Windows On Earth - TERC 2

Aditi Dass, Greg Frasco, Frederick Joossens
{aditid, frascog, fjansen}@bu.edu

1. Project Task

In this project classified images that are taken from the International Space Station are to be categorized into galleries on http://www.windowsonearth.org/. The set of tags comes from the Windows on Earth organization. One of the difficulties of this project is that only a small subset of the images is labeled, while the total size of the dataset is over 1.3 million images. See Figure 1 for example image and tags.

Figure 1: Example Instance
Tags: Dragon, Dock Undock, Structure

2. Related Work

Much work has already been completed in the space of computer vision and image recognition, allowing us to draw from a rich body of prior research. Furthermore, a vast set of resources is available on the Internet, ranging from blog posts, tutorials, … etcetera, to freely available books and online lectures.

For this project specifically, we rely upon the work of Inception-v3 (Szegedy et al. 2016), a state of the art deep convolutional neural network that is validated against the ImageNet Large Visual Recognition Challenge data set. This is paired with transfer learning, a technique where a small amount of newly labeled data is applied to a much larger, previously trained model (Dai et al. 2007). Finally, we look at ways to modify the more common single-label classification techniques and move into the realm of multi-label image classification (Boutell et al. 2004).

3. Approach

Our approach is to create several different machine learning models for each image classification tag and identify which ones fit best. To gain more insight into the data, we subject the given samples to a number of general classifiers. We also run experiments using transfer learning to leverage a vast amount of prior image classification efforts. Lastly, we experiment with various similarity and entropy indexes to find a suitable technique to identify "movie series".

3.1 Supervised Learning Classifier Definitions

Each supervised learning classifier is trained on the provided set of 10000 photos, and completes a five cross-fold validation on a set of classifiers for each tag to get the mean accuracy and standard deviation. The following classifiers are used:

- K Neighbors Classifier (kNN) with k=3 and uniform distance weighting
- Decision Tree with Gini impurity and entropy for the information gain
- Random Forest with five Decision Trees
- Feed-Forward Neural Network with one hundred neurons, ReLU activation function, and consistent learning rate of 0.0001 (Hastie et al. 2009)
- AdaBoost Classifier using the Decision Tree Classifier with fifty estimators and a learning rate of 1 (Glorot and Bengio 2010)
- Gaussian Naive Bayes
- SVM using radial basis function kernel (Gaussian)

3.2 Movie Series Research

One tag that does not respond very well to the other techniques mentioned in this paper is “movie”. A movie refers to a collection of time-lapsed images, where each participating image is tagged as such. As the concept of a time-lapse cannot not be characterized by the image features of a small sample, supervised learning is not suitable here.

Additionally:

- Times-lapses have to be explicitly set which causes the exposure program number to be ‘1’, indicating manual settings. This causes specific exposure indicators that the model would use to identify a “movie” image. However, many other photos are taken with manual settings resulting in false positives.

- Time lapses also take up 75% of the data but less than one percent are labeled, resulting in false negatives.

As a result, we instead test the probable success rates of various similarity metrics such as measuring entropy between two images (Grandvalet and Bengio 2005). We assume that in a time-lapse, an image and its successor would be more similar than two random images.

Therefore, we use the below-mentioned similarity metrics on both the tagged images and a sample of random images. By comparing these sets with a two sample t-test, we are able to determine if the similarity metric is statistically significant enough to provide a proper threshold by which a movie series can be accurately identified.

We utilise the following comparison methods:

- Mean Square Error
- Naive / Maximum Likelihood Entropy Estimate
- Panzeri-Treves / Bayesian Bias Correction (Panzeri and Treves 1996)
- Quadratic Extrapolation (Strong et al. 1998)
- Image-Histogram Entropy Comparison

3.3 Transfer Learning

For transfer learning we use the pre-trained Inception-v3 model and retrain the final fully connected layer. We attempt three different approaches in order to support single label classification as well as multi-label classification.
For the first goal we utilize softmax for the activation function and a softmax cross entropy loss. This matches the approach that Inception-v3 takes to train the model.

In order to support multiple labels, we change the last layer to a sigmoid nonlinearity with a sigmoid cross entropy loss. This matches a technique employed by Google for the OpenImages dataset (Krasin et al. 2017).

Finally, we attempt to obtain a multi-label classification result by modifying the single-label classification technique. In order to accomplish this we train a separate model for each individual label. This allows us to use a softmax activation function instead of sigmoid, which makes fine-tuning the dataset per label easier. Given the limited amount of labels that have to be supported (in the order of 10-20), the overhead is acceptable.
We use ~10000 images and train for 10000 epochs, with a learning rate of 0.01. The distribution is 80% training, 10% validation and 10% testing. A batch size of 100 images is used for training and validation, with a batch size of 1 for testing.

4. Dataset and Metric

Dataset	Images	Size
BUSampleDataSet	1,990	7.25 GB
WinEarthPhotosByKeyword	975	3.2 GB
BU10000Set	10207	23.69 GB

Table 1: Dataset descriptions

The WinEarthSampleSetTags and BU10000Set CSV files are provided with tags for images in BUSampleDataSet and BU10000Set respectively.

To determine our accuracy we use cross validation, which splits the data sets into k parts equally, and trains on k-1 parts and tests on 1 part. It repeats this k times until the full dataset has been tested. This helps lower bias in our testing data.

The following requirements were provided:

Tags	Useful Accuracy	Good Accuracy
Day, Night, Movie	80%	95%
Aurora, Cupola, DockUndock, Stars, SunriseSunset, Structure, Volcano, Agriculture, Clouds, Hurricane	70%	80%

Table 2: Baseline required accuracy provided by TERC.

For all images, we parse the labels (either through IPTC data or a separate CSV file) and resize them to be maximum 500px in either dimension. This is done as a separate step which only needs to happen once. For the supervised learning classifiers, we also convert the images to grayscale. See Appendix 10.3 for examples.

5. Evaluation

With ~10,000 images and ~33% of them labeled as "day", the "dumb" baseline for single label classification would be 33% accuracy if we always guess an image has that label. Alternatively, never guessing volcano would result in ~99.4% accuracy given the very limited occurrence of that tag. This is not a very useful metric to compare our solution to however. Though there is a bias towards certain tags, in our evaluation we always use an equal distribution. This makes 50% our starting baseline for test accuracy.

5.1 Supervised Learning

Each model is evaluated using a five cross-fold validation with the large multi-labeled dataset. Doing this cross validation for results returns accuracy, precision, loss, and training time. Since the classification methods surpass the aforementioned baselines, it serves as another baseline for our transfer learning method as the general classifiers represent the minimum machine learning can achieve. See appendix 10.2 for the accuracy table where each accuracy is the mean of the five cross-fold accuracy.

	Aurora	Cupola	Volcano
Best Accuracy	99.5%	99.3%	99.2%
Model	AdaBoost	kNN	Neural Net

Table 3: Testing accuracy of various testing methods. See Appendix 10.2 for the full test results.

All classifiers are set up with many standard parameters such as kNN where k=3 with uniform distance weights. The classifier parameters are not tuned to increase accuracy or precision to compare our results to transfer learning. We will only tune the parameters of the best default approach. This model succeeds very well except in movie classification, since it detects features within the image, where movies do not have distinct features.

5.1 Movie Series Identification

Method	p-value
Mean Square Error	1.2349×10^{-29}
Naive Entropy	5.9614×10^{-57}
PT Entropy	5.0377×10^{-68}
QE Entropy	2.0626×10^{-66}
Image-Histogram Entropy	$\mathbf{6.425 \times 10^{-76}}$

Table 4: P-values of 2 sample T test for various similarity metrics

Values for successive images provided above have extremely small p-values which means we are able to **reject** the premise that the difference between the similarity in random images and similarity in time-lapsed images was random with a very high confidence interval.

As the Image-Histogram Entropy measure has the smallest value (by a factor of 10^{10} compared to the second smallest), we recommend its usage to find a threshold that would differentiate successively similar images from random ones so that they can be tagged as "movie".

5.3 Transfer Learning

	Aurora	Cupola	Volcano
Single*	93%	93%	93%
Multi-sig*	98%	98%	98%
Multi-soft	99.6%	100%	100%

Table 5: Final testing accuracy of various methods. See Appendix 10.1 for the full test results.

*A limited dataset of ~1000 images is used for these techniques.

For single-label softmax and multi-label sigmoid, no per-label accuracy has been calculated. While the multi-label sigmoid classification appears impressive at first sight, the accuracy is calculated in terms of the Hamming Loss. I.e. the fraction of labels that are predicted incorrectly. With ten labels and most images only having one label, guessing zero for each label would still result in 90% accuracy.

Finally, multi-label classification with a separate model for each tag provides the best final testing accuracy. While a certain bias is present in the provided dataset, additional testing on a set of random images from Google matches the bias inherent in the human tagging. I.e. a hurricane image is detected as a hurricane and not as clouds, identical to the dataset.

6. Conclusion

Our methodology consists of a breadth of approaches from which we pick the best results. Besting the general classifiers, Transfer Learning is the most accurate with the highest precision. Image-Histogram Entropy is the most statistically significant similarity metric.

7. Roles

Task (lines of code)	Lead
Implement transfer learning (300)	Frederick Joossens
Implement image labeling for end-user (200)	Frederick Joossens
Traditional Supervised Learning Models (500)	Greg Frasco
Unsupervised learning research (0)	Aditi Dass
Movie similarity metric models (500)	Aditi Dass
Prepare report and presentation	All

Table 6: Documentation of major tasks and leads

8. Link to GitHub repo

https://github.com/gregfrasco/WindowsOnEarth

9. References

Boutell, Matthew R., Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. “Learning Multi-Label Scene Classification.” *Pattern Recognition* 37 (9):1757–71.

Dai, Wenyuan, Qiang Yang, Gui-Rong Xue, and Yong Yu. 2007. “Boosting for Transfer Learning.” In *Proceedings of the 24th International Conference on Machine Learning - ICML ’07*. https://doi.org/10.1145/1273496.1273521.

Glorot, Xavier, and Yoshua Bengio. 2010. “Understanding the Difficulty of Training Deep Feedforward Neural Networks.” *International Conference on Artificial Intelligence and Statistics*.

Grandvalet, Yves, and Yoshua Bengio. 2005. “Semi-Supervised Learning by Entropy Minimization.” In *Advances in Neural Information Processing Systems 17*, edited by L. K. Saul, Y. Weiss, and L. Bottou, 529–36. MIT Press.

Hastie, Trevor, Saharon Rosset, Ji Zhu, and Hui Zou. 2009. “Multi-Class AdaBoost.” *Statistics and Its Interface* 2 (3):349–60.

Krasin, Ivan, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, et al. 2017. “OpenImages: A Public Dataset for Large-Scale Multi-Label and Multi-Class Image Classification.” *Dataset Available from Https://github.com/openimages*.

Panzeri, Stefano, and Alessandro Treves. 1996. “Analytical Estimates of Limited Sampling Biases in Different Information Measures.” *Network - Comp. Neural. 7*, 87–107.

Strong, S. P., Roland Koberle, Rob R. de Ruyter van Steveninck, and William Bialek. 1998. “Entropy and Information in Neural Spike Trains.” *Physical Review Letters* 80 (1):197–200.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. “Rethinking the Inception Architecture for Computer Vision.” In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.308.

10. Appendix

10.1.1 Final testing accuracy per label with transfer learning

	Aurora	Clouds	Cupola	Day	Satellite	Dock	Glacier	Hurricane
Single label*	93%	N/A	93%	93%	N/A	93%	N/A	N/A
Multi-label sigmoid*	98%	N/A	98%	98%	N/A	98%	N/A	N/A
Multi-label softmax	99.6%	99.5%	100%	97.7%	99.9%	99.2%	100%	99.5%

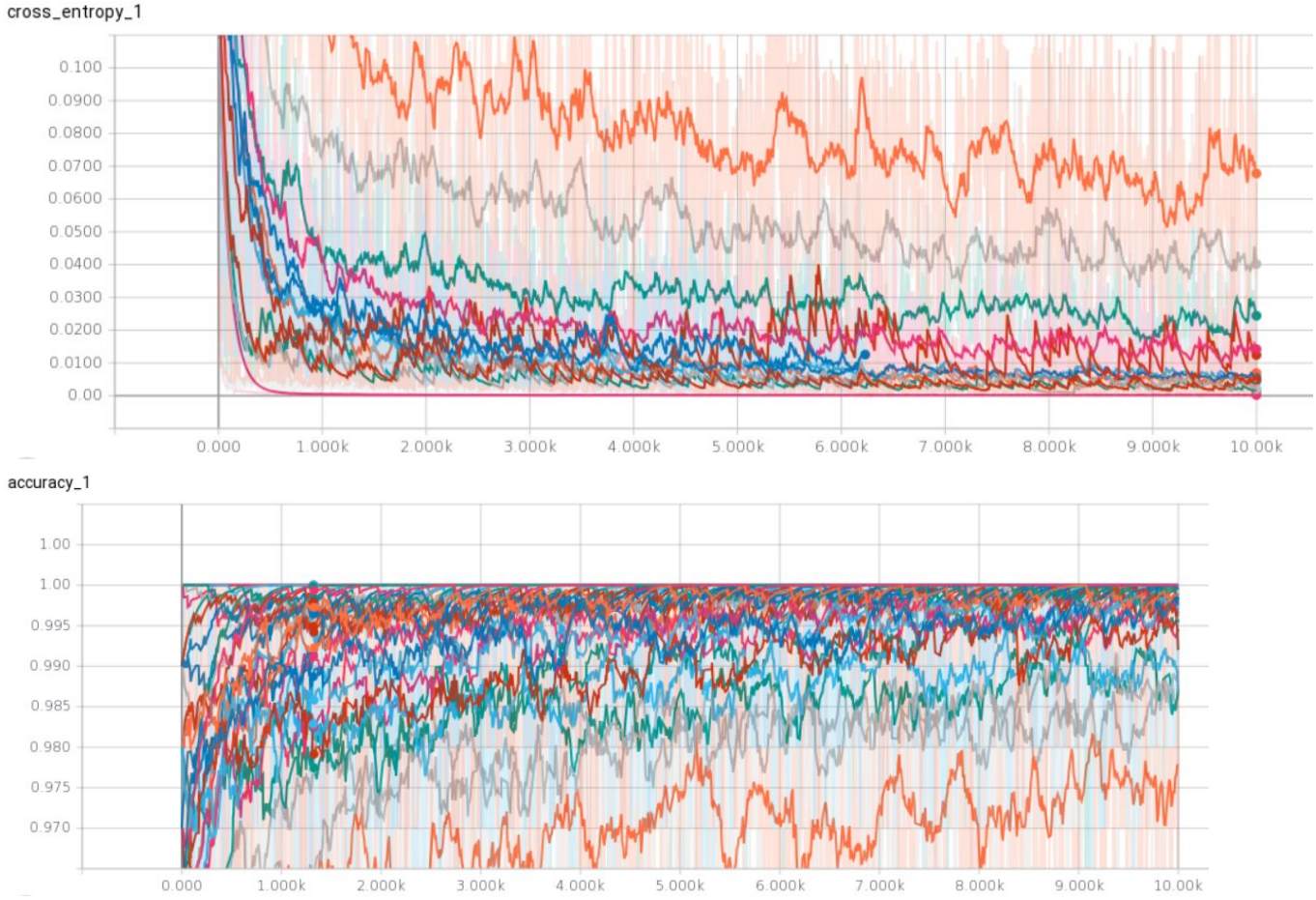
	Moon	Night	Panels	Eclipse	Stars	Structure	Sunrise	Volcano	Average
Single label*	93%	93%	N/A	N/A	93%	93%	93%	93%	93%
Multi-label sigmoid*	98%	98%	N/A	N/A	98%	98%	98%	98%	98%
Multi-label softmax	99.9%	99.4%	99.8%	99.8%	100%	98.8%	100%	99.7%	99.6%

10.1.2 Cross entropy loss for multi-label softmax training

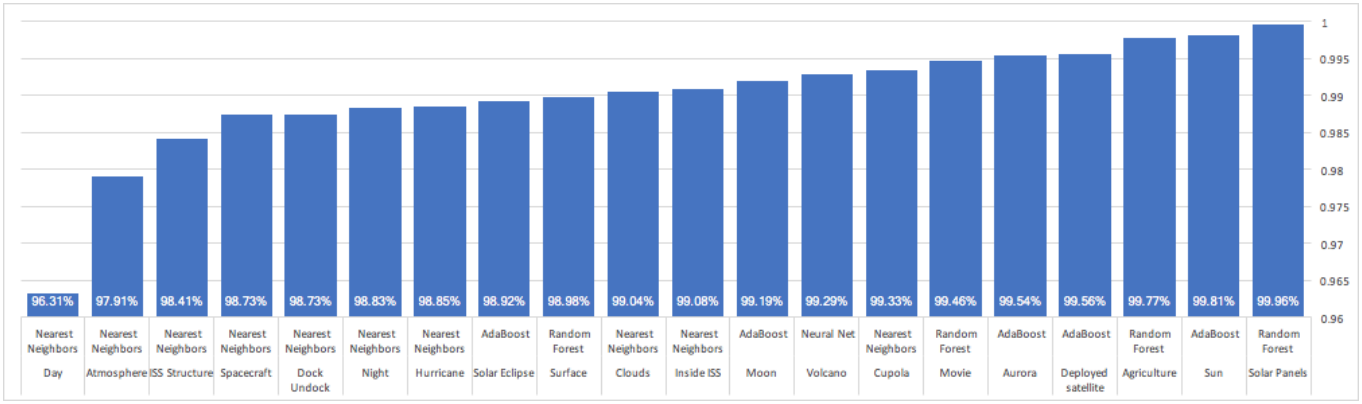
	Aurora	Clouds	Cupola	Day	Satellite	Dock	Glacier	Hurricane
Multi-label softmax	0.001155	0.016439	0.001902	0.051830	0.000560	0.008017	0.000405	0.005286

	Moon	Night	Panels	Eclipse	Stars	Structure	Sunrise	Volcano	Average
Multi-label softmax	0.000911	0.005291	0.005827	0.013217	0.000154	0.032848	0.000047	0.240020	0.023994313

10.1.2 Cross entropy and testing accuracy for multi-label classification per label



10.2 Validation accuracy for multi-label classification with multi supervised learning classifiers

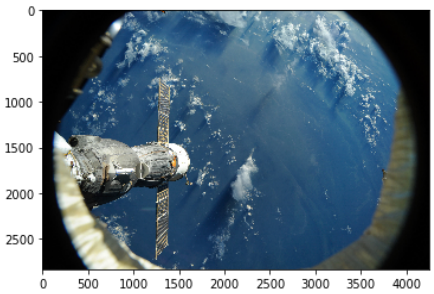


Tag	Neural Net	Random Forest	SVM	AdaBoost	Decision Tree	Naive Bayes	Nearest Neighbors	Best Classifier	Best Accuracy	Avg Per Tag
ISS Structure	95.33%	96.08%	95.74%	97.64%	96.77%	74.98%	**98.41%**	Nearest Neighbors	98.41%	93.56%
Day	85.79%	93.18%	73.48%	95.68%	94.57%	86.85%	**96.31%**	Nearest Neighbors	96.31%	89.41%
Movie	98.96%	**99.46%**	99.46%	99.42%	99.15%	69.72%	99.42%	Random Forest	99.46%	95.09%
Night	97.12%	96.74%	96.68%	98.23%	97.43%	83.10%	**98.83%**	Nearest Neighbors	98.83%	95.45%
Inside ISS	97.95%	97.37%	97.00%	99.08%	98.21%	76.80%	**99.08%**	Nearest Neighbors	99.08%	95.07%
Deployed satellite	98.91%	99.39%	99.33%	**99.56%**	99.27%	78.13%	99.25%	AdaBoost	99.56%	96.26%
Surface	98.73%	**98.98%**	98.98%	98.94%	98.71%	57.88%	98.85%	Random Forest	98.98%	93.01%
Agriculture	99.73%	**99.77%**	99.77%	99.69%	99.60%	65.58%	99.69%	Random Forest	99.77%	94.83%
Volcano	**99.29%**	99.29%	99.29%	99.17%	98.96%	76.49%	99.19%	Neural Net	99.29%	95.96%
Aurora	99.27%	99.27%	99.25%	**99.54%**	99.04%	86.25%	99.50%	AdaBoost	99.54%	97.45%
Clouds	97.62%	97.98%	97.98%	98.58%	97.71%	55.43%	**99.04%**	Nearest Neighbors	99.04%	92.05%
Sun	99.44%	99.08%	99.06%	**99.81%**	99.58%	99.52%	99.31%	AdaBoost	99.81%	99.40%
Solar Panels	99.92%	**99.96%**	99.96%	99.92%	99.96%	99.96%	99.96%	Random Forest	99.96%	99.95%
Hurricane	91.40%	91.63%	90.49%	96.95%	94.39%	74.75%	**98.85%**	Nearest Neighbors	98.85%	91.21%
Moon	96.24%	97.25%	89.61%	**99.19%**	98.35%	90.94%	98.71%	AdaBoost	99.19%	95.76%
Solar Eclipse	97.75%	97.89%	97.73%	**98.92%**	98.60%	79.70%	98.81%	AdaBoost	98.92%	95.63%
Spacecraft	97.23%	97.77%	97.73%	98.31%	97.98%	73.61%	**98.73%**	Nearest Neighbors	98.73%	94.48%
Dock Undock	97.22%	97.75%	97.73%	98.31%	98.04%	73.61%	**98.73%**	Nearest Neighbors	98.73%	94.49%
Atmosphere	89.15%	89.57%	88.48%	94.89%	92.41%	65.85%	**97.91%**	Nearest Neighbors	97.91%	88.32%
Cupola	97.91%	97.73%	97.23%	99.23%	98.23%	77.17%	**99.33%**	Nearest Neighbors	99.33%	95.26%

Average Per Classifier	96.75%	97.31%	95.75%	98.55%	97.85%	77.32%	**98.90%**	Nearest Neighbors	98.99%	94.63%

Accuracy table where each accuracy was the mean of the five cross fold accuracy

10.3 Examples of Image Preprocessing



Original Image A

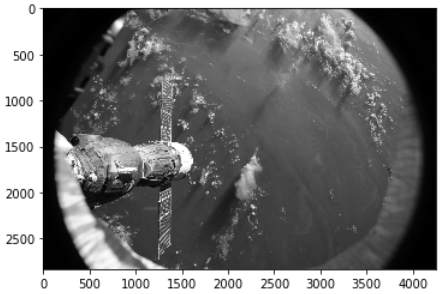


Image A Grayscale

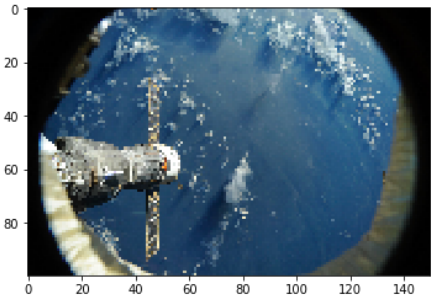


Image A Resized to (100,150)

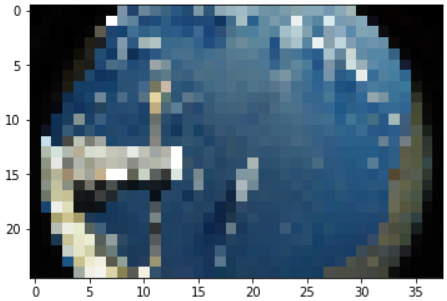
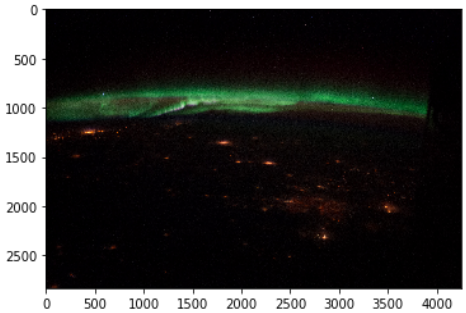


Image A Resized to (50,75)



Original Image B

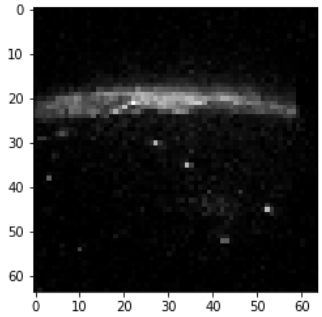


Image B Preprocessed