
SEMANTIC SEGMENTATION OF LOW EARTH ORBIT SATELLITES USING CONVOLUTIONAL NEURAL NETWORKS

Justin Fletcher, et al.

Air Force Research Laboratory
550 Lipoa Parkway
Kihei, HI 96753

08 February 2022

Technical Paper

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION IS UNLIMITED.



AIR FORCE RESEARCH LABORATORY
Directed Energy Directorate
3550 Aberdeen Ave SE
AIR FORCE MATERIEL COMMAND
KIRTLAND AIR FORCE BASE, NM 87117-5776

REPORT DOCUMENTATION PAGE

1. REPORT DATE 08 February 2022		2. REPORT TYPE Presentation		3. DATES COVERED	
				START DATE	END DATE
4. TITLE AND SUBTITLE SEMANTIC SEGMENTATION OF LOW EARTH ORBIT SATELLITES USING CONVOLUTIONAL NEURAL NETWORKS					
5a. CONTRACT NUMBER		5b. GRANT NUMBER		5c. PROGRAM ELEMENT NUMBER	
5d. PROJECT NUMBER		5e. TASK NUMBER		5f. WORK UNIT NUMBER No associated work unit number	
6. AUTHOR(S) Justin Fletcher Julia Yang Trent Kyono Andrew Van Berg Jacob Lucas Michael Abercrombie					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Air Force Research Laboratory 550 Lipoa Parkway Kihei, HI 96753				8. PERFORMING ORGANIZATION REPORT NUMBER AFRL-RD-PS-TP-2023-0011	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)			10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/RDSM		11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-2022-1080
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution A: Approved for public release; distribution is unlimited. Public Affairs release approval # AFRL-2022-1080					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT Semantic segmentation is taking an image that has different components, like sky, tree, cat, cow, etc. and identifying which parts of the image belong to each component. For this study, we wanted to see if we could transfer this task over to images of satellites to identify which parts of the images were the different components of the satellite such as bus, solar panel, antenna, payload, etc. This can be very helpful for a number of applications including providing health and status of a satellite, or understanding its behavior. With the pristine images, this might not be too hard. By eye, we can roughly produce a segmentation, but with turbulent images, this becomes much, much harder. And for either turbulent or pristine images, semantics segmentation becomes very tedious with large volumes of images. However, semantic segmentation of images in other fields such as biology and even day-to-day applications has seen revolutionary improvements through convolutional neural networks, a machine learning technique. So we wanted to apply CNN to semantic segmentation of satellites as well.					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT SAR		18. NUMBER OF PAGES 31
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			
19a. NAME OF RESPONSIBLE PERSON Justin Fletcher				19b. PHONE NUMBER <i>(Include area code)</i>	



Semantic Segmentation of Low Earth Orbit Satellites using Convolutional Neural Networks

Julia Yang

The Boeing Company

Jacob Lucas (Boeing), Trent Kyono (Boeing), Michael Abercrombie (Boeing),
Andrew Vanden Berg (AFRL), Justin Fletcher (USSF SSC)

March 09, 2022



LEO Imaging in Maui

- **The Air Force Maui Optical and Supercomputing (AMOS) site produces resolved images of LEO Satellites**
 - 3.6-m telescope on Mt. Haleakala (10k feet)
 - Confounded by varied viewing geometries and turbulence conditions
 - Turbulence mitigation includes adaptive optics (AO) and various algorithms



AFRL

LEO Imaging in Maui

- **The Air Force Maui Optical and Supercomputing (AMOS) site produces resolved images of LEO Satellites**
 - 3.6-m telescope on Mt. Haleakala (10k feet)
 - Confounded by varied viewing geometries and turbulence conditions
 - Turbulence mitigation includes adaptive optics (AO) and various algorithms



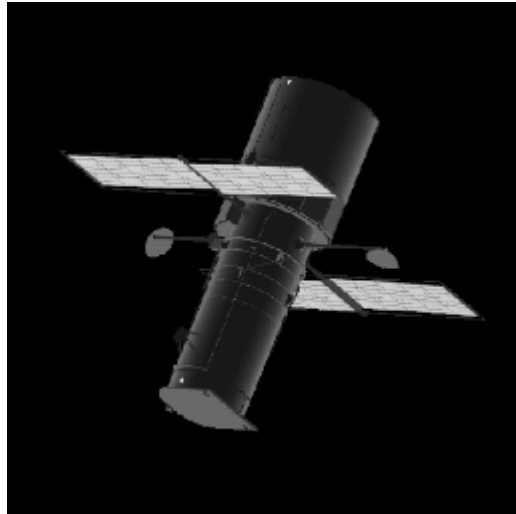
THE AIR FORCE RESEARCH LABORATORY

Approved for public release; distribution unlimited.

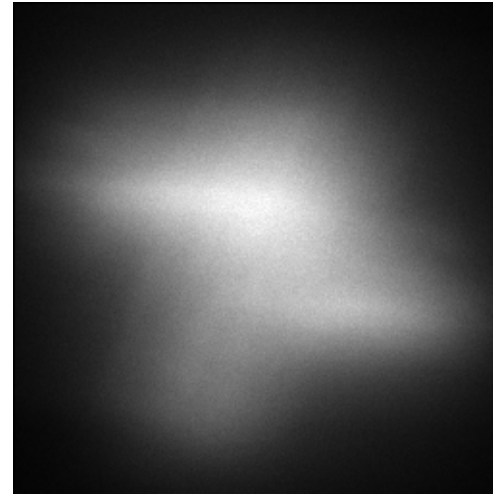
2



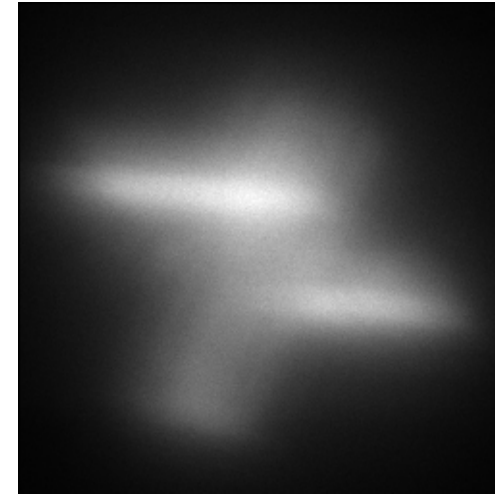
Representative Seeing Conditions



Pristine (no turbulence)



Poor ($r_0 = 10\text{cm}$)



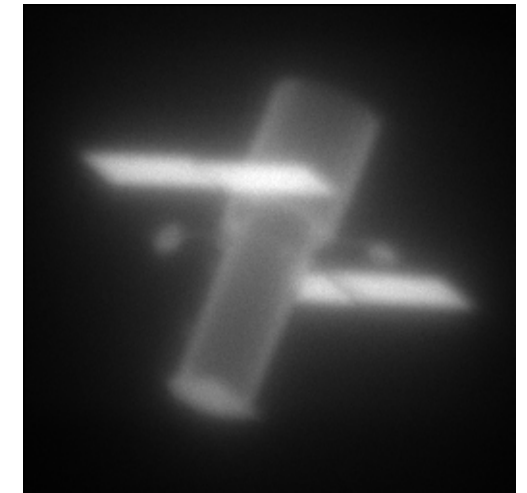
Average ($r_0 = 15\text{cm}$)



Good ($r_0 = 25\text{cm}$)



Exceptional ($r_0 = 40\text{cm}$)



Typical AO ($r_0 = 80\text{cm}$)

And to get a sense of how much the turbulence and adaptive optics can affect image quality, lets take a look at some example images. So on the left here, we have a render of the hubble space telescope in the absence of any turbulence. Throughout this presentation, I'll be using the term pristine to refer to these original images without any turbulence. and I'll refer to the turbulent images by their r_0 number in terms of centimeters, r_0 is just a quantifiable way to measure the level of blurriness in an image.

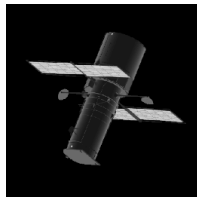
Now on the right here, we have the same render of HST, but with turbulence. We have the same image with a r_0 of 10cm, which is representative of images taken at AEOS telescope on days with poor seeing conditions. Then we have r_0 of 15cm, equivalent to an Average day, r_0 of 25cm- a good day, r_0 of 40cm - an exceptional day, and finally r_0 of 80cm - which represents images taken from AEOS telescope after typical application of Adaptive Optics.

As you can tell, these images can get pretty difficult to decipher, and especially when collecting large volumes of images, gathering information from these images can become very difficult and increasingly susceptible to error.

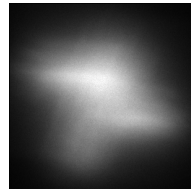
One such task of gathering information is the task of semantic segmentation of satellites, And for those who are unfamiliar with semantic segmentation - a quick crash course...

AFRL

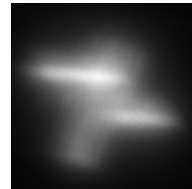
Representative Seeing Conditions



Pristine (no turbulence)



Poor ($r_0 = 10\text{cm}$)



Average ($r_0 = 15\text{cm}$)



Good ($r_0 = 25\text{cm}$)



Exceptional ($r_0 = 40\text{cm}$)



Typical AO ($r_0 = 80\text{cm}$)

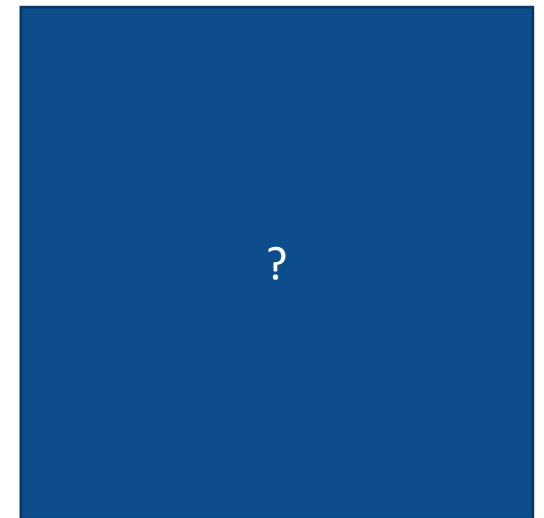
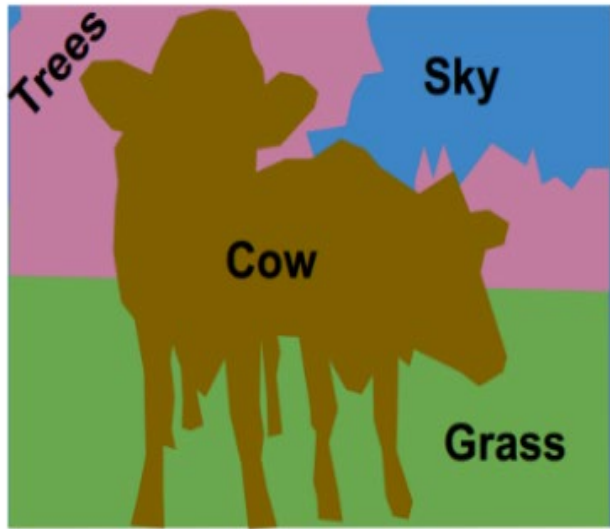
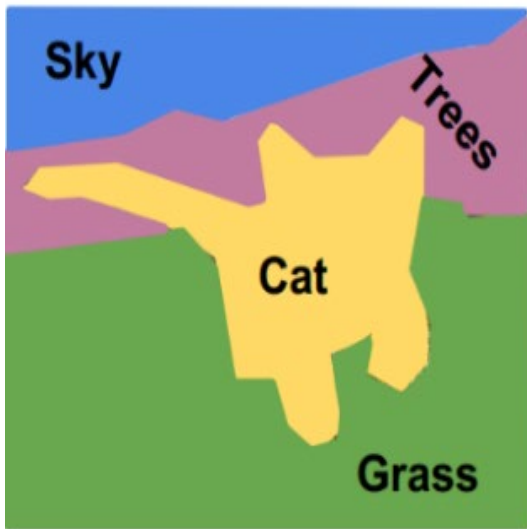
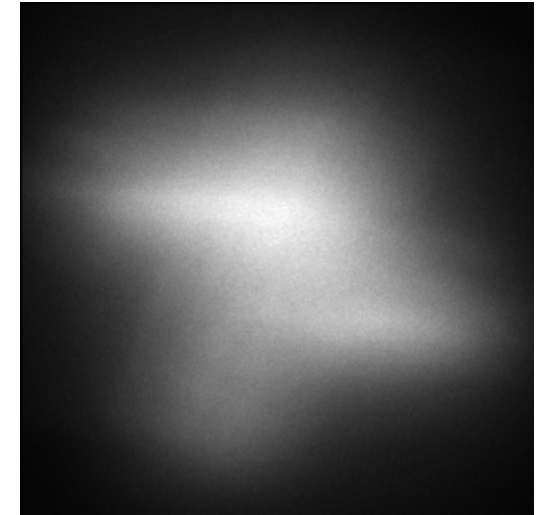
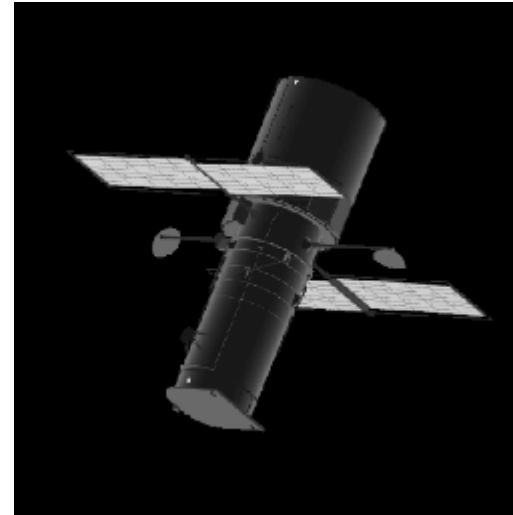
THE AIR FORCE RESEARCH LABORATORY

Approved for public release; distribution unlimited.

3



Semantic Segmentation



Semantic segmentation is taking an image that has different components, like sky, tree, cat, cow, etc. and identifying which parts of the image belong to each component. For this study, we wanted to see if we could transfer this task over to images of satellites to identify which parts of the images were the different components of the satellite such as bus, solar panel, antenna, payload, etc. This can be very helpful for a number of applications including providing health and status of a satellite, or understanding its behavior.

With the pristine images, this might not be too hard. By eye, we can roughly produce a segmentation, but with turbulent images, this becomes much, much harder. And for either turbulent or pristine images, semantics segmentation becomes very tedious with large volumes of images. However, semantic segmentation of images in other fields such as biology and even day-to-day applications has seen revolutionary improvements through convolutional neural networks, a machine learning technique. So we wanted to apply CNN to semantic segmentation of satellites as well.

So now that we know our starting point and our end goal, how did we get from point A (a bunch of images of satellites) to point B (machine learning model that could semantically segment them)?

AFRL

Semantic Segmentation

This image is CC0 public domain

THE AIR FORCE RESEARCH LABORATORY

Approved for public release: distribution unlimited.



Experiment Outline

- **Scenario A: Single Satellite, No Turbulence-** Established a baseline
- **Scenario B: Single Satellite, Single Turbulence Level**—Evaluate performance in turbulence
- **Scenario C: Single Satellite, Multi Turbulence Level**—Increase generalizability by training a model using the full range of turbulence levels
- **Scenario D: Reproducibility with Hubble**
- **Scenario E: Generalizing Poses**
- **Scenario F: Multiple Satellites**— Finally we attempted to create a model which can generalize across satellites

We broke the problem down into a few stages, starting with the simplest task and moving on to the most difficult. In the simplest task, we trained our machine learning model with pristine images of one satellite from different angles. This helped us establish a baseline of performance for how much the model could learn.

After that, we upped the ante a bit and added in turbulence, but kept it relatively simple for the model by training a model for each turbulence level to simply test whether the model could segment turbulent images.

For the next scenario, we gave the model a harder task, training it with images of the satellite at all the different turbulence levels. The goal was to create a model that could generalize across the spectrum of turbulence.

And as a bit of an intermission, we verified our results that we saw with our first satellite by running the same tests with HST.

We also verified that the machine learning models were learning to segment the images across any position, even a new position, rather than mapping to the "most similar" image from the training set.

Finally, we attempted to create a model which could generalize across all satellites.

So how did these experiments turn out? Did the models do a good job?

AFRL

Experiment Outline

- **Scenario A: Single Satellite, No Turbulence-** Established a baseline
- **Scenario B: Single Satellite, Single Turbulence Level**—Evaluate performance in turbulence
- **Scenario C: Single Satellite, Multi Turbulence Level**—Increase generalizability by training a model using the full range of turbulence levels
- **Scenario D: Reproducibility with Hubble**
- **Scenario E: Generalizing Poses**
- **Scenario F: Multiple Satellites**— Finally we attempted to create a model which can generalize across satellites



Network Metrics

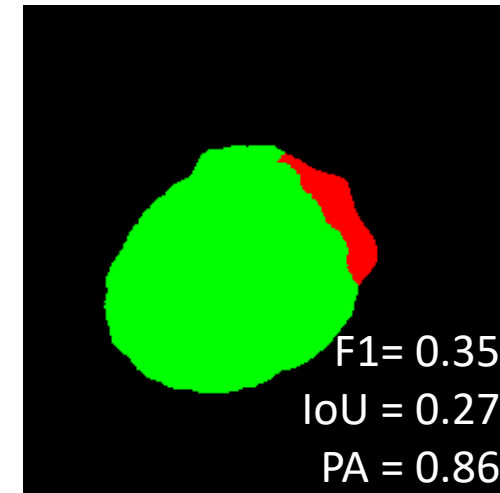
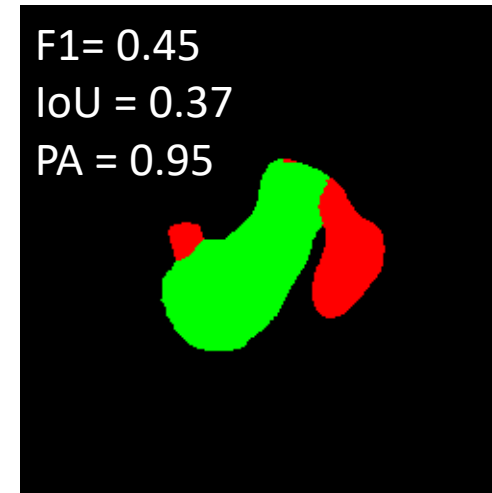
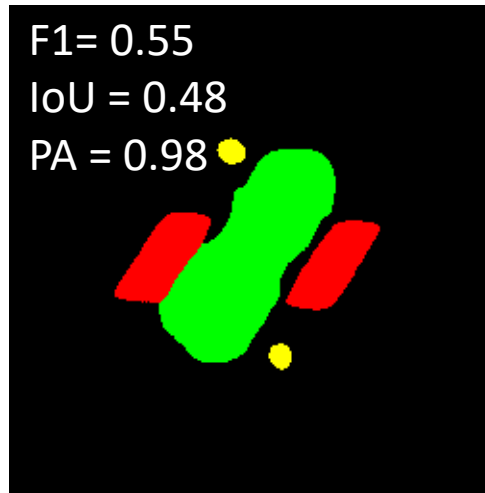
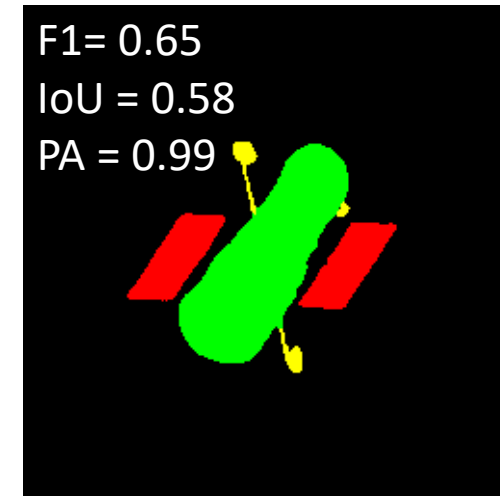
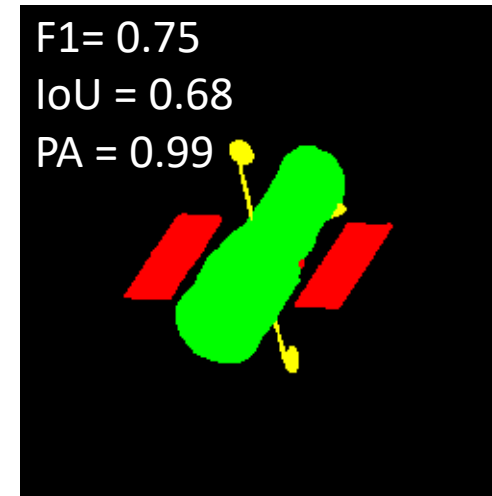
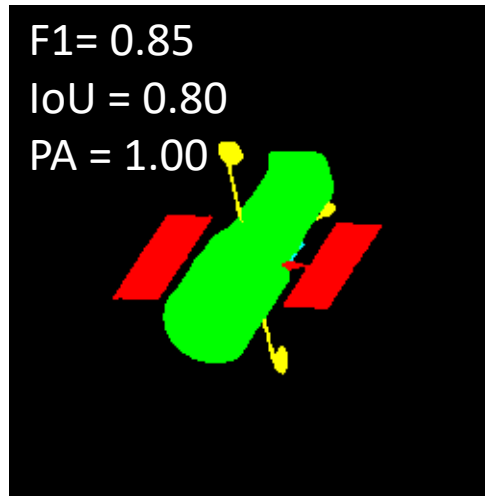
- F1 Score (Sørensen–Dice coefficient)

$$F_1 = \frac{2(\textit{precision} * \textit{recall} + \epsilon)}{\textit{precision} + \textit{recall} + \epsilon}$$

- Intersection over Union (IoU)

$$IoU = \frac{p_i * y_i + \epsilon}{p_i + y_i - p_i * y_i + \epsilon}$$

- Pixel Accuracy



For our study, we used three metrics to measure how well a model was performing at semantic segmentation. F1 score also known as the Dice coefficient, Intersection over Union (IoU) and Pixel Accuracy.

Network Metrics

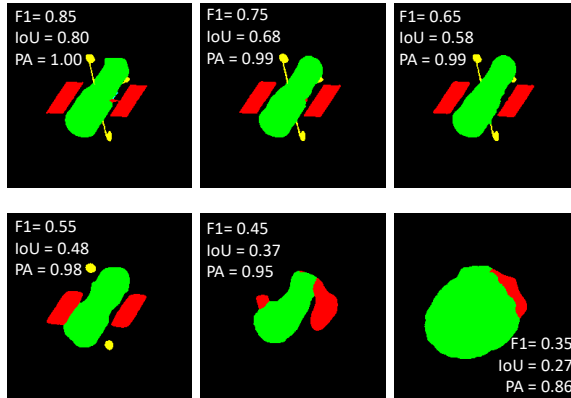
- F1 Score (Sørensen–Dice coefficient)

$$F_1 = \frac{2(\textit{precision} * \textit{recall} + \epsilon)}{\textit{precision} + \textit{recall} + \epsilon}$$

- Intersection over Union (IoU)

$$IoU = \frac{p_i * y_i + \epsilon}{p_i + y_i - p_i * y_i + \epsilon}$$

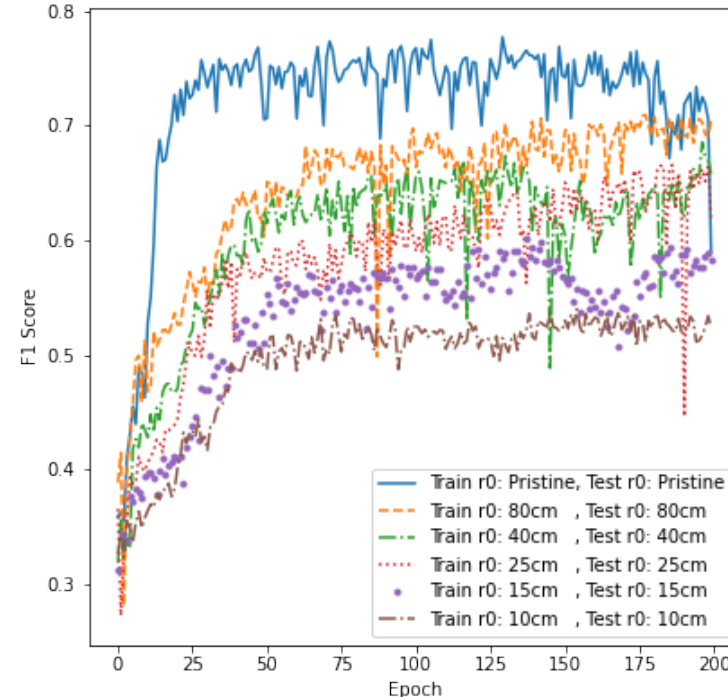
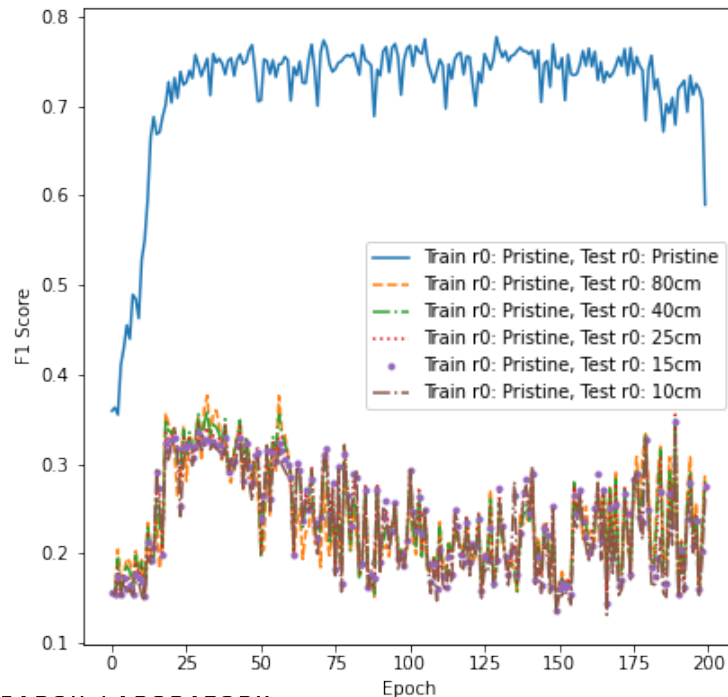
- Pixel Accuracy





Scenario A/B: Single Turbulence Levels

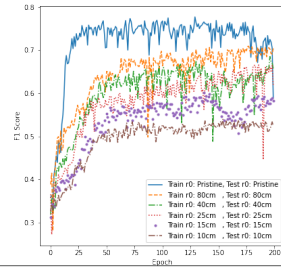
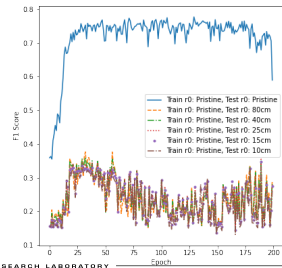
		Training r_0 (cm)					
		Pristine	80	40	25	15	10
Test r_0 (cm)	Pristine	0.77, 0.73, 1.00	0.38, 0.31, 0.96	0.32, 0.26, 0.95	0.33, 0.28, 0.94	0.23, 0.20, 0.93	0.29, 0.26, 0.93
	80	0.28, 0.23, 0.70	0.69, 0.62, 0.99	0.59, 0.51, 0.98	0.58, 0.50, 0.98	0.35, 0.28, 0.95	0.30, 0.24, 0.94
	40	0.31, 0.26, 0.70	0.59, 0.51, 0.99	0.62, 0.55, 0.99	0.64, 0.56, 0.98	0.43, 0.35, 0.96	0.33, 0.26, 0.94
	25	0.31, 0.28, 0.71	0.47, 0.40, 0.97	0.54, 0.47, 0.98	0.66, 0.60, 0.99	0.51, 0.43, 0.97	0.38, 0.30, 0.95
	15	0.31, 0.28, 0.72	0.36, 0.30, 0.95	0.39, 0.33, 0.96	0.53, 0.44, 0.97	0.57, 0.50, 0.99	0.48, 0.40, 0.97
	10	0.30, 0.27, 0.71	0.30, 0.24, 0.93	0.31, 0.25, 0.92	0.39, 0.31, 0.94	0.44, 0.37, 0.96	0.53, 0.46, 0.98



AFRL

Scenario A/B: Single Turbulence Levels

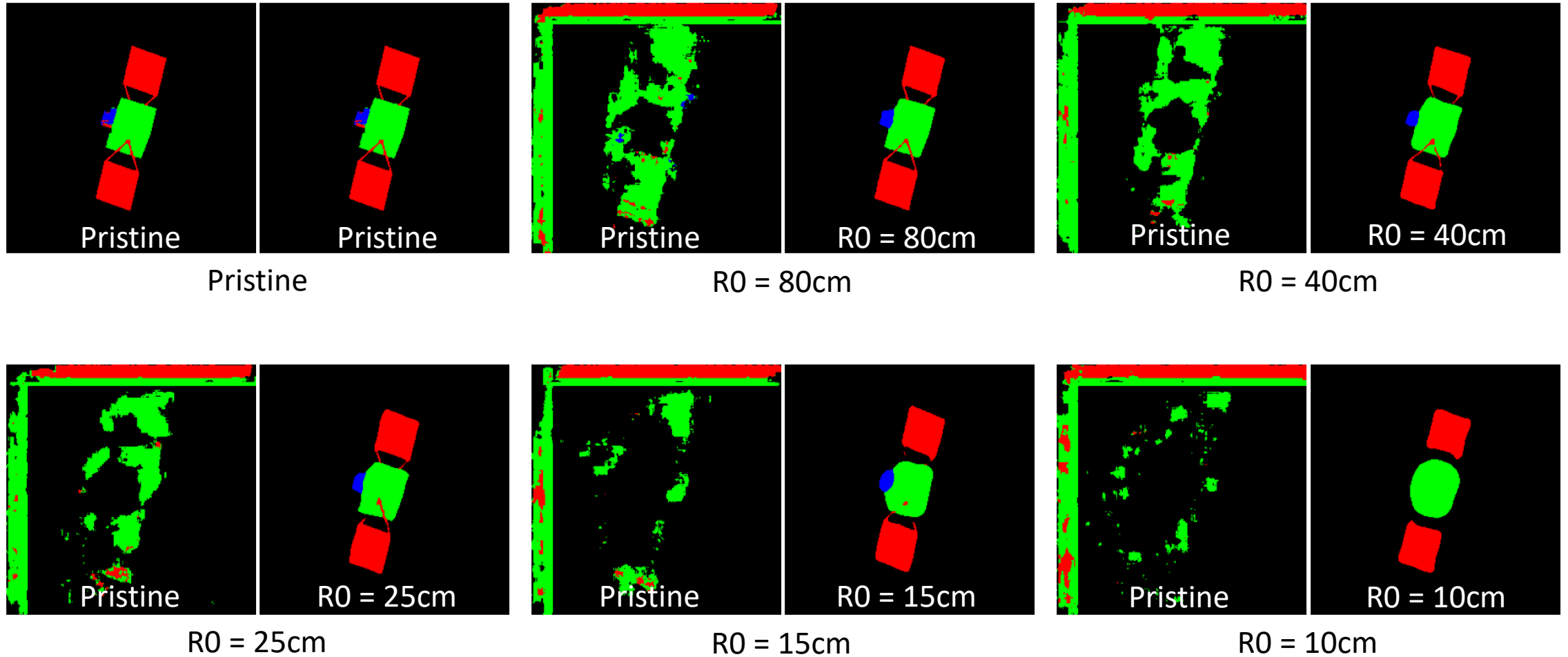
Test r_0 (cm)	Training r_0 (cm)					
	Pristine	80	40	25	15	10
Pristine	0.77, 0.73, 1.00	0.38, 0.31, 0.96	0.32, 0.26, 0.95	0.33, 0.28, 0.94	0.23, 0.20, 0.93	0.29, 0.26, 0.93
80	0.28, 0.23, 0.70	0.69, 0.62, 0.99	0.59, 0.51, 0.98	0.58, 0.50, 0.98	0.35, 0.28, 0.95	0.30, 0.24, 0.94
40	0.31, 0.26, 0.70	0.59, 0.51, 0.99	0.62, 0.55, 0.99	<i>0.64, 0.56, 0.98</i>	0.43, 0.35, 0.96	0.33, 0.26, 0.94
25	0.31, 0.28, 0.71	0.47, 0.40, 0.97	0.54, 0.47, 0.98	0.66, 0.60, 0.99	0.51, 0.43, 0.97	0.38, 0.30, 0.95
15	0.31, 0.28, 0.72	0.36, 0.30, 0.95	0.39, 0.33, 0.96	0.53, 0.44, 0.97	0.57, 0.50, 0.99	0.48, 0.40, 0.97
10	0.30, 0.27, 0.71	0.30, 0.24, 0.93	0.31, 0.25, 0.92	0.39, 0.31, 0.94	0.44, 0.37, 0.96	0.53, 0.46, 0.98



THE AIR FORCE RESEARCH LABORATORY

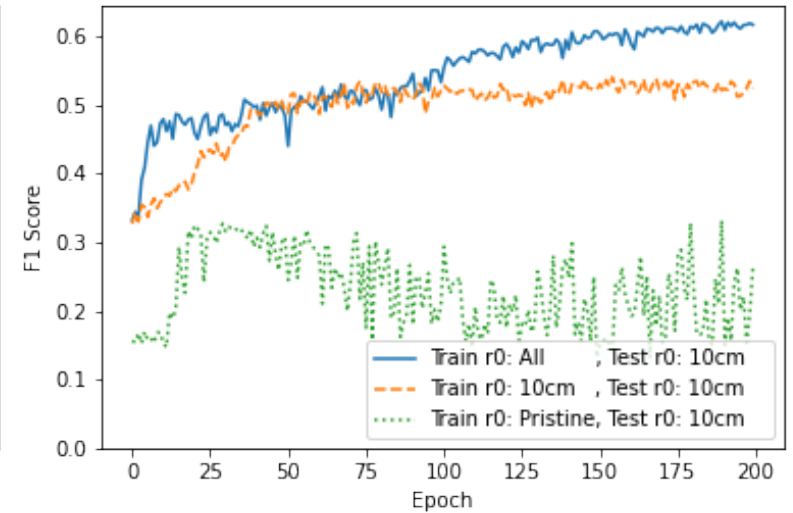
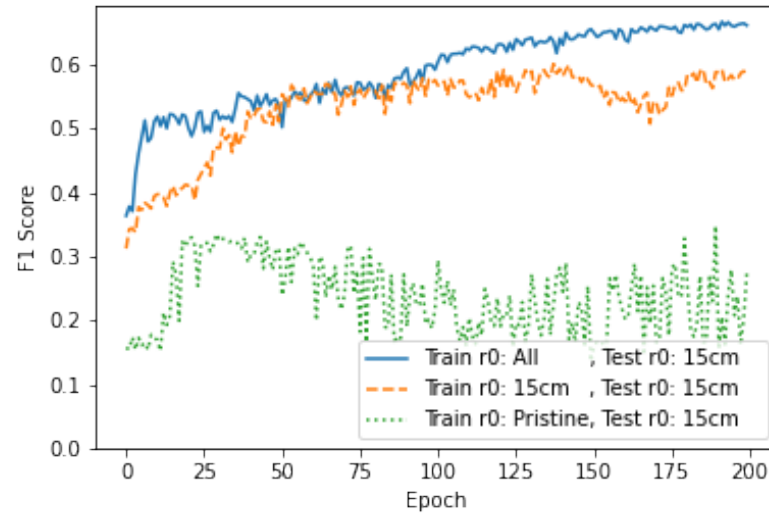
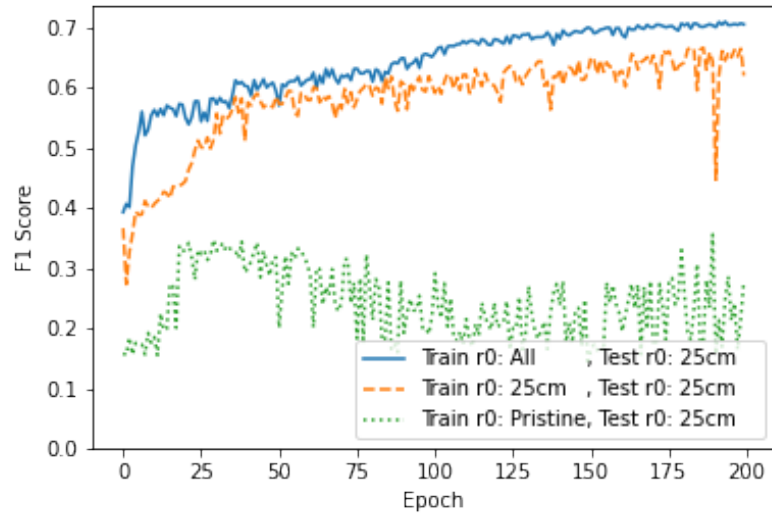
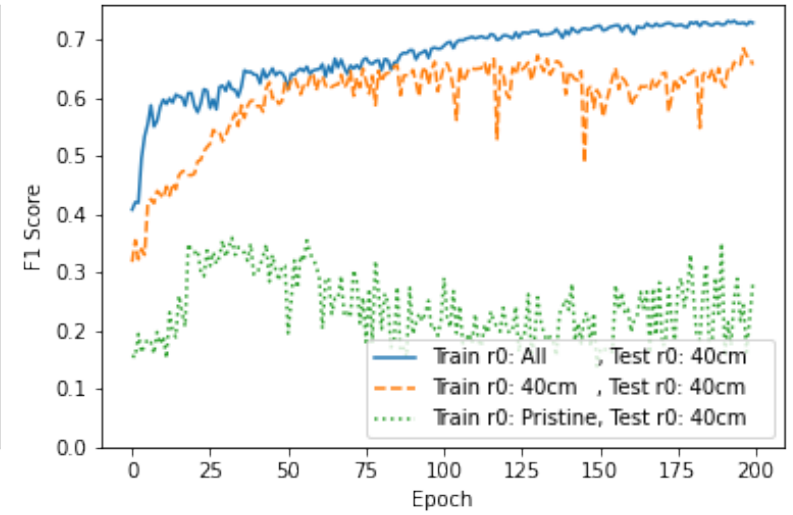
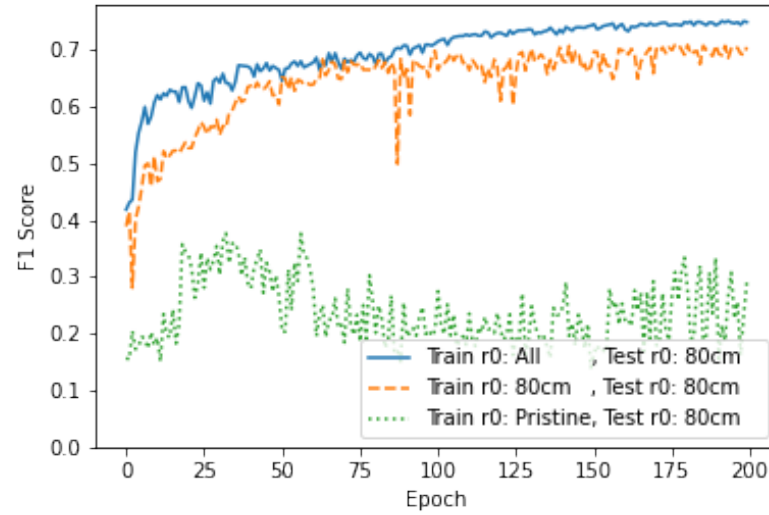
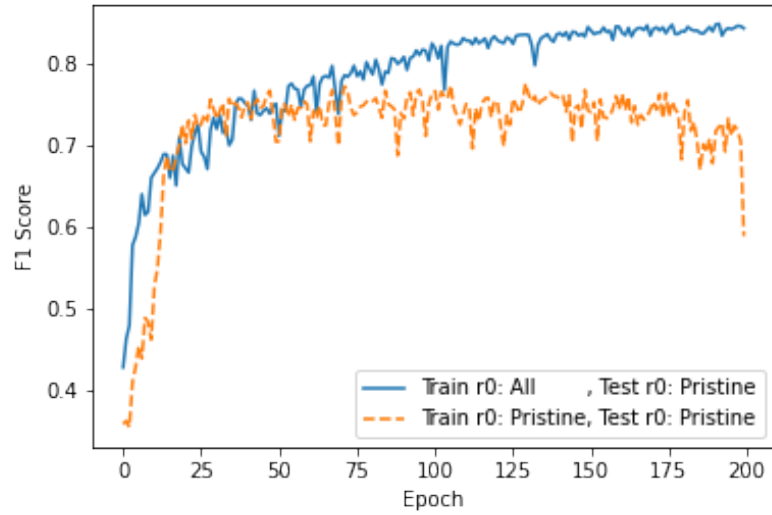
Approved for public release; distribution unlimited.

Scenario A





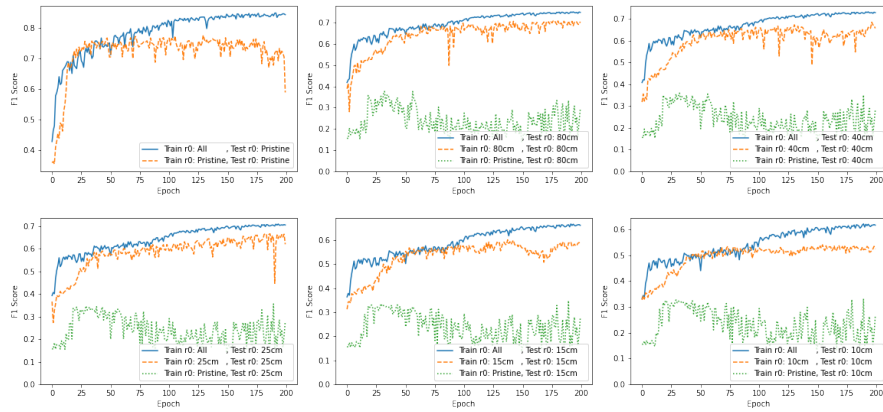
Scenario C: All turbulence Levels



The model trained on all turbulence levels consistently performs better than the model trained at single turbulence level. Furthermore, the models trained on multiple turbulence levels continued to learn even at the later epochs while the models trained on single turbulence levels leveled off towards the latter half of training. This indicates that the model trained on all turbulence levels could have seen even better performance if it had been trained with more epochs or a higher learning rate. The model trained on single turbulence levels could have also benefited from further regularization.

AFRL

Scenario C: All turbulence Levels

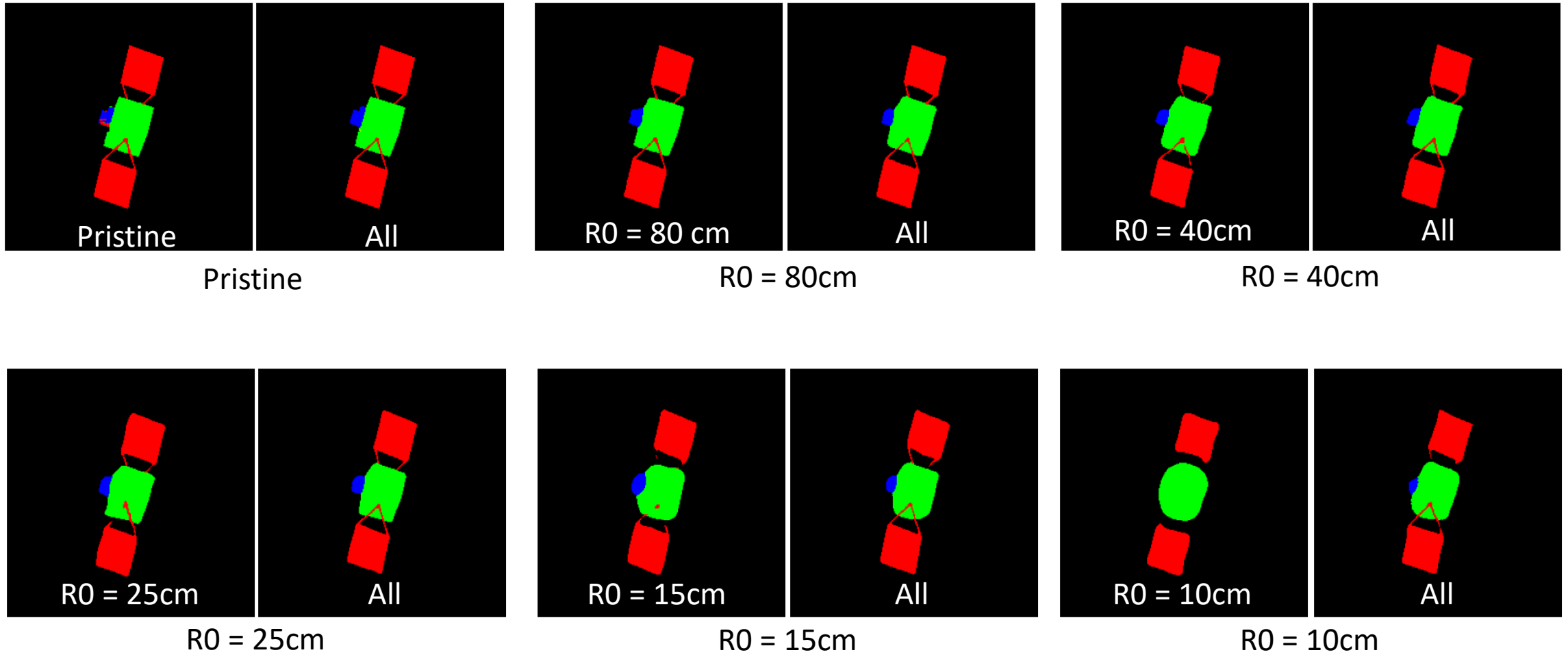


THE AIR FORCE RESEARCH LABORATORY

Approved for public release; distribution unlimited.

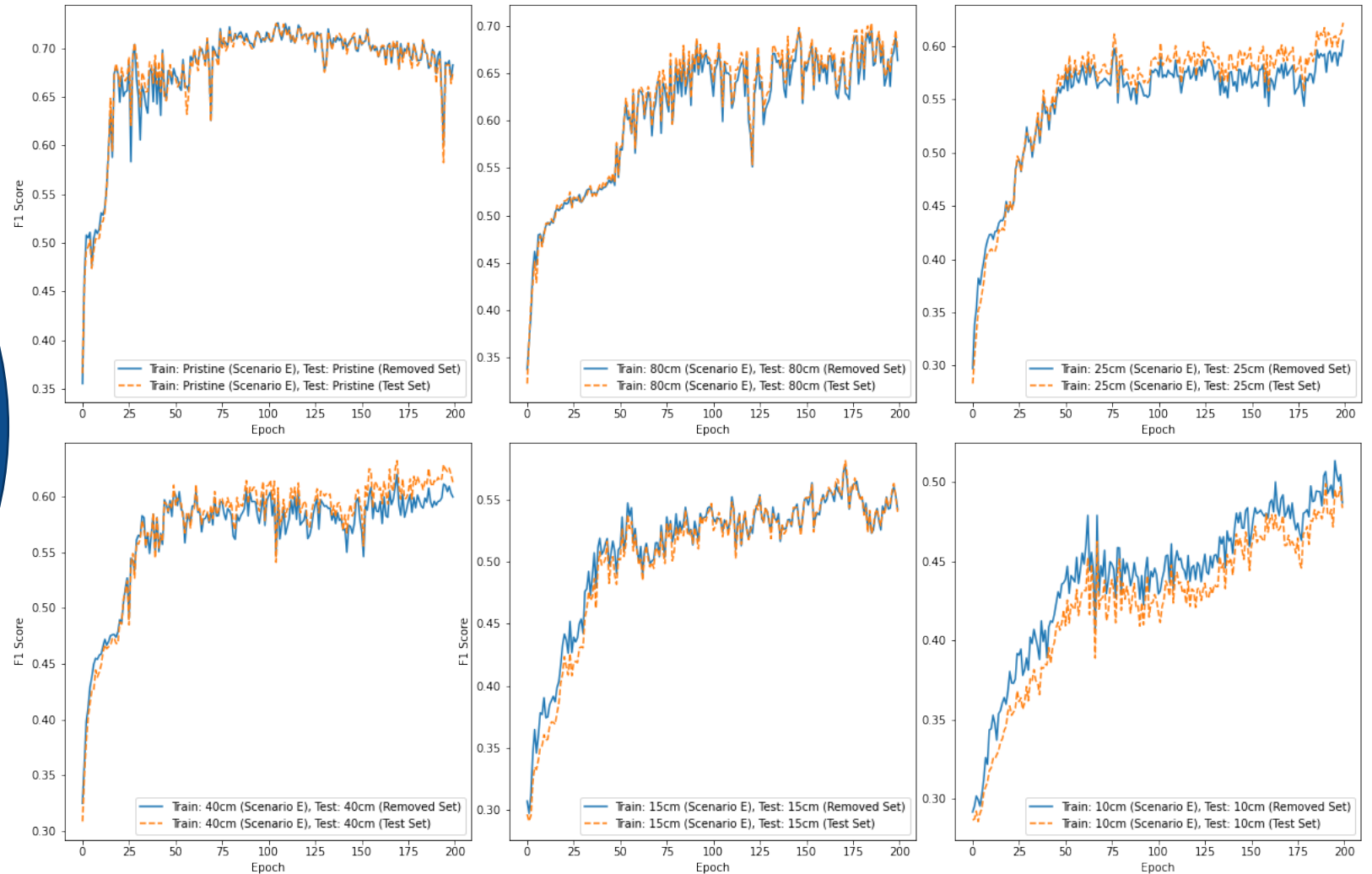
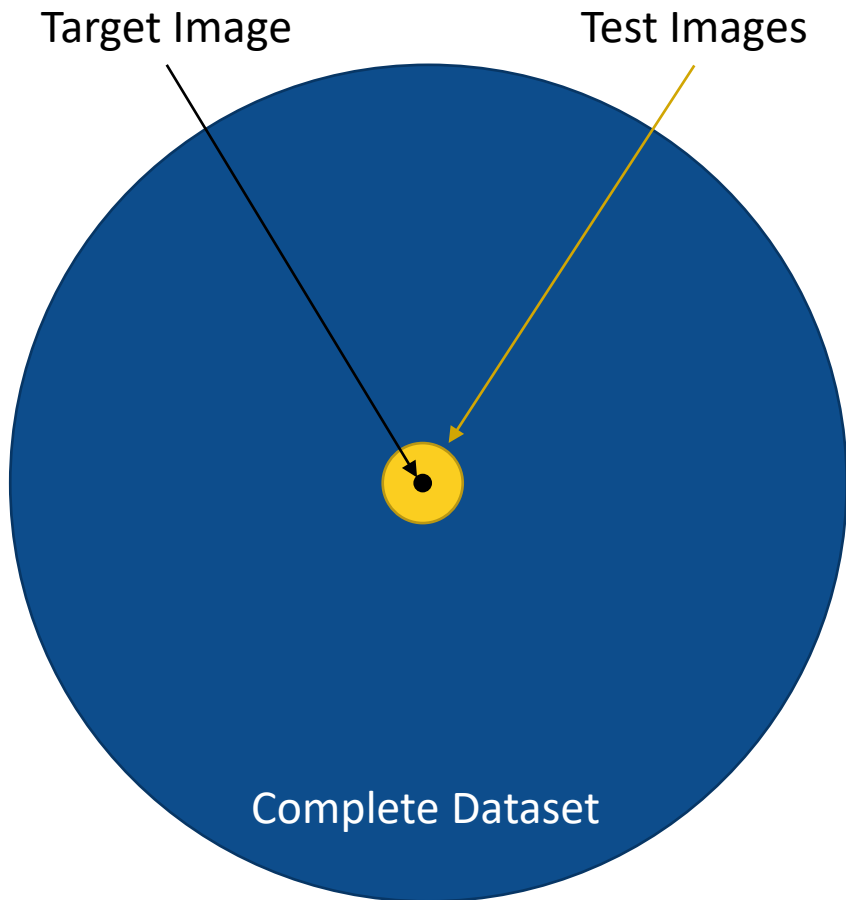
9

Scenario B vs C: Trained on single turbulence level vs All





Scenario E: Generalizing Poses



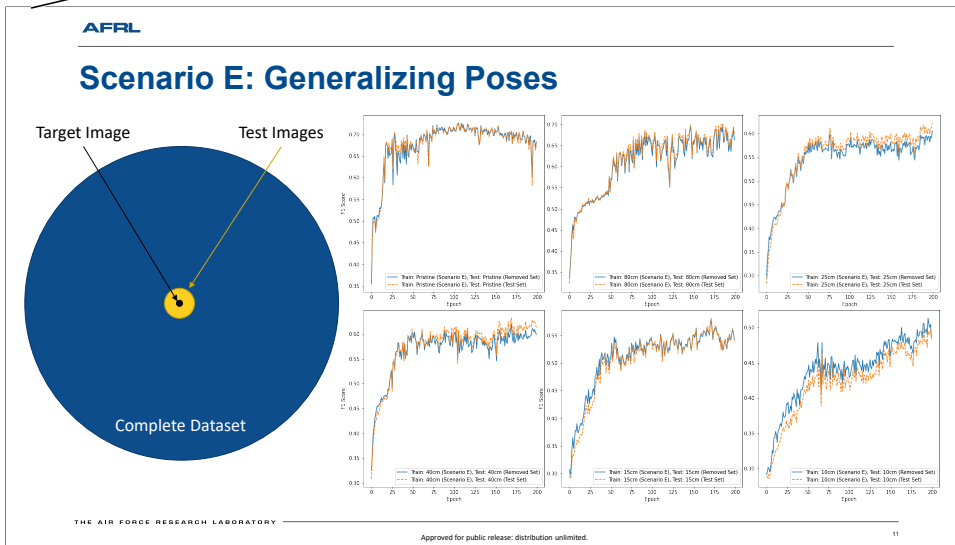
Ok so, we can segment across turbulence levels, but because our training data was randomly selected from a uniform distribution of viewing angles, we wanted to make sure that the model was for sure learning how to segment these images of the satellite we were giving it rather than just creating a giant lookup table to the closest image in the training set.

So what we did was pick a "target image" in the set of images, and find the 200 most similar images to that image from the dataset (each dataset consisted of approximately 10,000 images from a uniform distribution of discrete viewing angles). And we measured "similarity between images" using our standard F1 metric.

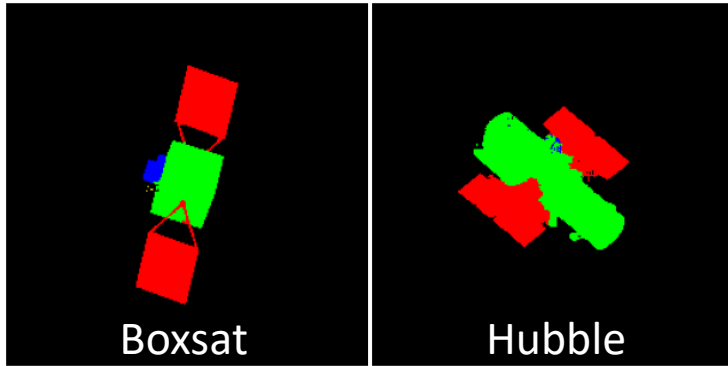
We held out these 200 images, and trained the model on the remaining images. In other words, the model was only able to train off images that were very different from the target test image. It was the same satellite and the same turbulence, but different poses.

We then tested on the trained model on these 200 held out images, and the model was able to semantically segment the images at effectively the same performance as the models trained on the completely random selection.

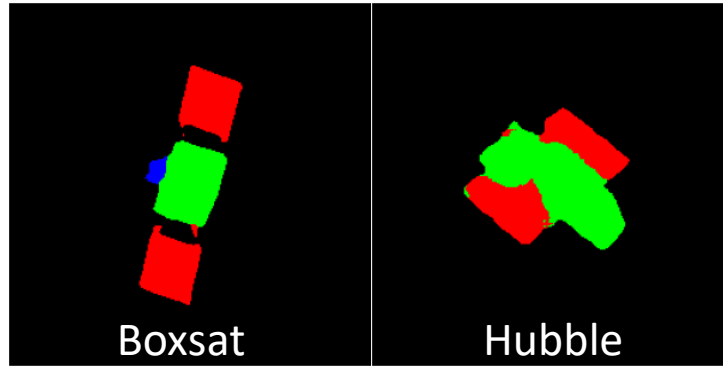
As you can see in the plots on the right, the F1 scores being plotted for each turbulence level is nearly identical for the two lines. The performance is nearly identical, so we can conclude that the model isn't just looking for the most similar image from the test dataset, it's actually segmenting the images.



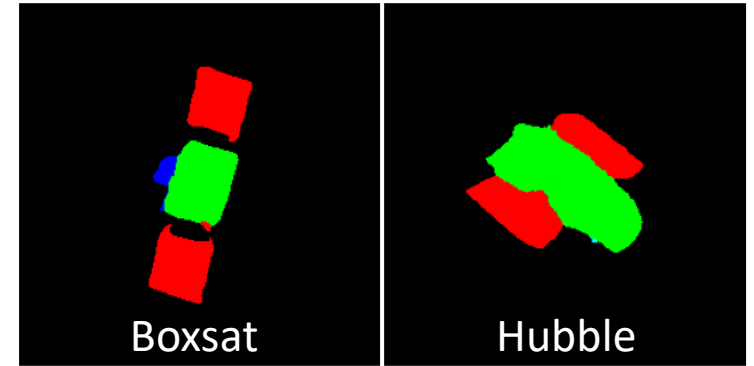
Scenario F: Multiple Satellites (Boxsat vs Hubble)



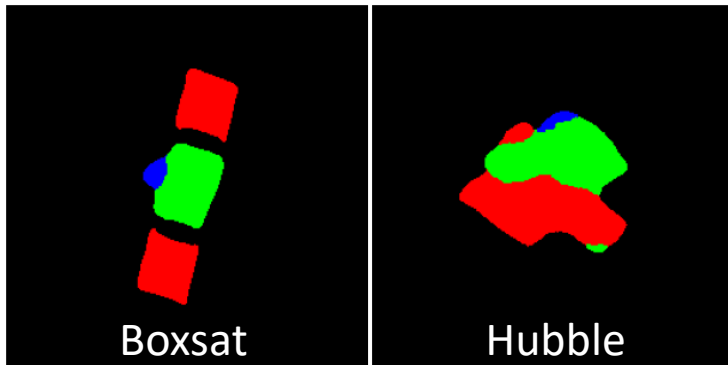
Pristine



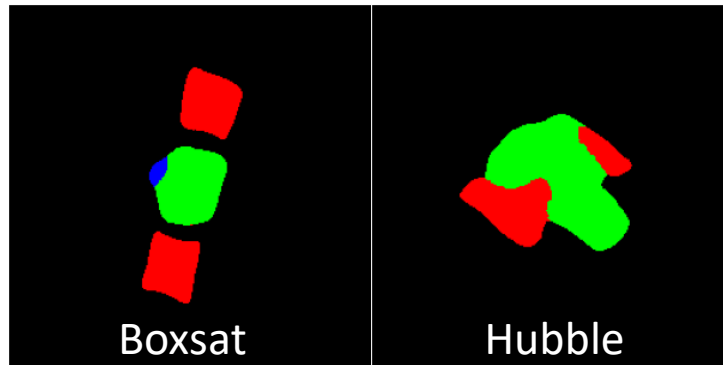
R0 = 80cm



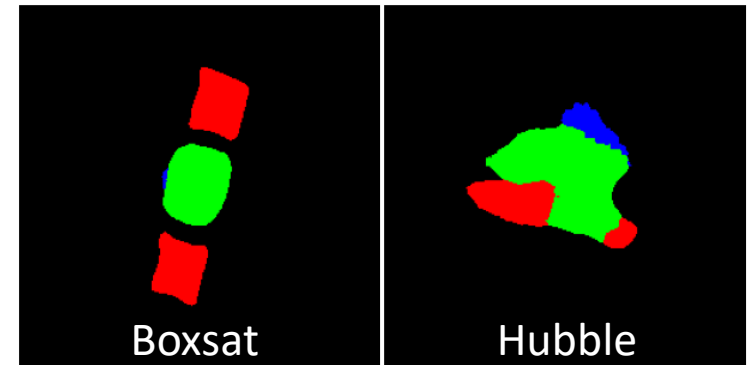
R0 = 40cm



R0 = 25cm

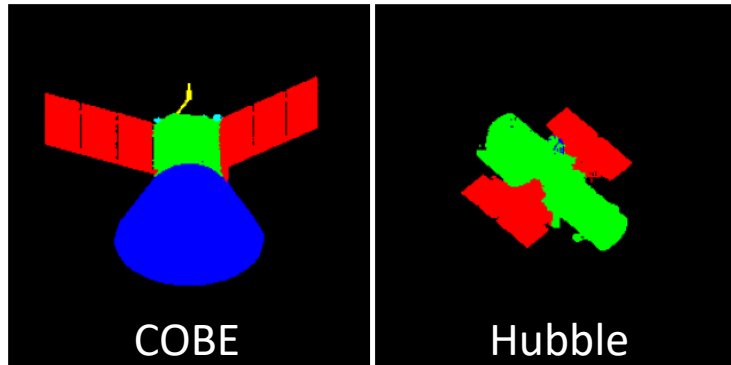


R0 = 15cm

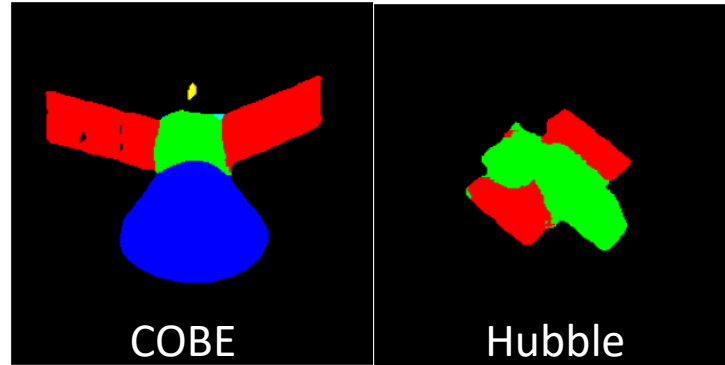


R0 = 10cm

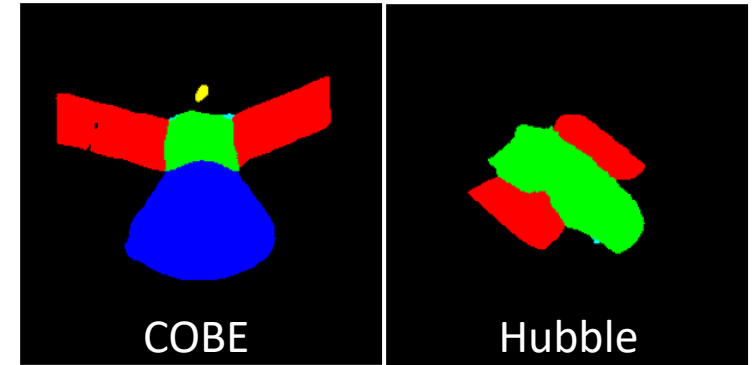
Scenario F: Multiple Satellites (COBE vs Hubble)



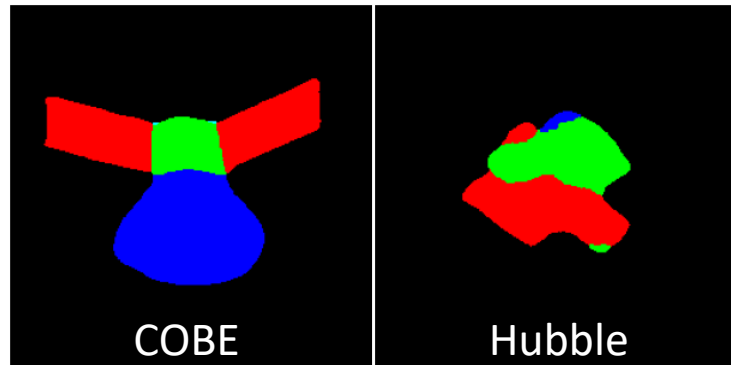
Pristine



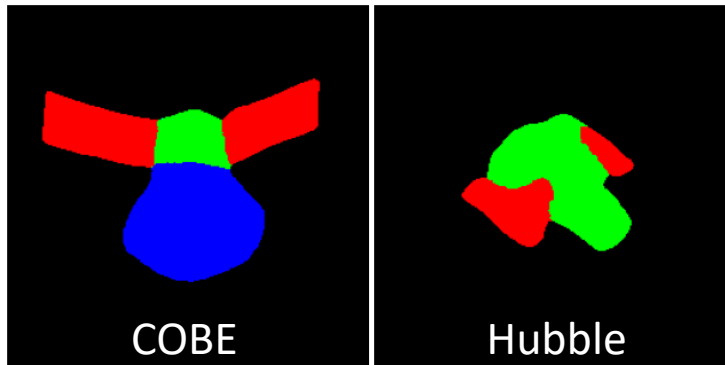
R0 = 80cm



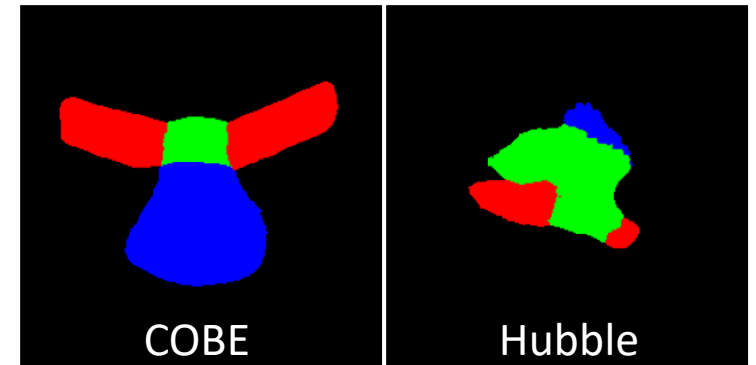
R0 = 40cm



R0 = 25cm



R0 = 15cm



R0 = 10cm

Scenario F: Multiple Satellites

		Test r_0 (cm)					
		Pristine		80		40	
		All	Hub.	All	Hub.	All	Hub.
Training r_0	Pristine	0.67, 0.64, 1.00	0.45, 0.39, 0.97	0.11, 0.08, 0.36	0.10, 0.07, 0.40	0.10, 0.07, 0.34	0.09, 0.07, 0.37
	80cm	0.36, 0.30, 0.93	0.27, 0.23, 0.93	0.60, 0.55, 0.99	0.41, 0.36, 0.96	0.54, 0.48, 0.97	0.39, 0.33, 0.95
	40cm	0.28, 0.22, 0.85	0.22, 0.18, 0.87	0.55, 0.50, 0.98	0.37, 0.32, 0.96	0.58, 0.53, 0.99	0.39, 0.33, 0.96
	25cm	0.27, 0.22, 0.89	0.18, 0.16, 0.90	0.45, 0.39, 0.95	0.31, 0.26, 0.94	0.52, 0.46, 0.97	0.35, 0.30, 0.95
	15cm	0.20, 0.16, 0.84	0.17, 0.15, 0.89	0.26, 0.21, 0.89	0.19, 0.17, 0.90	0.37, 0.30, 0.93	0.24, 0.21, 0.92
	10cm	0.21, 0.18, 0.87	0.18, 0.16, 0.90	0.19, 0.16, 0.87	0.16, 0.15, 0.90	0.22, 0.18, 0.88	0.17, 0.16, 0.90
		25		15		10	
		All	Hub.	All	Hub.	All	Hub.
Training r_0	Pristine	0.10, 0.06, 0.31	0.09, 0.06, 0.33	0.08, 0.05, 0.25	0.07, 0.04, 0.22	0.07, 0.04, 0.16	0.05, 0.03, 0.11
	80cm	0.44, 0.38, 0.94	0.32, 0.27, 0.91	0.29, 0.23, 0.83	0.25, 0.21, 0.85	0.21, 0.17, 0.77	0.21, 0.18, 0.83
	40cm	0.52, 0.46, 0.97	0.36, 0.30, 0.94	0.36, 0.30, 0.89	0.29, 0.24, 0.89	0.27, 0.21, 0.81	0.25, 0.20, 0.84
	25cm	0.55, 0.50, 0.98	0.37, 0.31, 0.95	0.45, 0.39, 0.94	0.32, 0.27, 0.92	0.31, 0.25, 0.84	0.27, 0.22, 0.86
	15cm	0.45, 0.39, 0.95	0.30, 0.25, 0.93	0.52, 0.47, 0.98	0.32, 0.27, 0.94	0.41, 0.34, 0.92	0.29, 0.24, 0.90
	10cm	0.28, 0.23, 0.90	0.19, 0.17, 0.90	0.41, 0.34, 0.94	0.25, 0.21, 0.92	0.49, 0.44, 0.97	0.29, 0.24, 0.92

Conclusion

- The overall performance of these experiments has shown that semantic segmentation of ground-based images of LEO satellites through turbulence is viable using a convolutional neural network approach.
- We have shown that a U-Net approach for this semantic segmentation task can segment images through a wide range of atmospheric turbulence levels.
- We have also shown that the network can segment multiple different satellites.
- Future work:
 - Incorporate procedural satellite generation for better regularization and avoid overfitting to just the satellites within the training dataset
 - Unique conditions such as glints, smear, and jitter
 - Real images

Acknowledgements

The authors would like to thank the Air Force Research Laboratory (AFRL) for supporting this effort. The views, opinions, and/or findings expressed here are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Q/A

Backup

CNN Architecture: U-Net

