Segmenting Sky Pixels in Images

Cecilia La Place
Arizona State University
cecilia.laplace@asu.edu

Aisha Urooj Khan
University of Central Florida
aishaurooj@gmail.com

Ali Borji
University of Central Florida
aborji@crcv.ucf.edu

Abstract

Outdoor scene parsing models are often trained on ideal datasets and produce quality results. However, this leads to a discrepancy when applied to the real world. The quality of scene parsing, particularly sky classification, decreases in night time images, images involving varying weather conditions, and scene changes due to seasonal weather. This project focuses on approaching these challenges by using a state-of-the-art model in conjunction with a non-ideal dataset. We focus specifically on sky segmentation, the action of determining sky and not-sky pixels, and improving upon an existing state-of-the-art model: RefineNet. As a result of our efforts, we have seen an improvement of 10-15% from the average MCR compared to the prior methods. We have also improved from an off-the shelf-model in terms of average mIOU by nearly 35%.

1. Introduction

Inspired by Mihail *et al*. [7] paper that compared existing sky segmentation methods, and sought to bring attention to the problem of ideal datasets, we focused upon approaching the challenges they mentioned through existing models. The challenges of outdoor scene parsing lie in the time of day, year, and the varying weather conditions. Their Skyfinder dataset allowed us to pursue these challenges in order to obtain improved results.

Sky segmentation is a part of the scene parsing world, in which algorithms seek to label or identify objects in images. However, due to being trained on ideal datasets, these algorithms face difficulty in non-ideal conditions [7]. As deep learning methods might become more involved in real world applications, it becomes apparent that off-the-shelf methods are not always effective and reliability comes into question. This work highlights the importance of this problem. We offer an improved model that will aid in challenging sky segmentation in the real world.

2. Related Work

Current applications of sky segmentation range from personal to public use and more. Often times they are used in scene parsing [2, 10]. Other applications include weather classification [6], image editing [3, 9], weather estimation [11, 8], and more.

The history of scene parsing has evolved from its beginnings as convolutional networks. From convolutional networks spawned fully convolutional networks which moved away from pixel level algorithms and into whole-image-at-a-time methods [5]. Afterward, the introduction of skip layers led to deep residual learning [1]. Residual nets became the backbone to models such as Refinenet [4], RefineNext, and more. Other scene parsing algorithms such as PSPNet [12] have also come into the field returning to the idea of a fully convolutional network.

3. Approach

3.1. Dataset

The Skyfinder dataset is a subset of the Archive of Many Outdoor Scenes (AMOS) dataset. Due to the full Skyfinder dataset not being available, we used 45 of the 53 cameras shared. This entailed 60-70,000 images, with an average of 1500 images per camera. These images were of varying dimensions, quality, time of day, season, and weather. However, some cameras contained images which indicated the camera was experiencing technical difficulties or repairs. These few images were removed to focus on the challenges we wished to address. We then changed the sizes of the images to introduce a form of uniformity and make test evaluation faster by resizing them to be within the ranges of 320x480.

A singular segmentation map was associated with each camera. This is due to each camera being stationary for at least a year in order to be included within in the dataset. We then duplicated the ground truth for each original image, and changed the names of the copies to match the file name of each original image in order to use RefineNet.

Method	S1-MCR	S1-mIOU	S2-MCR	S2-mIOU	ave-MCR	ave-mIOU
Hoiem *et al*.	21.28%	-	20.68%	-	20.98%	-
Lu *et al*.	25.38%	-	21.67%	-	23.52%	-
Tighe *et al*.	17.48%	-	20.33%	-	18.91%	-
(RefineNet-Res101) Cityscapes	17.12%	49.31%	18%	46.31%	17.56%	47.81%
(RefineNet-Res101) Finetuned	5.17%	79.48%	7.48%	72.18%	6.33%	75.83%
(RefineNet-Res101) ImageNet	**5.08%**	**87.07%**	**7%**	**73.84%**	**6.04%**	**80.46%**

Table 1. Results from the testing splits

3.2. Refinenet

The model we used, Refinenet, was created by Lin *et al*. [4]. This model seeks to to retain detail throughout the reconstruction of the image and its segmented output unlike its predecessor and backbone, ResNet.

3.3. Trained on Cityscapes

In order to establish a baseline in Refinenet, we used Refinenet's Res101 model that was trained on Cityscapes, a dataset of European cities for scene segmentation. The Cityscapes dataset is of course ideal, and as a result does well in ideal conditions in the Skyfinder dataset. After setting up the model and using it on a singular Titan X GPU, we ran each of the 45 cameras through the model and evaluated solely on their sky classification ability.

3.4. Finetuning

To obtain proof of concept prior to running the entire dataset we focused on a smaller subset of the dataset. We took 10 random cameras and broke it into a train-val-test split. From each camera we took 75 files for training, 25 files for validation, and between 175-300 images for testing and evaluation.

Following the success of running the model on the subset, we finetuned on the model trained on Cityscapes by initializing from it and using Refinenet's code along with a different train-val-test split than our proof of concept version and the one incompletely referred to in Mihail *et al*. [7] as it was not clarified which were used. The split consisted of 13 cameras used for testing, 4 cameras used for validation, and the remaining cameras used for training. We then shuffled the cameras in each section but kept the same number of cameras for each training, validation, and testing set, and repeated the aforementioned finetuning process. We used a learning rate of 5e-5 for each instance and ran the testing section of the dataset on the model for 10 epochs.

3.5. Training with Skyfinder

Finally, we used ImageNet's pre-trained resnet with 101 layers to initialize from, but trained the model on solely the Skyfinder dataset. We used the same exact same two train-val-test splits as mentioned above (to allow for proper comparison) and trained at a learning rate of 5e-4 for over 10 epochs. Due to the flattening of the learning curve, we used the model at epoch 10 for testing in both instances.

4. Evaluation

To evaluate the accuracy of our results we used both the misclassification rate (MCR) defined in Mihail *et al*. [7], and the mean intersection over union (mIOU). The use of both of these allowed for the ability to compare our results to the results found in Mihail *et al*. [7], which focused primarily on MCR results, and determine the overlapping accuracy of the segmentation outputs.

$$MCR = \frac{FP + FN}{imageresolution} \tag{1}$$

$$mIOU = \sum \frac{TP}{TP + FP + FN} \tag{2}$$

5. Results

For reference, in saying Cityscapes we mean the model trained on Cityscapes.

The results shown in Table 1 show the results between the cameras used in testing in first and second splits of data, the average between them, the average of the same cameras' results from Mihail *et al*. and their overall average. Hoiem *et al*., Lu *et al*., or Tighe *et al*. in Mihail *et al*.'s data do not report their mIOU.

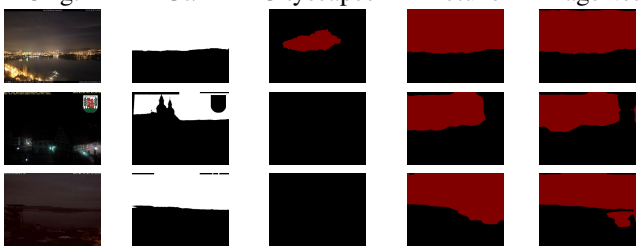


Table 2. Improvement in night time images

For visual examples, the figures are pulled from a few of the first test set of cameras and do not include visual results

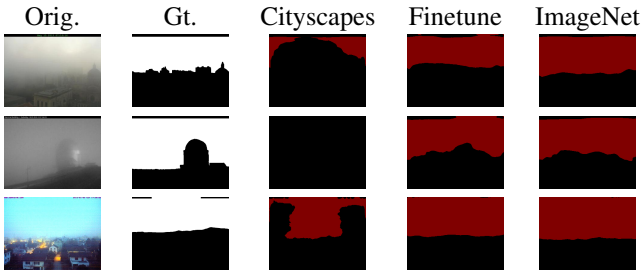


Table 3. Improvement in weather obscured images

from Hoiem *et al*., Lu *et al*., or Tighe *et al*. In Cityscapes, the images have been made binary to focus on the sky segmentation. Red is sky, black is everything else. Similarly in the other results the red marks the sky and black the ground. While Mihail *et al*. did create their own ensemble using the combination of the three methods, their results were unreported for individual images.

6. Discussion

We should first direct our focus to the original results of Refinenet's res101 model trained on Cityscapes in Table 1. Both results incorporate the pretrained model's incapability to properly access non-ideal images, particularly night images as addressed earlier. Upon finetuning on the model using the Skyfinder dataset, we note a drastic improvement in the mIOU and just over 10% decrease in the MCR. This is clearly as a result of including the non-ideal images. Upon training it on the dataset but initializing from ImageNet, we see a less drastic improvement in one split, but improvement nonetheless.

Despite the averages in Table 1, focusing on a singular camera we can see a vast improvement. Camera 858, consists of 200 images, and was not seen by the finetuned or ImageNet initialized model during training. The mIOU from the model trained on Cityscapes was 69.21%, and had an MCR of 18.16%. However, after finetuning the model, the mIOU and MCR results respectively quickly jumped up to 91.41% and 4.69%. Finally, after training on only 28 cameras of the data (again, not including camera 858) and initializing from ImageNet, the results improved slightly. The mIOU increased to 95.76% and the MCR dropped a little more to 2.30%. Other cameras show similar rates of improvement. There are some cameras however that prove difficult to segment for all methods and show a smaller rate of improvement.

The importance of these results lies in the ability to use these in real-world applications. Using the original pretrained model would result in poor quality segmentation outside of the ideal circumstances. Off-the-shelf methods must be modified in order to be used most effectively in the real world.

Understanding the impacts of these results, we can now incorporate the results from Mihail et al's findings. While their own model's results which reported an MCR of 12.96% across their own testing split, we cannot report on the MCR of images consisting of our testing split. In spite of this, we can still further prove the idea that existing models are still effective–so long as they have been modified to suit the task's needs.

7. Future Work

Following this work, we intend to look at other scene parsing models to evaluate their off-the-shelf methods, finetune on them, and possibly train it on the Skyfinder dataset in order to compare the results to the above. We also plan to develop our own end-to-end sky segmentation model to also use for comparison. To possibly improve general results, cleaning of the dataset may need to occur, such as removing timestamps.

Future plans also include developing applications for these findings, such as in weather classification and weather forecasting. We intend to also find use for the weather metadata associated with Skyfinder images alongside these segmentations.

8. Acknowledgements

We would like to thank the National Science Foundation for funding the Research Experience for Undergraduates program at the Center for Research in Computer Vision at UCF. We would also like to thank Dr. Shah, Dr. Lobo, and Dr. Borji, for their help and support throughout this program.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2016.

[2] D. Hoeiem, A. A. Efros, and M. Hebert. Geometric context from a single image, 2005.

[3] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes, 2014.

[4] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, 2016.

[5] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation, 2015.

[6] C. Lu, D. Lin, J. Jia, and C. K. Tang. Two-class weather classification, 2014.

[7] R. Mihail, S. Workman, Z. Bessinger, and N. Jacobs. Sky segmentation in the wild: An empirical study. *Journal of Foo*, 12(1):234–778, 2002.

[8] L. Tao, Y. Luo, and X. Zheng. Weather recognition based on images captured by vision system in vehicle, 2009.

[9] L. Tao, L. Yuan, and J. Sun. Skyfinder: Attribute-based sky image search, 2009.

[10] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels, 2010.

[11] Wei-TaChu, Xiang-YouZheng, and D.-S. Ding. Camera as weather sensor: Estimating weather information from single images. *Journal of Visual Communication and Image Representation*, 46(1):233–249, 2017.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network, 2017.