transient attributes (bright, day, midday) are related to ground truth hour of day. Right: transient attributes (dry, winter, summer) are related to ground truth month of year.

of day and day of year. This provides strong support for using such features as a surrogate for weather data.

We now explore labeler robustness with respect to transient attributes as predictor of high labeler failure rates. In Figure 6 we show that Hoiem et al. and Tighe et al. are more robust to cloudiness conditions than Lu et al. The best predictors of high labeler failure are the "gloomy", "night" and "cold" transient attributes. The methods we evaluated seem to be robust to cloudiness. In Figure 8 we show sample images with high "cold" and "gloomy" attributes.

5. Deep Ensemble Approach

Based on our experiments and the insights we gained from the data analysis, we show improvements can be made by combining the outputs of the three methods we evaluated. While Tighe et al., achieved best overall performance, on a per-image count, they are outperformed on nearly half of the images in the dataset (Figure 3). This suggests that an ensemble method which uses the three methods' output,

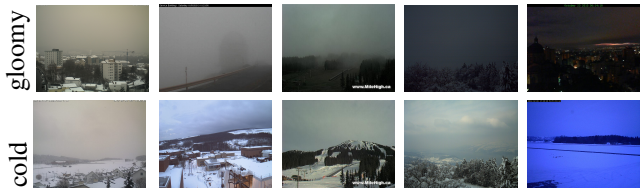

Figure 8. Sample images with "gloomy" and "cold" transient attributes above 0.6. These images are difficult due to sky color and clouds/haze/fog.

combined with raw image data, could outperform individual methods on our challenging dataset. We now describe an ensemble method based on a deep recurrent convolutional neural network (rCNN) architecture.

5.1. Recurrent CNN

We use a recurrent convolutional neural network, similar to that of Pinheiro et al. [13]. The network consists of three convolutional layers with hyperbolic tangent activation functions. The recurrent architecture involves the composition of three instances of a convolutional neural network, with each instance sharing identical parameters. Input to the rCNN is a 3D matrix with the smallest dimension indexing over the color channels. The full model file, solver definition and learned weights of our networks are available at [1].

We trained two rCNNs, one only with raw image data and another with the raw image data augmented by the outputs of the methods we evaluate. For the second network, we augmented the RGB input with the binary output of the three methods we evaluate, which results in a width $\times$ height $\times$ 6 input matrix. The output of the networks are probability maps of the same size as the input image. We threshold the output to obtain a binary label. Our data was split into a training set of 40 cameras and 13 test cameras. We trained each network on an NVIDIA Tesla M2075 for two days.

5.2. Evaluation

The results show rCNNs can be successfully used as ensemble methods to learn a nonlinear combination of raw image data and the outputs of other methods to improve accuracy. Overall, our ensemble outperforms the three methods evaluated. On the test cameras, Tighe et al. averaged an MCR of 18.19%, Hoiem et al. averaged 23.08%, and Lu et al. averaged 30.60%, while our rCNN ensemble averaged an MCR of 12.96%, a relative improvement of 28.75% from Tighe et al. The rCNN trained on raw image data alone achieved an average MCR of 17.28%. We compare this baseline approach to the ensemble in Figure 10.

To gain a better understanding of where improvements are most significant, we aggregate with respect to month of year and hour of day, as seen in Figure 9. We note

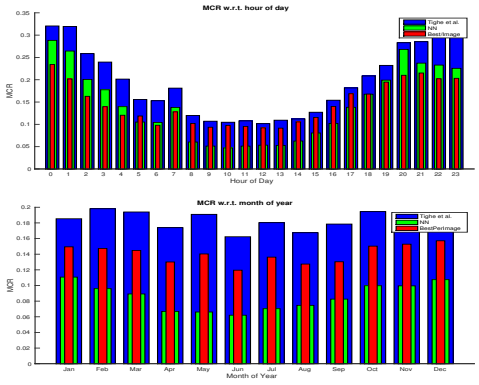

Figure 9. Comparison of our recurrent neural network ensemble with three compositions. The comparisons are made w.r.t. Tighe et al. and the best MCR per-image. rCNN outperforms the best method per-image metric when aggregated by month. Aggregation by hour of day reveals highest performance gains during daylight hours.

significant improvements compared to the best per image MCR[1] during daylight hours. Aggregation with respect to month of year shows improvements for all months, with higher performance gains during summer months, possibly attributable to longer days and "easier" weather conditions.

6. Conclusion

To analyze sky segmentation performance in real outdoor scenes, we created a new challenging dataset from static outdoor webcams over long periods of time with known ground truth location to supplement local weather observations. We extensively evaluated the performance of three sky labeling methods (Tighe et al. [17], Hoiem et al. [3] and Lu et al. [12]) under real-world weather and illumination conditions. This exploratory study was driven by the importance of accurately segmenting sky pixels from outdoor imagery, as it serves as input to many high-level vision algorithms. Our results show that sky labeling algorithm performance varies most significantly with respect to illumination conditions, i.e., sun position as indicated by time of day. In addition, we found that certain weather conditions and time of day are good predictors of current labeler errors.

We proposed a deep ensemble method that combines the output of existing methods with raw image data using an rCNN. Our model achieves better overall performance than any of the individual methods. Additionally, we aggregated results with respect to hour of day (the most important factor driving sky labeler performance) and compared our ensemble with a metric that uses an oracle to select the best method, showing improved performance during daylight hours. This work suggests two directions for future

[1]The best per-image metric gives us an empirical performance boundary treating all three input methods equally.

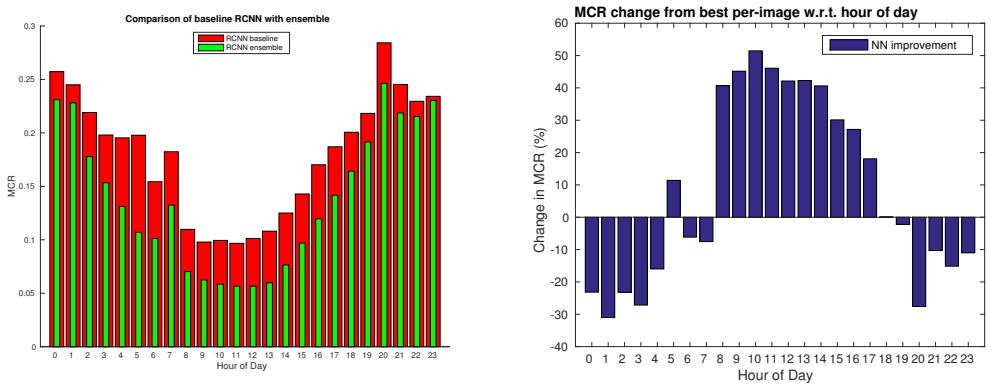

Figure 10. Left: We compare the baseline recurrent neural network with the deep ensemble. We note improvements for all hours of the day. Right: We show relative changes in the deep ensemble MCR when aggregated by hour of day and compared to the best MCR per image. Positive change represents a decrease in MCR, while negative change reflects an increase in error.

work. One is in exploring alternative methods for integrating weather metadata into the sky segmentation algorithms. In particular, we think that adding the metadata earlier in the processing pipeline would be beneficial. The other is in exploring how incorporating temporal context and weather metadata can improve performance on other vision tasks.

References

[1] http://mypages.valdosta.edu/rpmihail/skyfinder/. 2, 5

[2] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. *ACM Transactions on Graphics (SIGGRAPH)*, 24(3):577–584, 2005. 2

[3] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *IEEE International Conference on Computer Vision*, 2005. 1, 2, 5

[4] N. Jacobs, B. Bies, and R. Pless. Using Cloud Shadows to Infer Scene Structure and Camera Calibration. In *Computer Vision and Pattern Recognition*, 2010. 2

[5] N. Jacobs, J. King, D. Bowers, and R. Souvenir. Estimating Cloud Maps from Outdoor Image Sequences. In *IEEE Winter Conference on Applications of Computer Vision*, 2014. 2

[6] N. Jacobs, N. Roman, and R. Pless. Consistent Temporal Variations in Many Outdoor Scenes. In *Computer Vision and Pattern Recognition*, 2007. 2

[7] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating static cameras. In *IEEE International Conference on Computer Vision*, 2007. 2

[8] N. Jacobs, S. Workman, and R. Souvenir. Scene Geometry from Several Partly Cloudy Days. In *International Conference on Distributed Smart Cameras*, 2013. 2

[9] H. Katsura, J. Miura, M. Hild, and Y. Shirai. A view-based outdoor navigation using object recognition robust to changes of weather and seasons. In *IEEE International Conference on Intelligent Robots and Systems*, 2003. 2

[10] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (SIGGRAPH)*, 33(4), 2014. 2

[11] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *Computer Vision and Pattern Recognition*, 2009. 2

[12] C. Lu, D. Lin, J. Jia, and C.-K. Tang. Two-class weather classification. In *Computer Vision and Pattern Recognition*, 2014. 1, 2, 5

[13] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *International Conference on Machine Learning*, 2014. 5

[14] M. Roser and F. Moosmann. Classification of weather situations on single color images. In *IEEE Intelligent Vehicles Symposium*, 2008. 2

[15] A. Stylianou, A. Abrams, and R. Pless. Characterizing feature matching performance over long time periods. In *IEEE Winter Conference on Applications of Computer Vision*, 2015. 2

[16] L. Tao, L. Yuan, and J. Sun. Skyfinder: attribute-based sky image search. *ACM Transactions on Graphics (SIGGRAPH)*, 28(3):68, 2009. 2

[17] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *European Conference on Computer Vision*, 2010. 1, 2, 5

[18] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *Computer Vision and Pattern Recognition*, 2013. 2

[19] J. Tighe, M. Niethammer, and S. Lazebnik. Scene parsing with object instances and occlusion ordering. In *Computer Vision and Pattern Recognition*, 2014. 2

[20] S. Workman, R. P. Mihail, and N. Jacobs. A pot of gold: Rainbows as a calibration cue. In *European Conference on Computer Vision*, 2014. 2

[21] S. Workman, R. Souvenir, and N. Jacobs. Scene Shape Estimation from Multiple Partly Cloud Days. *Computer Vision and Image Understanding*, 134:116–129, 2015. 2

[22] S. Workman, R. Souvenir, and N. Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision*, 2015. 2

[23] X. Yan, Y. Luo, and X. Zheng. Weather recognition based on images captured by vision system in vehicle. In *International Symposium on Neural Networks*, 2009. 2