Computer Vision Approaches to Burned Area Image Segmentation

Serena Zhang
Stanford University
Computer Science
serena2z@stanford.edu

Matthew Villescas
Stanford University
Computer Science
mattjv22@stanford.edu

Iris Xia
Stanford University
Math and Comp. Sciences
ixxia@stanford.edu

1. Abstract

Understanding the extent and severity of burned areas following wildfires is an important goal and focus of ongoing research for scientists interested in the impact of climate change. In our project, we experiment with six image segmentation models to work towards this task, using remote sensing images from the lower resolution Landsat satellites as input and outputting segmentation masks that indicate what parts of the images are burned and not burned. Our models are built on the Deeplabv3 segmentation model from the Pytorch model library and an open-source U-Net model; we use pretrained versions of these models as baselines, and further experiment with adding infrared bands as input and with MAE loss functions that aim to reduce the impact of noise and low-resolution. We find that all methods achieve high accuracies, but the 5-band model that includes NIR (near-infrared) and SWIR (short-wave infrared) bands in addition to RGB performs the best for both Deeplabv3 and U-Net architectures. Our models work well for California wildfires in our dataset, but we hope to further generalize them to out-of-distribution fires in other parts of the world where fires aren't as well-documented. These advances would greatly help emergency preparedness, wildfire recovery, and climate science in these areas.

2. Introduction

In recent years, climate change has been significantly pronounced in the Western United States through the spread of wildfires, which have increased fivefold in California since 1971 [1]. Understanding the extent of burned area in wildfires is essential for modeling the impact that fires have on human health and vegetation growth, as well as impacts on greenhouse gas emissions and the carbon budget. Researchers have traditionally calculated burned area using bands from satellite products such as MODIS [2] or Landsat, but these calculations can often be noisy or low-resolution and haven't been generalized to areas outside of the United States and Canada.

For our project, we use deep learning and computer vision methods to automate the process of detecting burn scars in satellite images. Concretely, we feed into our models images taken from the Landsat satellites and output a mask that outlines burned areas in the image. The mask is a 0-1 array indicating where image pixels are part of the burned area. For data, we used images from California wildfires on a variety of topographies in the state. We trained a total of six models, each reflecting a technique across two standard architectures, the Deeplabv3 [2] and the U-Net [6]. For our baseline methods, we fine-tuned both Deeplabv3 and U-Net model for image segmentation with pre-trained resnet backbones. We then trained from scratch 5-Banded Deeplabv3 and Unet, architectures that take in RGB and the NIR and SWIR channels as input from data. We also trained an RGB-based Deeplabv3 and Unet model with Mean Absolute Error for the loss function [5].

Another goal of our project is to learn to generalize our models to out of distribution areas outside of California, where our dataset is from. While California has a relatively robust system for detecting and analyzing wildfire burn scars, other parts of the world, such as the Amazons, have regular wildfires that aren't as well-documented. With our segmentation models, we're able to provide a lot more information about wildfires in these areas.

3. Related Work

3.1. SVM Classification

[4] compared a rules-based approach with a supervised classification model using Sentinel-2 and Landsat 8 optical data. For the rules-based approach, the authors calculated the NIR ratio and RBR spectral indices using the NIR (near infrared) and SWIR (short-wave infrared) bands on the satellite images, and identified thresholds for these values that would determine whether segments of the image were part of the burned area. This represents a fairly traditional method of using spectral indices to approximate features of burn scars. The supervised classification model trained a SVM classification algorithm to recognize segments of the images as ”burned” and ”unburned” and merged features

of the same classified value to create a final mask. When comparing the output masks with wildfire maps from the Emergency Management Service, the supervised learning algorithm did slightly better, with an average overlap percentage of 89.65, while the rules-based method had an average overlap percentage of 86.15. This paper demonstrated the effectiveness of using deep learning methods over traditional rules-based methods, and the SVM classification model is one of the potential baseline models we can use. This model also used Landsat images for training, which we'll also be using in our project.

3.2. Deep U-Nets

[3] explored deep learning architectures to specifically address the task of segmenting satellite images of burned areas over various topographies and during various seasons. Cultivating a custom dataset on high resolution Sentinel 2 satellite images of burned areas, they trained a random forest (RF) and U-Net CNN to segment burned and non-burned areas of their images. Both methods achieved scores of over 0.85 on accuracy, precision, recall, and F1 metrics. The U-Net CNN outperformed the RF in all metrics, achieving scores of at least 0.95, making it a very attractive baseline model to compare to. Notably, the U-Net was not pretrained on any data. Motivated by the results of this paper, we experimented with the model architecture, with both pretrained and non-pretrained versions.

[8] also compares several deep learning approaches applied to their new dataset BurnedAreaUAV, which includes remote sensing images taken from aerial videos. This paper contrasted a few U-Net models, including a U-Net RED model that only takes as input the red channels of the images, and U-Net 3D that performs 3D convolutions. They found that their base U-Net model with 2D convolutions performed the best on their new dataset with an intersection over union (IoU) of 95 percent on the test set.

4. Datasets

Our dataset comprises of wildfire images from California obtained from Landsat Collection 2. The labels are derived from the Monitoring Trends in Burn Severity (MTBS) project, which assesses burn severity and areas affected by wildfires across the United States. The dataset is organized into pre-fire and post-fire imagery, capturing scenes before and after each fire event. Consequently, fire scars are visible in all post-fire images. Landsat imagery also has an average revisit interval of 8 to 16 days, resulting in multiple images for many of the fire events. The dataset includes both visible and infrared bands for a total of 6 channels (with one channel being a quality assurance mask to mask out any clouds and other debris that may obscure parts of the image), which can be used to calculate various vegetation and burn spectral indices.

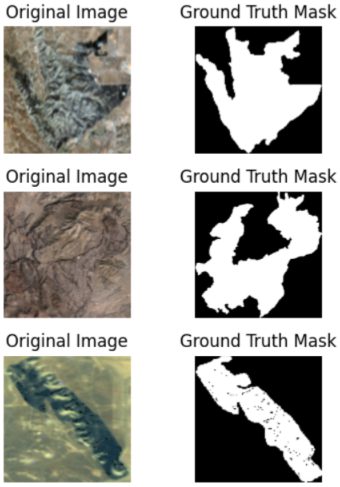

Figure 1. Shown here are some example images and ground truth segmentation masks from our dataset. The black regions represent burned areas and white regions represent non-burned.

Our initial dataset size comprises of approximately 1700 images, which represents a smaller subset of the MTBS dataset. This reduced size was chosen intentionally to streamline computational resources and reduce storage while still maintaining the integrity of the dataset for our analysis. Furthermore, these images represent the median of multiple pictures taken of the same area over time, to reduce the probability of thick clouds and other clutter from obstructing the burned areas. For dataset pre-processing, we first use the QA channel to clean up pixels that aren't informative for fire burn segmentation (i.e. snow or clouds that are pictured). Next, we generate binary pixel masks of the burn area (0 for non-burned, 1 for burned) from the MTBS labels. For our RGB 3-channel models, We employ data augmentation to increase the robustness and diversity of our dataset, thereby enhancing the model's ability to generalize across various scenarios and improve overall performance. In particular, we employ random cropping, random horizontal flip (p = 0.5), random vertical flip (p = 0.5), random affine transformations (rotation, translation, scale and shear with p = 0.3), and color jitter. Lastly, we split our dataset into train/val/test splits using 80/10/10 proportions.

5. Methods

We implement two baseline segmentation models on our processed dataset, Deeplabv3 and U-Net. For each of these two models, we keep the parameters constant and change only the model architecture. On top of these two models, we also experiment with two different changes: creating a 5-band segmentation model to include not only the RGB bands but also the NIR and SWIR bands, and using a loss function that is more robust to noise (MAE loss).

5.1. Baseline #1: Deeplabv3 Segmentation

Deeplabv3 is a semantic segmentation model that incorporates specialized modules that utilize dilated convolution either sequentially or concurrently, allowing for the capture of multi-scale contextual information.

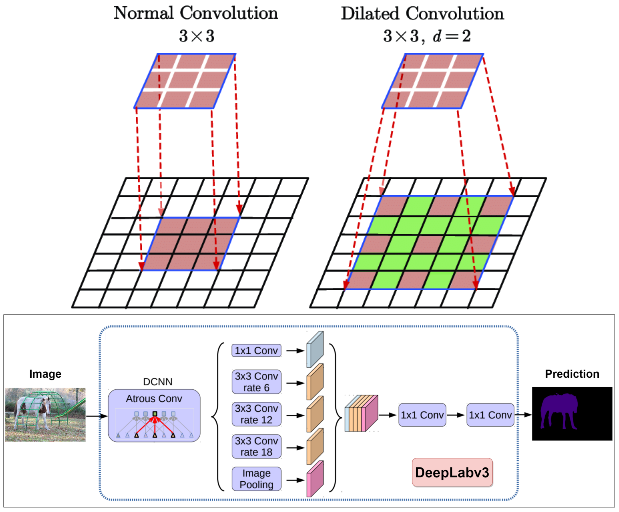

Figure 2. Dilated convolution and architecture of Deeplabv3 model

We chose to use Deeplabv3 with a Resnet-50 backbone pretrained on ImageNet as our first baseline because it is one of two easily accessible segmentation models from the Pytorch model library. In studies comparing Deeplabv3 with other segmentation models like Fully Convolutional Networks (FCN), also from the model library, Deeplabv3 tended to do better [1]. We confirmed these studies as well when experimenting with Fully Convolutional Networks as our baseline, finding that the Deeplabv3 model performs better. Other studies in our reviewed literature have also used Deeplabv3 models for segmentation of Sentinel 2 satellite imagery [2].

For our baseline implementation, we used cross entropy loss as our loss function, as defined below

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{1}$$

where M is the number of classes, y is the binary indicator (0 or 1) if class label c is the correct classification for observation o, and p is the predicted probability observation o is of class c. Note that p is calculated as a softmax distribution over the inputs x.

5.2. Baseline #2: U-Net Segmentation

U-Net is a widely used semantic segmentation model, especially in biomedical image segmentation. Its architecture features a symmetric encoder-decoder structure: the encoder captures the image context by down-sampling through convolutional and pooling layers, while the decoder restores the spatial resolution by up-sampling and combining feature maps from the encoder via skip connections. These skip connections retain fine-grained spatial details, leading to more accurate segmentation results. Thus, in contrast to Deeplabv3 which uses dilated convolutions to capture multi-scale context, U-Net directly merges high-resolution features from the encoder with upsampled features in the decoder. This allows U-Net to excel in tasks requiring precise boundary delineation by preserving detailed spatial information.

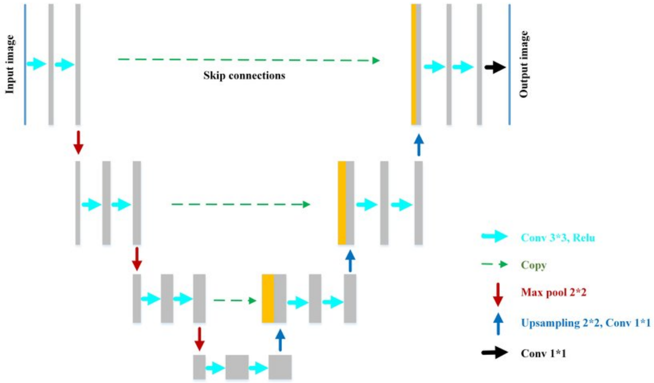

Figure 3. U-Net model architecture

We picked U-Net for our project due to its proven effectiveness for diverse use cases and in scenarios where high precision is critical. Its architecture is particularly advantageous when dealing with limited training data, as the symmetric design and skip connections help the model generalize better and leverage the available data more efficiently. And as described in our literature review section, U-Nets have been demonstrated to be effective in wildfire burn mask segmentation.

For our baseline implementation, we used the pretrained U-Nets from [6], and specifically used the U-Net pretrained on a resnet-50 backbone. We also used cross entropy loss for the loss function.

5.3. 5-Band Segmentation

On top of our baseline, we decided to experiment with our input bands. Satellite images usually contain bands from across the electromagnetic spectrum (in addition to visible RGB bands), and our Landsat data contains images with five bands with respect to this: red, green, blue, NIR (near-infrared), and SWIR (short-wave infrared). The NIR band is particularly useful for vegetation analysis, as it can help differentiate between healthy and stressed vegetation due to its sensitivity to chlorophyll. By incorporating the NIR band into our model, we aim to improve the accuracy of our segmentation tasks, particularly in distinguishing vegetation from other land cover types. Additionally, the SWIR band provides valuable information on soil and moisture content, which can further enhance the model's ability to classify different surface materials accurately. Experimenting with these multi-banded inputs allows us to leverage the full spectrum of available data, potentially leading to more robust and insightful segmentation results.

To incorporate the NIR and SWIR bands into the input, we train both of our baseline models from scratch, changing the input layer from 3 to 5 channels. Since we are no longer using the pretrained ImageNet models, we normalize each channel of the image based on its own unique distribution (by calculating the mean and standard deviation of each channel) to between 0 and 1.

5.4. MAE Loss

Finally, we experiment with implementing a different loss function more robust to noise. In particular, we wanted to accomodate the lower resolution satellite images we have. Besides being prone to noise such as clouds/atmospheric noise, the lower resolution of Landsat 2 data compared to datasets such as Sentinel 2 may pronounce or even introduce noise, making its reduction a pertinent challenge. Specifically, we experimented with using mean absolute error (MAE) loss [5], known for its tolerance to label noise, and implemented a data sampling approach to ensure that only low-loss training samples contribute to gradient updates [7] [10].

$$\sum_{i=1}^{D} |x_i - y_i| \tag{2}$$

In this loss sampling strategy, the model processes a batch of inputs as usual, but before averaging the loss across samples, we identify the training sample with the highest loss in the batch. The loss for this outlier is then set to 0, preventing it from influencing the weight updates. According to [10], this method operates on the assumption that as model performance improves, particularly noisy samples will produce high losses, which can significantly impact weight updates. This training strategy helps mitigate the influence of such samples.

5.5. Experiments

We ran a total of six experiments using the models described above: Deeplabv3 baseline, U-Net baseline, Deeplabv3 with 5 band input, U-Net with 5 bands, Deeplabv3 with MAE loss, and U-Net with MAE loss. The results of these experiments are described in the following section.

In all of our experiments, we use Adam as our optimizer, and a learning rate of 0.001. We implemented learning rate decay, with a step size of 25 and a decay rate of 0.1. All of our models were trained for 50 epochs, and the models with the lowest validation loss and highest validation IoU score were saved. Fifty epochs were chosen as it provides a balance between training time and model performance. This duration allows the models to learn sufficiently from the data without overfitting, ensuring that we capture the optimal weights for accurate segmentation.

6. Results

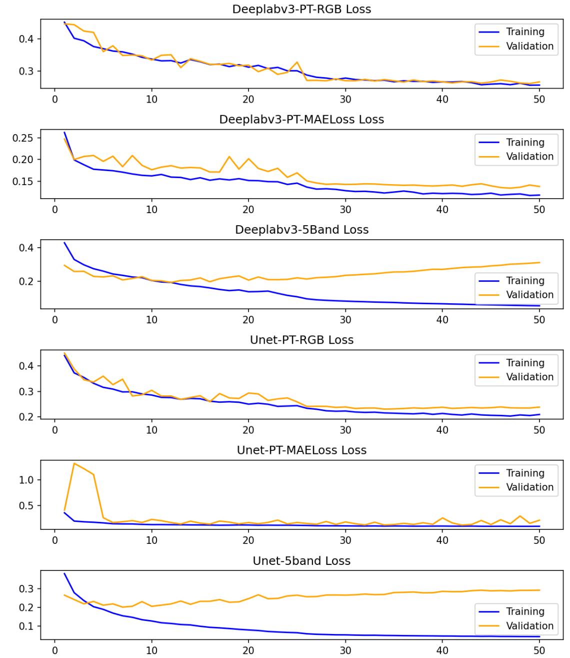

Figure 4. Loss Curves by Model

The main metrics we used to evaluate our model performance are the intersection over union (IoU) score and the Dice score.

IoU is a standard evaluation metric for segmentation tasks, measuring the overlap between the predicted segmentation mask and the ground truth. It is defined as the ratio of the intersection area to the union area of the predicted and ground truth masks. Mathematically, IoU can be expressed as:

$$\text{IoU} = \frac{|\text{Predicted} \cap \text{Ground Truth}|}{|\text{Predicted} \cup \text{Ground Truth}|}$$

Here, $|\text{Predicted} \cap \text{Ground Truth}|$ represents the number of pixels common to both the predicted and ground truth segments, while $|\text{Predicted} \cup \text{Ground Truth}|$ represents the total number of pixels in either the predicted or ground truth segments. A higher IoU score indicates better performance, with a value of 1 indicating perfect overlap and 0 indicating no overlap at all.

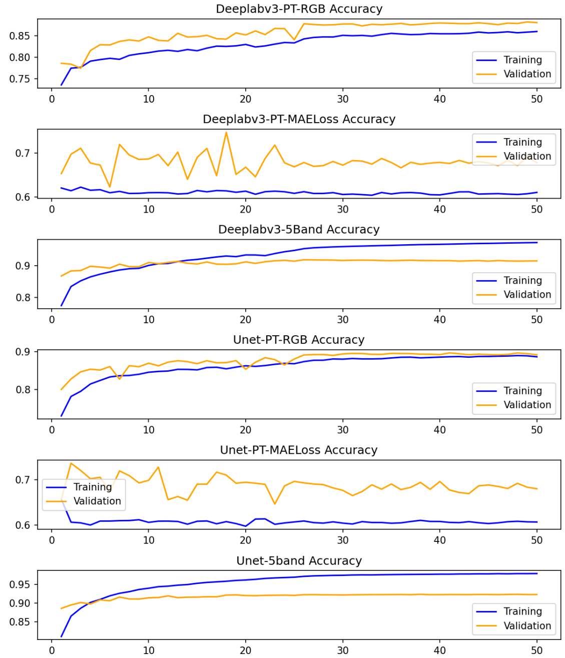

Figure 5. Accuracy (Dice) Curves by Model

The Dice score, also known as the Dice coefficient or F1 score, is another metric used to measure the similarity between the predicted and ground truth segments. It is calculated as:

$$\text{Dice Score} = \frac{2 \times |\text{Predicted} \cap \text{Ground Truth}|}{|\text{Predicted}| + |\text{Ground Truth}|}$$

The Dice score takes into account both the precision and recall of the prediction, providing a balanced measure of overlap. Like IoU, a Dice score of 1 indicates perfect agreement between the predicted and ground truth segments, while a score of 0 indicates no overlap.

Using both IoU and Dice score as our evaluation metrics ensures a comprehensive assessment of our models' segmentation performance, capturing both the extent of overlap and the balance between false positives and false negatives.

6.1. Training

Examining the training trajectories across models presented in Figures 4 and 5, we can see consistent behaviour with respect to specific augmentations.

In both 5Band-based models, we see that the validation loss is steadily increasing while the training loss is decreasing. This implies that overfitting is occurring, and there could be several reasons for why. One possible explanation is the size of the dataset. In particular, around 1300 training images may simply not be enough to produce more robust behaviour for models that take almost double the number of input channels as a normal model.

Another possibility is the content of the dataset itself. There are multiple topographies at play, and the randomly sampled training set may be dominated by certain features not pronounced in the validation dataset. It is notable that in the baseline models, both exhibit relatively healthy behaviour in the development of the loss curves. This implies that the more dominant features at play in the training data would be in the extra NIR and SWIR bands, and so proper diversification across these channels ought to be considered for future runs. We also had to recompute statistics across our training samples to normalize this data. The means and standard deviations for channels R-G-B-NIR-SWIR respectively were (0.2583, 0.2491, 0.2366, 0.3691, 0.3544) and (0.1187, 0.1079, 0.0969, 0.1400, 0.1395). Although the standard deviations seem relatively close, the NIR and SWIR bands do have higher values. There is also considerable difference between the means of the NIR and SWIR bands than the other channels, which may reflect a dominating significance of those channels that took place during training.

The loss curves for the MAELoss models are better behaved and exhibit similar development in training as the baseline models. The primary differentiating factor is in the accuracy observed, which exhibits consistent behaviour across Deeplab and Unet variations. The validation accuracy is relatively unstable for early epochs before settling towards a middling result despite the loss still decreasing. The scale of the MAELoss accuracy is strictly less than that of the baseline, but the validation accuracy is considerably greater than training accuracy. This makes the MAELoss a promising option for robustness, but its overall low accuracy implies a hybrid approach with another loss function may be required, if it is to be used at all.

6.2. Comparing Model Performance

Method	Accuracy (IoU)	Dice Score (F1)
Unet-Pretrained-RGB	0.82	0.89
Unet-5band	**0.84**	**0.91**
Unet-MAEloss	0.80	0.88
Deeplabv3-Pretrained-RGB	0.80	0.88
Deeplabv3-5band	**0.83**	**0.90**
Deeplabv3-MAEloss	0.78	0.87

Table 1. Comparing IoU and Dice Score across the six segmentation models

Utilizing 5-band input data enhances model performance, as evidenced by the improved IoU and Dice Score for both Deeplabv3 and Unet models. Deeplabv3-5band achieves the second highest performance with an IoU of 0.83 and a Dice Score of 0.90, while Unet-5band achieves the highest performance with an IoU of 0.84 and a Dice Score of 0.91. This suggests that additional spectral information is beneficial for segmentation tasks with remote sensing images. Conversely, models trained with Mean Absolute Error (MAE) loss show slightly lower performance, with Deeplabv3-MAEloss and Unet-MAEloss achieving IoUs of 0.78 and 0.80, and Dice Scores of 0.87 and 0.88, respectively.

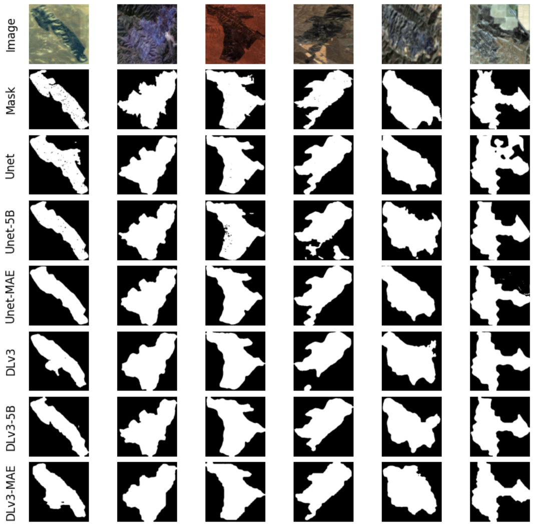

Figure 6. Segmentation results compared to ground truth for each model. An enlarged version of this image is included below in the Appendix section.

For a qualitative analysis of the segmentation results, we compare the output masks across models on samples of the data (an enlarged version of the image has been provided in the appendix). It seems 5Band can be helpful in picking up on finer details of the image, as displayed in columns 1 and 6. Both the U-Net and Deeplab baseline models pick up on significant clusters of pixels that drastically change the contiguous shape of the mask, which the extra bands help to smooth out. However, for tailed features, as in columns 4 and 5, 5Band seems to overcorrect or even misrepresent the thin shape. The MAELoss models frequently take on roughly similar shapes as the baseline models with certain features exaggerated as demonstrated in columns 1 and 3, but column 4 is an interesting instance in which both MAELoss models handle tailed features relatively well compared to 5Band counterparts.

We also looked to compare our results with those in the literature, but it was hard to directly compare IoU and dice score results as we used very different datasets from those used in the literature. One avenue of future work could be testing the models from these papers on our dataset, or testing our models on datasets from these papers to create a better comparison.

6.3. Discussion

These observations highlight the advantages and disadvantages of using pretrained models. Arguably, pretrained models typically perform better for out-of-distribution (OOD) data due to their extensive training on large and diverse datasets. However, one disadvantage is that these models may not be well-suited for a specific task. Models trained from scratch may outperform in specific fields or industries; in our case, satellite imagery differs a lot from the images in ImageNet that our models were pretrained on. In such cases, the additional spectral bands available in satellite imagery may provide useful information that pretrained models on RGB 3-channel data may not fully exploit.

It is notable that the validation accuracy (which was used to pick the model) of the 5Band model marginally increased after only a few epochs of training, so while the potential for using more bands is very promising, yielding more successful results will require either a larger abundance of data or potentially more balanced data across all bands. Still, even with limited amounts of data, 5Bands was the best performing model.

Additionally, the choice of loss function significantly affects model performance in segmentation tasks. Mean Absolute Error (MAE) loss, which measures the average absolute difference between predicted and actual values, treats all errors equally; there is less sensitivity to the error size to ensure that the trained model isn't over-penalizing large errors, especially if the dataset contains outliers or unlikely-to-occur data samples, and this makes MAE loss robust to noise. However, this can make it less effective for precise segmentation compared to Cross-Entropy (CE) loss. CE loss is better at handling binary classification tasks by heavily penalizing misclassifications, making it more suitable for tasks that require fine distinctions. Our findings aligned with previous research [9] indicating that MAE generally underperforms compared to CE in segmentation tasks, despite the noise in loss resolution satellite imagery.

7. Conclusion & Future Work

Satellite imagery can be very challenging to process, as it comes at different resolutions and is prone to obstructions. The traditional context of computer vision data, that being the RGB channels which represent electromagnetic radiation that is explicitly visible, produce strong results for classifying the images of forest burns, even when taken from so far away. However, tapping into the other invisible electromagnetic spectra explicitly adds more information that is demonstrably beneficial to semantic segmentation. This very directly opens future work that utilizes data with more electromagnetic spectra past NIR and SWIR such as panchromatic or thermal infrared. There are also open questions about the specific significance of the visible light spectra. Although a model that only used NIR and SWIR channels would probably be worse, the success of a model that didn't use the visible light spectrum at all would be very illuminating as to the importance of including these invisible bands, whether their contribution is marginal or very drastic.

We hoped that we could combat the noisy nature of loss resolution and obstructed satellite images with a different loss function, but this ultimately gave suboptimal results. It remains to be seen whether this would drastically hinder a 5Band model, but at the very least, the question is open over whether a different loss function could considerably improve results. Further, there was some promise in select examples of the MAELoss function in picking up thinner, tailed patches of masks, which may imply that a hybrid approach across loss functions via some sort of weighted average or other consolidation method could indeed be helpful.

Finally, higher resolution implicitly encodes more data that would likely allow models to make more accurate predictions. Attempting both approaches in a higher resolution context would be beneficial in at least evaluating to what degree these methods can succeed in combating image resolution challenges. If the benefits are marginal, than lower resolution data (which is generally more economical) may suffice for practical purposes. There are ongoing projects in Professor Burke's lab working on using computer vision methods to increase the resolution of satellite images, and these results may help us determine if our approaches work on higher-resolution images.

As we work on extending our project over the summer, some other future work we hope to do include using pre-fire images and generalizing our model to out of distribution data. We were limited in this project by our data size and compute power, but we had hoped to bring in more sources of data to test our model in an OOD context. One way that we're planning to do this using the data that we have is by splitting our data between different vegetation types; California is dominated by conifers and shrublands, so one way to achieve OOD testing is by training only on images from conifer areas and test on shrubland images. We also plan to use pre-fire images as negative examples (with empty masks) to add to our dataset, which may extend our model to also determine if the image has no burn scar at all.

8. Contribution & Acknowledgement

Serena: Worked on preprocessing the data and setting up training and testing code infrastructure; also implemented the Deeplabv3 model, MAE loss functions, and worked on model evaluation.

Matthew: Worked on preprocessing data and training both the Deeplabv3 and U-Net baseline models; also implemented the 5-band input model and worked on output visualizations. Babysat way too many epochs by now.

Iris: Worked on setting up data storage and compute resources, preprocessing data and training the U-Net baseline model; also experimented with fully convolutional networks (not included in our final set of models) and worked on output visualizations.

All contributed equally to writing and editing the final report, with Iris focusing on the abstract/introduction, literature search, and methods, Serena focusing on the dataset, methods, and results, and Matthew focusing on the discussion and conclusions sections.

We would also like to acknowledge Iván Higuera-Mendieta, a Ph.D. student in Professor Burke's lab, for helping us conceive the project idea and mentoring us throughout the quarter. Iván supplied us with the data and ideas for using MAE loss, and also helped identify articles for our literature search.

We also used an open-source Github implementation for our U-Net baseline, which can be found here: https://github.com/mberkay0/pretrained-backbones-unet/tree/main. This is also cited in our references section.

9. Appendices

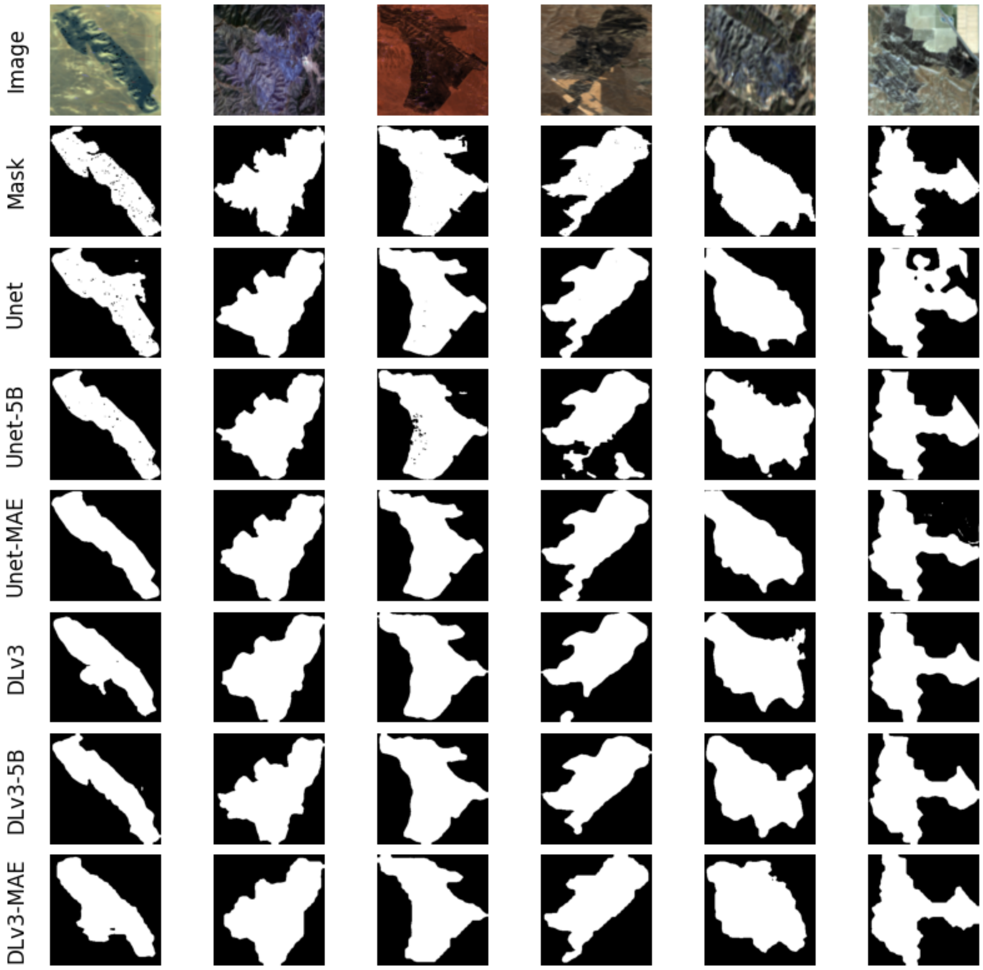

Figure 7. [ENLARGED] Segmentation results compared to ground truth for each model

References

[1] I. Ahmed, M. Ahmad, F. A. Khan, and M. Asif. Comparison of deep-learning-based segmentation models: Using top view person images. *IEEE Access*, 8:136361–136373, 2020. 3

[2] Clabaut, S. Foucher, Y. Bouroubi, and M. Germain. Synthetic data for sentinel-2 semantic segmentation. *Remote Sensing*, 16(5), 2024. 1, 3

[3] L. Knopp, M. Wieland, M. Rättich, and S. Martinis. A deep learning approach for burned area segmentation with sentinel-2 data. *Remote Sensing*, 12(15), 2020. 2

[4] I. Kotaridis and M. Lazaridou. Integrating image segmentation in the delineation of burned areas on sentinel-2 and landsat 8 data. *Remote Sensing Applications: Society and Environment*, 30(100944), 2023. 1

[5] X. Ma, H. Huang, Y. Wang, S. R. S. Erfani, and J. Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 1, 4

[6] B. Mayali. pretrained-backbones-unet. https://github.com/mberkay0/pretrained-backbones-unet. 1, 3

[7] A. B. Savel, E. M.-R. Kempton, M. Malik, T. D. Komacek, J. L. Bean, E. M. May, K. B. Stevenson, M. Mansfield, and E. Rauscher. No umbrella needed: Confronting the hypothesis of iron rain on wasp-76b with post-processed general circulation models. *The Astrophysical Journal*, 926(1):85, Feb. 2022. 4

[8] J. M. R. L. d. C. C. Tiago F.R. Ribeiro, Fernando Silva. Burned area semantic segmentation: A novel dataset and evaluation using convolutional networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202, 2023. 2

[9] J. Wen and M. Burke. Wildfire smoke plume segmentation using geostationary satellite imagery, 2021. 6

[10] F.-F. Xue, J. Peng, R. Wang, Q. Zhang, and W.-S. Zheng. Improving robustness of medical image diagnosis with denoising convolutional neural networks. In D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 846–854, Cham, 2019. Springer International Publishing. 4