# Supplementary Material: Active Learning for Improved Semi-Supervised Semantic Segmentation in Satellite Images

Shasvat Desai *
Orbital Insight
shasvat.desai@orbitalinsight.com

Debasmita Ghose *
Yale University
debasmita.ghose@yale.edu

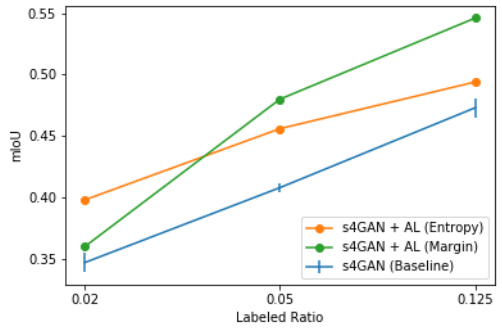
## 1. Ablation Study

### 1.1. Active Learning Parameters

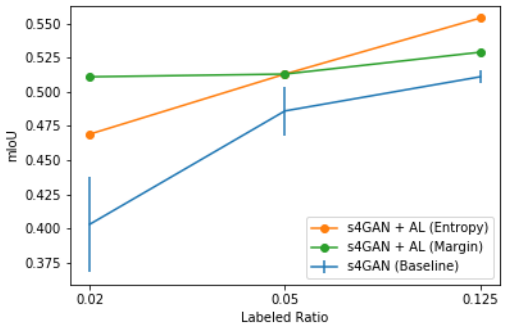
Tables 1 and 2 show the results of our experiments with different combinations of $\alpha_{init}$ and $\beta_Q$ on UC Merced Land Use Classification [9] and the DeepGlobe Land Cover Classification [1] datasets respectively. We vary both the parameters between 0.1 and 0.9 for both entropy and margin-based sampling strategies for three different labeled ratios. We found the best performing $\alpha_{init}$ and $_Q$ values to be 0.1 and 0.5 respectively. Overall we noticed out method to be sensitive to changes in $\alpha_{init}$ and $\beta_Q$ as the average difference in the worst performing and best performing model across all labeled ratios and sampling techniques is 4 mIoU points for the UC Merced Land Use Classification Dataset and 2.4 mIoU points for the DeepGlobe Land Cover Classification Dataset.

### 1.2. Network Capacity of Active Learner

Tables 3 and 4 show the results of our experiments with different backbone networks on UC Merced Land Use Classification [9] and the DeepGlobe Land Cover Classification [1] datasets respectively. We experiment with VGG-16 [6], ResNet-50 [2] and ResNet-101 [2] which have different network capacities. We found the best performing backbone network to be ResNet-101 for the UC Merced Land Use Classification dataset and ResNet-50 for the DeepGlobe Land Cover Classification dataset. As shown by the results, the image classification network's capacity for the *learner* is crucial in determining the quality of the selected samples. Any network with low capacity with respect to the size of the dataset and the number of classes tends to underfit, while any network with a higher capacity than required could overfit and detrimentally affect the downstream task's performance. We noticed our method to be sensitive to networks with different capacities as the average difference in the worst performing and best performing model across all labeled ratios and sampling techniques is 3.3 mIoU points for the UC Merced Land Use Classification Dataset and 2.9 mIoU points for the DeepGlobe Land Cover Classification Dataset. Notably, we see that in most cases, VGG-16 performed significantly performed poorly across all labeled ratios in both the datasets as compared to the ResNet-50 and ResNet-101 models reinforcing the hypothesis that models with insufficient network capacity underperform at the downstream task.




a)




b)

Figure 1: Visualization of quantitative results for different labeled ratios for the a) UC Merced Land Use Classification Dataset [9] and, b) DeepGlobe Land Cover Classification Dataset [1]

*Authors contributed equally

| Active Learning Parameters | | 2% | | 5% | | 12.5% | |
|---|---|---|---|---|---|---|---|
| $\alpha_{init}$ | $\beta_Q$ | **Entropy** | **Margin** | **Entropy** | **Margin** | **Entropy** | **Margin** |
| 0.1 | 0.1 | 0.381 | 0.358 | 0.423 | 0.450 | 0.484 | 0.497 |
| **0.1** | **0.5** | **0.398** | **0.36** | **0.456** | **0.48** | **0.494** | **0.546** |
| 0.9 | 0.9 | 0.352 | 0.353 | 0.411 | 0.421 | 0.478 | 0.478 |

Table 1: Ablation Study for the different Active Learning parameters on the UC Merced Land Use Classification Dataset [9]

| Active Learning Parameters | | 2% | | 5% | | 12.5% | |
|---|---|---|---|---|---|---|---|
| $\alpha_{init}$ | $\beta_Q$ | **Entropy** | **Margin** | **Entropy** | **Margin** | **Entropy** | **Margin** |
| 0.1 | 0.1 | 0.464 | 0.497 | 0.507 | 0.502 | 0.549 | 0.513 |
| **0.1** | **0.5** | **0.469** | **0.511** | **0.513** | **0.513** | **0.554** | **0.529** |
| 0.9 | 0.9 | 0.449 | 0.462 | 0.495 | 0.498 | 0.527 | 0.512 |

Table 2: Ablation Study for the different Active Learning parameters on the DeepGlobe Land Cover Classification Dataset [1]

## 2. Quantitative Evaluation of Diversity

In this paper, we proposed a method which aims to select the most diverse and representative set of samples to serve as an initial labeled set of data for the semi-supervised network. We empirically showed the success of the proposed method on different datasets. In this section, we evaluate the robustness of our method using statistical indices which measure the diversity of the selected samples. To achieve this, we choose two diversity indices which are frequently used in ecological studies that measure species diversity, but the same analysis can also be applied to measure diversity of any set of random samples.

### 2.1. Shannon's Diversity Index

The Shannon index [5] was developed from information theory and is based on measuring uncertainty. Shannon's index accounts for both abundance and evenness of the samples present. Shannon index is defined in Equation 1:

$$H(x) = -\sum_{i=1}^{N} p_i \log p_i \tag{1}$$

In our case, each sample is a pixel. Hence, $p_i$ indicates the probability that a given pixel belongs to class $i$. N indicates the total number of classes that a given pixel can belong to. Thus, we are measuring how diverse are the samples selected by the active learning method as compared to samples selected randomly. Therefore, samples with a large number of pixels from different classes that are evenly distributed are the most diverse. On the other hand, samples that are dominated by pixels from one class are the least diverse. We report the value of Shannon diversity index for our baseline method averaged across our three experiments with different random seeds and for samples selected by both the active learning techniques. Intuitively, Shannon's index quantifies the uncertainty in predicting the class to which a given pixel belongs and hence a higher value of Shannon diversity index indicates a more diverse set of samples.

Our results for Shannon's diversity index are shown in Tables 5 and 6 for the UC Merced Land Use Classification [9] and DeepGlobe Land Cover Classification [1] datasets respectively. We notice a strong correlation between the mIoU values reported in the paper for the baseline and active learning strategies and the values of the Shannon's diversity index obtained for the respective experiments.

### 2.2. Simpson's Diversity Index

Traditionally, Simpson's Diversity Index [7] measures the probability that two individuals randomly selected from a sample will belong to the same species (or some category other than species). We extend it to our use case to measure the diversity of the selected samples. To make it easier and intuitive to understand the relevance of this index, we use the inverse Simpson index. Thus, greater the value, the greater the sample diversity. In this case, the index represents the probability that two individuals randomly selected from a sample will belong to different species. Thus, the inverse Simpson index is defined in 2:

$$D = 1 - \frac{\sum(n(n-1))}{N(N-1)} \tag{2}$$

where,
$n$ = the number of pixels belonging to class $i$,
$N$ = total number of classes that exist in the dataset.

Similar to Shannon's index in Section 2.1, we report results on the UC Merced Land Use Classification [9] and the DeepGlobe Land Cover Classification [1] datasets in Tables 7 and 8 We show that both the active learning sampling strategies used in this paper yield more diverse set of

| | 2% | | 5% | | 12.5% | |
|---|---|---|---|---|---|---|
| **Backbone** | **Entropy** | **Margin** | **Entropy** | **Margin** | **Entropy** | **Margin** |
| VGG-16 | 0.355 | 0.351 | 0.421 | 0.426 | 0.481 | 0.498 |
| Resnet-50 | 0.371 | 0.354 | 0.434 | 0.452 | 0.489 | 0.524 |
| **Resnet-101** | **0.398** | **0.36** | **0.456** | **0.48** | **0.494** | **0.546** |

Table 3: Impact of different network architectures for the active learner in UC Merced Land Use Classification Dataset [9] on mIoU values

| | 2% | | 5% | | 12.5% | |
|---|---|---|---|---|---|---|
| **Backbone** | **Entropy** | **Margin** | **Entropy** | **Margin** | **Entropy** | **Margin** |
| VGG-16 | 0.421 | 0.445 | 0.492 | 0.499 | 0.523 | 0.52 |
| **Resnet-50** | **0.469** | **0.511** | **0.513** | **0.513** | **0.554** | **0.529** |
| Resnet-101 | 0.443 | 0.482 | 0.505 | 0.492 | 0.534 | 0.518 |

Table 4: Impact of different network architectures for the active learner in the DeepGlobe Land Cover Classification Dataset [1] on mIoU values

| **Labeled Ratio(R)** | **2%** | **5%** | **12.5%** |
|---|---|---|---|
| s4GAN [4] (Baseline) | 1.96 ± 0.08 | 2.16 ± 0.02 | 2.14 ± 0.03 |
| s4GAN + Entropy (Ours) | 2.10 | 2.20 | 2.22 |
| s4GAN + Margin (Ours) | 2.08 | 2.22 | 2.25 |

Table 5: Shannon's Diversity Index for the UC Merced Land Use Classification Dataset [9] (Higher the better)

| **Labeled Ratio(R)** | **2%** | **5%** | **12.5%** |
|---|---|---|---|
| s4GAN [4] (Baseline) | 1.01 ± 0.04 | 1.16 ± 0.05 | 1.19 ± 0.14 |
| s4GAN + Entropy (Ours) | 1.06 | 1.25 | 1.38 |
| s4GAN + Margin (Ours) | 1.09 | 1.24 | 1.36 |

Table 6: Shannon's Diversity Index for the DeepGlobe Land Cover Classification Dataset [1] (Higher the better)

| **Labeled Ratio(R)** | **2%** | **5%** | **12.5%** |
|---|---|---|---|
| s4GAN [4] (Baseline) | 0.79 ± 0.03 | 0.83 ± 0.009 | 0.83 ± 0.008 |
| s4GAN + Entropy (Ours) | 0.85 | 0.84 | 0.85 |
| s4GAN + Margin (Ours) | 0.82 | 0.86 | 0.87 |

Table 7: Simpson's Diversity Index for the UC Merced Land Use Classification Dataset [9] (Higher the better)

| **Labeled Ratio(R)** | **2%** | **5%** | **12.5%** |
|---|---|---|---|
| s4GAN [4] (Baseline) | 0.55 ± 0.04 | 0.64 ± 0.01 | 0.65 ± 0.02 |
| s4GAN + Entropy (Ours) | 0.58 | 0.73 | 0.71 |
| s4GAN + Margin (Ours) | 0.62 | 0.71 | 0.68 |

Table 8: Simpson's Diversity Index for the DeepGlobe Land Cover Classification Dataset [1] (Higher the better)

samples and show strong correlation with the mIoU values reported on these datasets in the paper.

## 3. Discussion

### 3.1. Applicability of Our Method to Land Use Classification

The average number of semantic categories per scene in the UC Merced and DeepGlobe Landuse Classification datasets used in this paper is 3.39 and 2.51 respectively as depicted by figure 2. This implies that a given scene from the UCM dataset with a given image-level label will have 3 or more different pixel-level labels(semantic categories). Similarly, for the DeepGlobe dataset, we have about 2 or more semantic categories per scene on an average. UCM dataset has a total of 18 semantic categories and DeepGlobe has 6 semantic categories. Thus, each satellite scene in the UCM dataset has about 18% of all pixel level labels and similarly each satellite scene in the DeepGlobe dataset has about 42% of all pixel-level labels on an average. Figure 2 also shows us that about 90% of scenes in the UCM dataset have more than 1 semantic category and similarly about 80% of scenes in the DeepGlobe dataset have more than 1 semantic category. This number if quite high when we

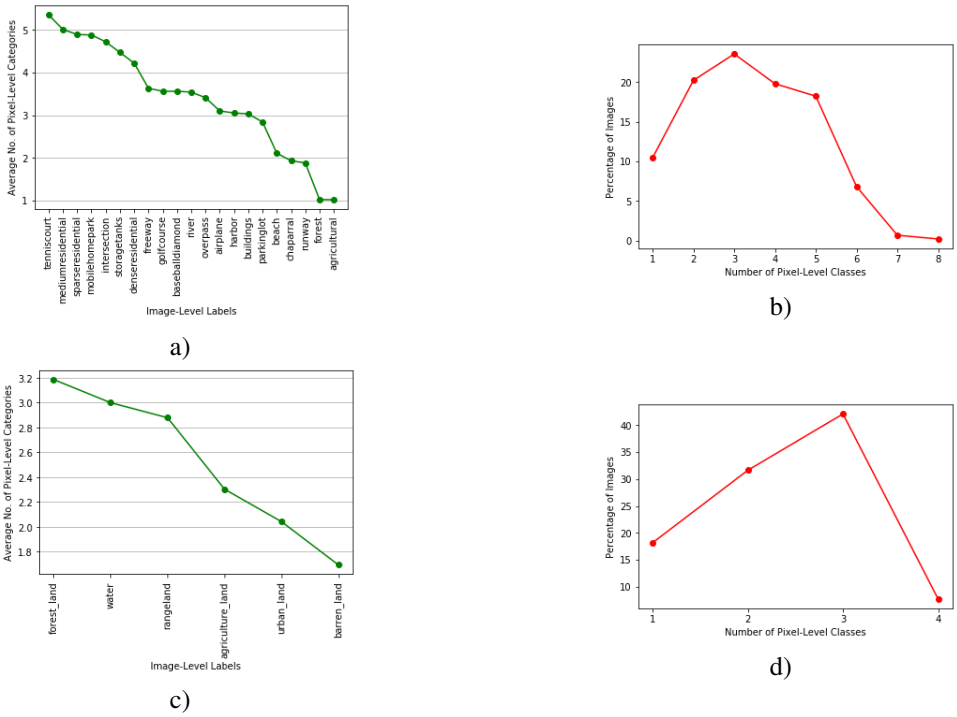



Figure 2: (a) Average number of semantic categories per image-level category for UC Merced Land Use Classification Dataset. (b) Percentage of images containing vs. number of pixel-level categories per image for UC Merced Land Use Classification Dataset. (c) Average number of semantic categories per image-level category for DeepGlobe Land Cover Classification Dataset. (d) Percentage of images containing vs. number of pixel-level categories per image for DeepGlobe Land Cover Classification Dataset

compare this statistics with that in some generic standard dataset. For instance, consider the COCO dataset [3]. Less that 30% of the images in the COCO dataset have more than 1 semantic category. This tells us that the landuse scenes in the domain of satellite imagery are inherently more diverse and hence our method is highly applicable specifically for land use classification in satellite images. We will get a more diverse set of samples for satellite domain as compared to using our method on generic datasets like COCO.

### 3.2. Suitability of s4GAN as our baseline

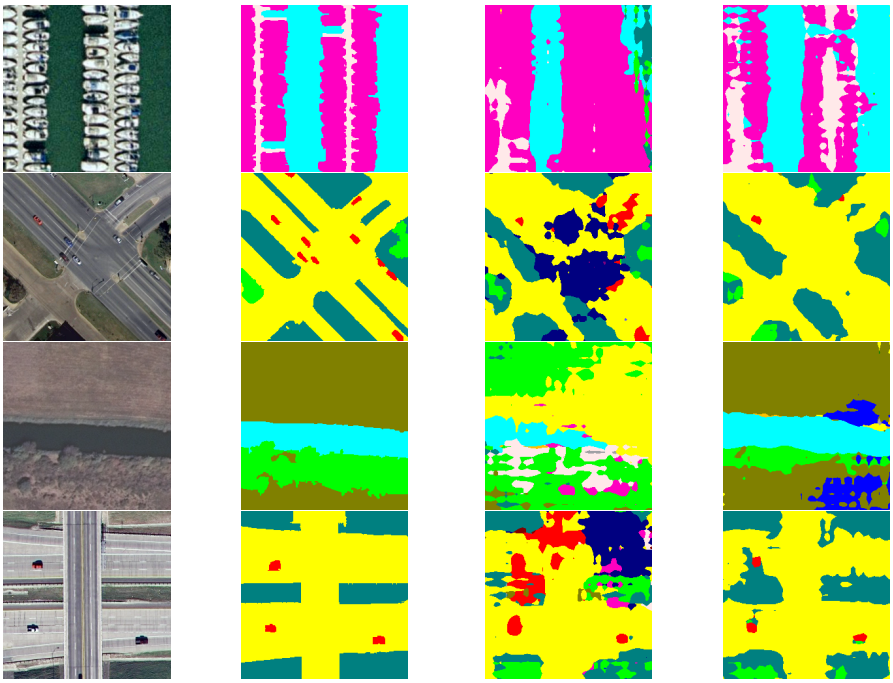
[4] propose to fuse the output of the s4GAN network with another image classification-based network called MLMT [8] during inference to reduce false positives. This MLMT branch uses an image classification network to output a confidence score for every category in the dataset. This output is combined with the pixel level output of the s4GAN network to reduce the number of false positives in the segmentation network. Therefore, one major constraint for using MLMT is that there should be a one-to-one correspondence between the image-level and the pixel-level labels. This would mean that the number of image-level categories should equal the number of pixel-level categories in a dataset. However, this does not always hold in the case of land use classification. An image-level label for land use classification in a satellite scene indicates predominant usage of land. However, the same scene can have multiple semantic categories. This prevents us from using MLMT as done by [4] as our baseline for the task of land use classification.

## 4. More Qualitative Evaluation

In this section, we provide more qualitative results from our best performing active learning strategies and compare them to our baseline for the UC Merced Land Use Classification Dataset [9].

Figure 3 compares the performance of our method with the baseline when trained with 2% labeled data. Row 1 shows how our method predicts the row of boats parked on the harbor better than the baseline method. Rows 2, 3, and 4 show that the baseline method gets confused between multiple unrelated classes, whereas our method reasonably predicts the correct classes.

Similarly, Figure 4 qualitatively compares the performance of our method with the baseline when trained with 5% labeled data. Rows 1 and 4 show an example of our



a) **Original Image** b) **Ground Truth** c) **Baseline** d) **Our Results**

Figure 3: Qualitative Results from the UC Merced Land Use Classification Dataset for 2% labeled data

method predicting the complex shape of airplanes better than the baseline method. Row 2 shows the baseline method being confused between cars in a parking lot and boats parked along a harbor, whereas our method predicts cars parked close together correctly. Row 3 shows how the baseline method completely misses the river and gets confused between multiple classes, while our method predicts the river reasonably well.

Finally, Figure 5 shows some qualitative examples of how our method outperforms the baseline when trained with 12.5% labeled data. Row 1 shows the baseline being confused between buildings and mobile homes, while our method predicts buildings in a dense residential setting more accurately. Rows 2 and 4 show our method predicting the baseball diamond structures accurately without being confused between other classes. Similarly, as shown by Row 3, our method predicts the contours of the airplane better than the baseline.

## References


[1] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018. 1, 2, 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 1

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[4] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3, 4

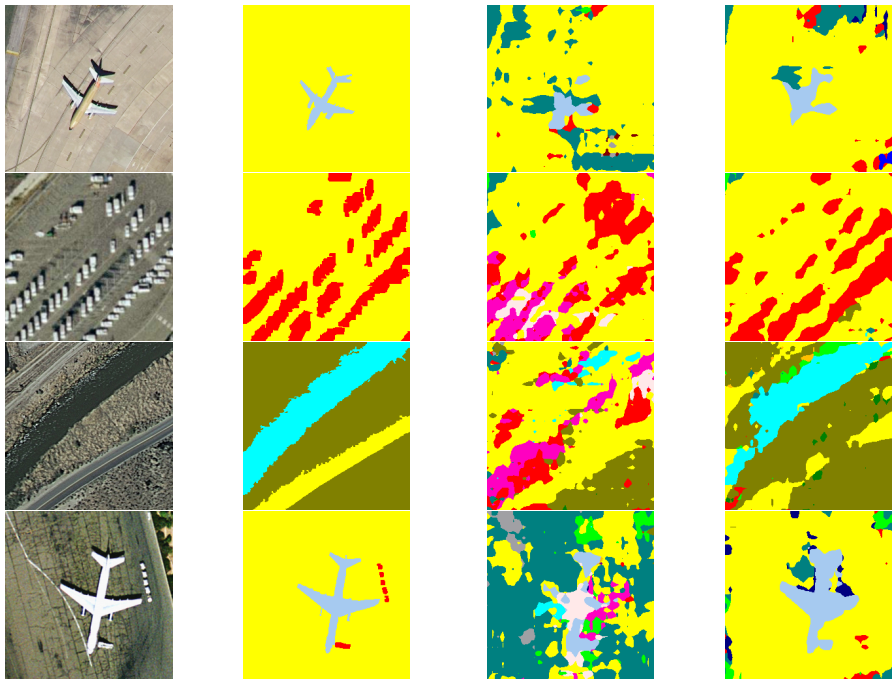
[5] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 2

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 1

[7] E Simpson. Medición de la diversidad. *Nature*, 163(688):1, 1949. 2

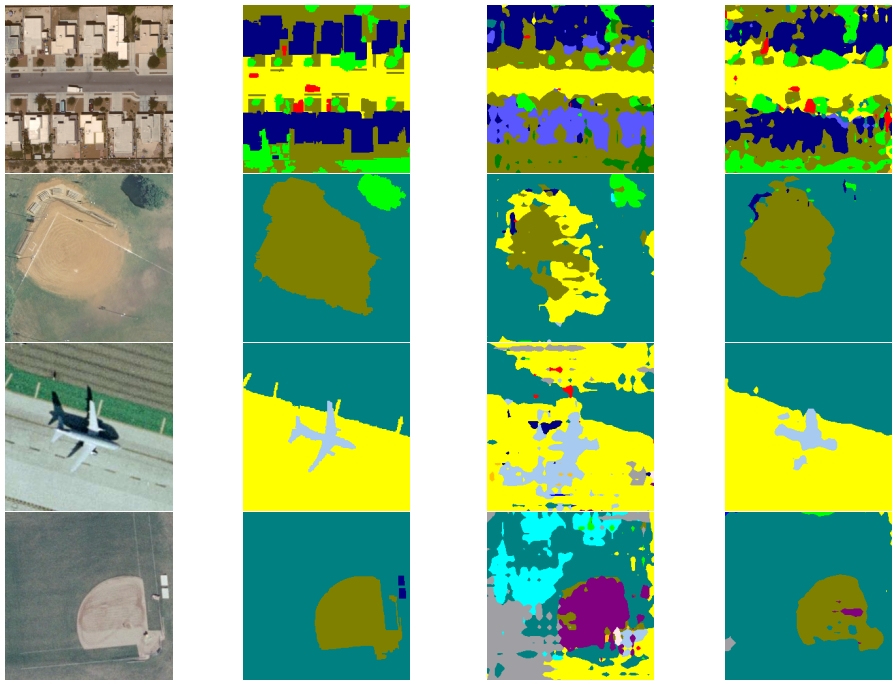
[8] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017. 4

[9] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 1, 2, 3, 4



a) **Original Image** b) **Ground Truth** c) **Baseline** d) **Our Results**

Figure 4: Qualitative Results from the UC Merced Land Use Classification Dataset for 5% labeled data



a) **Original Image** b) **Ground Truth** c) **Baseline** d) **Our Results**

Figure 5: Qualitative Results from the UC Merced Land Use Classification Dataset for 12.5% labeled data