# Automated 6DOF Satellite Pose Estimation from Resolved Ground-Based Imagery

**Thomas Dickinson, Derek Walvoord, Michael Gartley**
*Chester F. Carlson Center for Imaging Science, Rochester, New York, United States*

## ABSTRACT

Automated satellite pose estimation enhances spacecraft health assessment and behavior monitoring and is a valuable part of future Space Domain Awareness (SDA) architectures. Ground-based adaptive optics (AO) telescopes are underemployed for pose estimation due to limited image quality and current labor-intensive manual pose estimation processes. Learning-based pose estimation techniques excel using high-quality imagery and real training data but perform poorly when trained on simulated imagery and tested on real imagery – known as the Sim2Real domain gap. To overcome this limitation, a methodology was developed to simulate large amounts of realistic AO imagery, taking advantage of recent innovations in computer vision to regress 6 degrees of freedom (6DOF) pose. We propose a three-stage approach that first localizes a satellite, then estimates 6DOF pose, and finally updates the pose with temporal Kalman filtering. For the first two model stages tested on simulated 6DOF imagery we demonstrate 13.2° of mean rotational error (2.7° median), 31 cm of mean 2-axis translation error (18 cm median), and 1.4% or 13.7 km mean slant range error (1.2% median). Mean translation error can be expressed angularly as 300 nrad (170 nrad median). The model – entirely trained on synthetic data – achieves 4° of mean rotational error on 105 frames of real imagery, successfully bridging the domain gap.

## 1. INTRODUCTION

Automated satellite pose estimation has the potential to significantly enhance the assessment of spacecraft health and pattern-of-life, thereby becoming an integral component of future Space Domain Awareness (SDA) architectures. Large, ground-based electro-optical (EO) telescopes equipped with adaptive optics (AO) produce well-resolved imagery of Low Earth Orbit (LEO) satellites, but this imagery is rarely employed for pose estimation due to difficult visual interpretation and labor-intensive processes [1]. In recent years, the performance of learning-based approaches for automated pose estimation has surpassed that of conventional approaches [2, 3]. However, sufficient real, labeled training data is currently nonexistant and is challenging to create, and learning methods suffer when trained on simulated data for real applications. Our research focuses on developing a methodology to simulate realistic AO-compensated satellite imagery and train robust deep learning models, facilitating automated, real-time satellite pose estimation and bridging the simulation-to-real (Sim2Real) domain gap [2, 4]. This methodology would provide the ability to accurately track a satellite's position and orientation in real time, enhancing SDA. This work assumes a single instance of a single rigid object and an accurate, rigid CAD model for that object.

## 2. PRISTINE 3DOF POSE ESTIMATION

First, a training set of 200,000 pristine (zero blur or noise) and fully illuminated 32×32 pixel (780 nrad Instantaneous Field of View (IFOV)) images was rendered using Blender and a detailed Hubble Space Telescope satellite CAD model [5]. For comparison, the Rayleigh resolution for a 3.5-m telescope is 287 nrad at 825 nm. Random 3 degrees of freedom (3DOF) rotations were applied to the CAD model for each render, as shown in Fig. 1.

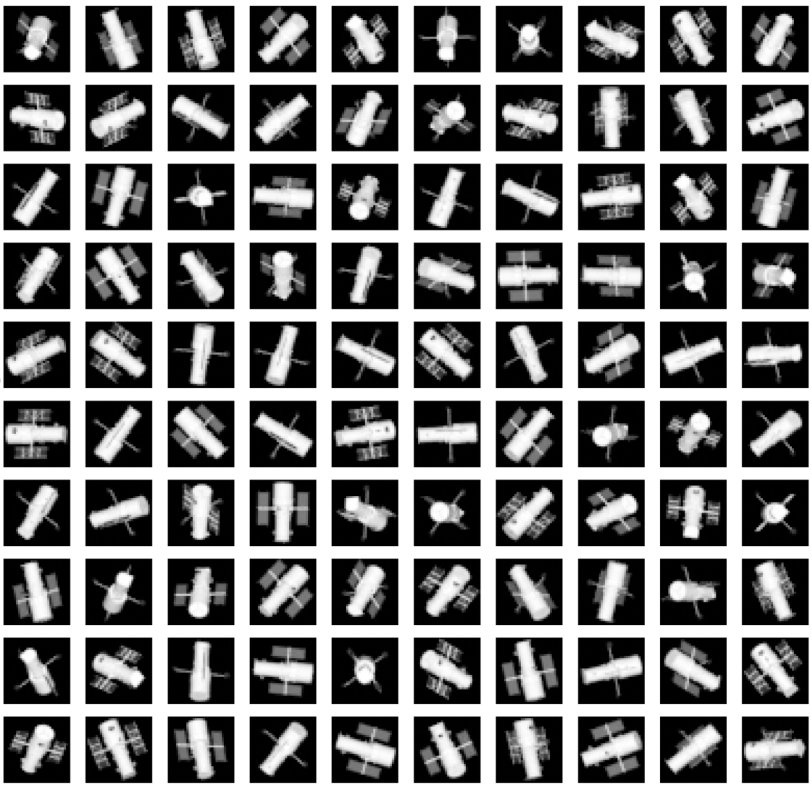

Fig. 1: The first 100 test images from the pristine, 3DOF, fully illuminated Hubble Space Telescope dataset showing random rotations drawn from a uniform spherical distribution.

The 6-dimensional (6D) representation for 3DOF rotation was used for training labels [6]. The mathematical details of this representation and its advantages are covered in [3, 6, 7]. A Convolutional Neural Network (CNN) with 4M trainable parameters was designed with Gram-Schmidt orthogonalization as the final layer [6]. The pose model was trained to directly regress 3DOF satellite pose from a single image, achieving 1.05° of mean quaternion distance rotational error on the test set as displayed in Fig. 2. This dataset and experiment are similar to the pristine data tested in [1], with the 6D representation and Gram-Schmidt orthogonalization resulting in reduced error despite the model using fewer than 20% of the parameters. In this work the quaternion distance rotational error $e_q$ was calculated as in Section 4.1 of [2], where

$$e_q = 2 \arccos\left(\left|\langle \hat{\boldsymbol{q}}, \boldsymbol{q} \rangle\right|\right) \tag{1}$$

and the mean rotational error $E_q$ is given by

$$E_q = \frac{1}{N} \sum_{i=1}^{N} e_q^{(i)}. \tag{2}$$

The rotational error is the minimum rotation required to align ground truth rotation $\hat{\boldsymbol{q}}$ and predicted rotation $\boldsymbol{q}$.

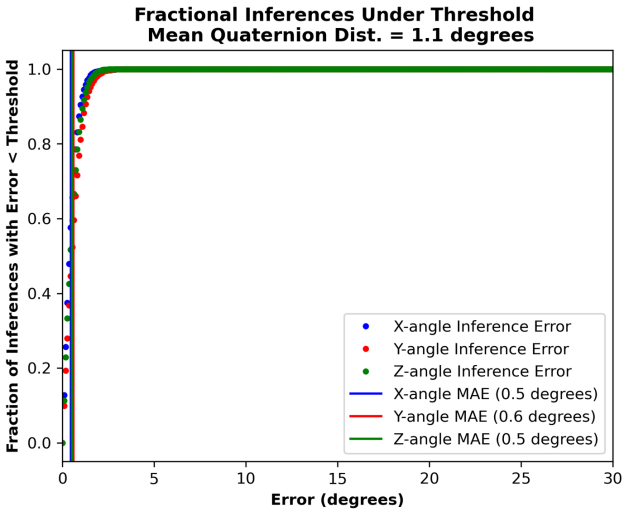


Fig. 2: Fractional inferences with error under different thresholds for the pristine, fully illuminated model. Vertical bars show the Mean Absolute Error (MAE) for each rotational axis. The mean rotational error for this network and dataset was 1.05°. Figure design based on [1].

## 3. PARTIALLY ILLUMINATED 3DOF POSE ESTIMATION

In real application, illumination is partial and variable since these SDA systems typically operate during terminator conditions. Additionally, shadows are extremely deep due to the lack of atmosphere and surroundings to scatter light onto parts of the object that are not directly illuminated, as shown in Fig. 4. Therefore, a training set of 410,000 pristine, 32×32 pixel, partially illuminated images was constructed with random illumination angles restricted to 25% of the celestial sphere in the fixed camera's coordinate frame. A higher capacity neural network was required to handle the partial and varying illumination. The network contained 17M trainable parameters and was trained from scratch in intermediate steps. The initial stage employed a pose classifier CNN, a sparse keypoint feature localizer CNN, and a latent representation of the input image produced by the encoder stripped from a pretrained Variational Autoencoder. The initial stage outputs were concatenated and fed into fully connected layers, then a final Gram-Schmidt layer. A mean rotational error of 1.8° (median = 1.5°) was achieved on the test set, as shown in Fig. 3.

An experiment was conducted to compare the efficacy of a learning-based method with a conventional computer vision approach. For the former, the sparse keypoint feature localizer CNN was connected to fully connected layers and the final Gram-Schmidt layer, totaling 8.7M trainable parameters. This model achieved a mean rotational error of 5.5° on a 6,000-sample test set. For conventional comparison, the same keypoint feature localizer was used to regress 40 keypoints, and the sparse 2D-3D keypoint correspondences were processed with Perspective-*n*-Point (P*n*P) and random sample consensus (RANSAC) algorithms to calculate 3DOF rotational pose. This conventional method produced a mean rotational error of 7.2°. The results demonstrated that the fully learning-based approach outperformed the indirect method for this specific application and implementation. These findings align with [3] which showed that while P*n*P/RANSAC can be more accurate than a learned approach with unrealistically accurate keypoint localizations, the learned approach is more robust to noise and outperforms P*n*P/RANSAC when dealing with noisy keypoint localizations.

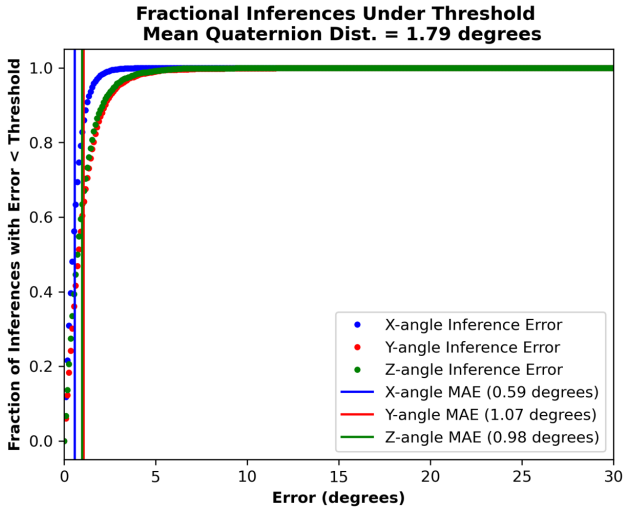


Fig. 3: Fractional inferences with error under different thresholds for the pristine, partially illuminated model. Vertical bars show the Mean Absolute Error (MAE) for each rotational axis. The mean rotational error for this network and dataset was 1.79°. Figure design based on [1].

## 4. 6DOF POSE ESTIMATION

Ultimately, practical relevance demands a pose model capable of 6 degrees of freedom (6DOF) pose estimation – 3DOF rotation, 2DOF translation, and 1DOF universal scale – since satellites will not be perfectly centered on the telescope's focal plane array with fixed scale. Additionally, the model must handle the challenges of partial illumination along with blur and low signal-to-noise ratio, as shown in Fig. 4. A model inspired by [3] and [8] was designed for this purpose. A Seasat (SATCAT 10967) CAD model was used to render 110,000 unique 6DOF poses and illuminations (256×256 pixel images, IFOV = 140 nrad), with an 80%/10%/10% split for training, validation, and testing. Illumination conditions were chosen randomly and constrained to be physically valid for a full year of terminator conditions, for Seasat passes over a given ground site. Random double-Gaussian blur, noise, bias, gamma, contrast, brightness, random Gaussian glints, and coarse dropout were used to degrade the pristine renders and create an augmented training set representative of real imagery using [9] based on [4]. This random augmentation tripled the number of images in the training set, resulting in 274,000 images but only 88,000 unique poses and illuminations. The same random process was used to degrade the validation set and test set, which each contained 11,000 images (all with unique poses and illuminations).

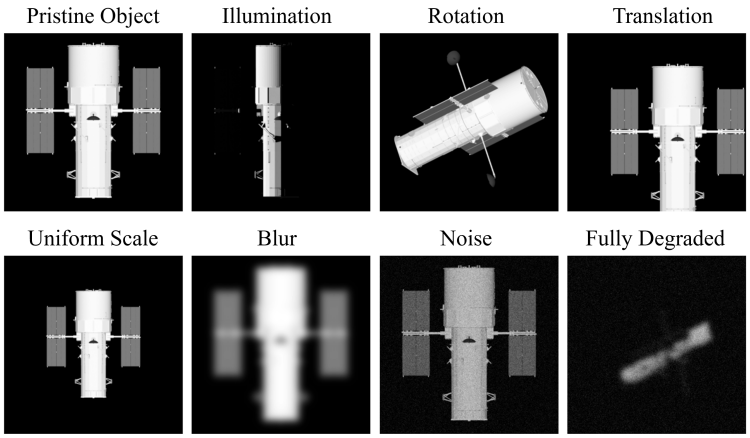


Fig. 4: Representation of partial illumination, 6DOF pose, and image degradations.

First, a coarse localizer network containing roughly 30M parameters was trained to regress amodal satellite bounding boxes and satellite center in image coordinates. The satellite center was pre-defined for the CAD model. Images (256×256 pixels) were fed into a U-Net to predict a binary amodal (fully illuminated) object mask. The original image and mask were then concatenated along the channel dimension and fed into two separate CNN prediction heads, with one regressing bounding boxes and the other regressing satellite center coordinates. The bounding box regressor achieved mean Intersection over Union (IoU) = 0.84 (median IoU = 0.86) for the degraded test set. The predicted bounding boxes and satellite center coordinates (shown in Fig. 5) were used to crop the initial images to the satellite region of interest (RoI) and resize them to 64×64 pixels for input to the 6DOF pose model.

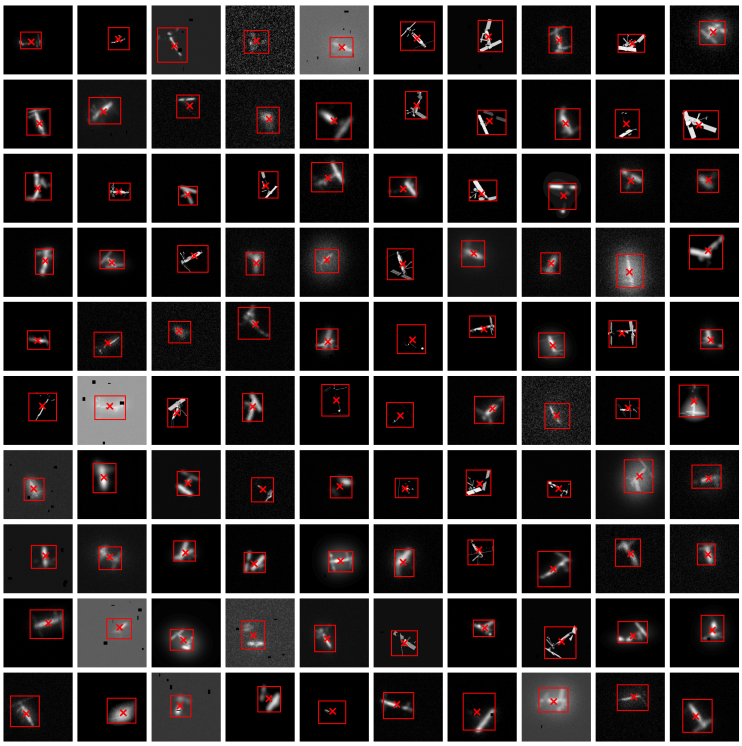

Fig. 5: The first 100 test images from the augmented 6DOF Seasat dataset, with bounding boxes and object centers predicted by the coarse localizer model shown in red.

This dynamic "zoom and crop" algorithm was designed around three rules: 1) the cropped RoI image must be square prior to resizing; 2) the RoI must contain the full bounding box; and 3) the predicted satellite center coordinates must be placed at the center of the RoI. Images were zero-padded when necessary to adhere to these rules. The intent of these the rules, respectively, was to 1) preserve aspect ratio; 2) avoid losing any information by cropping out parts of the object; and 3) minimize residual translation so the 6DOF pose model could primarily learn the 3DOF rotation of the object about the cropped image center. Reference [3] used an off-the-shelf localizer to regress only modal bounding boxes. It therefore lacks amodal bounding box or object center predictions, requiring a simpler dynamic zoom algorithm that may not minimize residual translation [3]. As with the Dynamic Zoom-In (DZI) algorithm in [3], the training process was decoupled from testing. However, our implementation was more closely aligned with [8]. During training, the target object's center coordinates and bounding box dimensions (width and height) were randomly perturbed using values drawn from a truncated normal distribution.

The 6DOF pose model totals 83M parameters and was trained from scratch in stages. Inspired by [3], U-Nets were used to regress Surface Region Attention Maps and normalized $X/Y/Z$ world coordinates corresponding to each image pixel resulting in dense 2D-3D correspondences. Finally, these correspondences and Surface Region Attention Maps were fed into a feature extractor CNN (analogous to P$n$P) to regress pose. Only the 3DOF rotation predictions were passed through the Gram-Schmidt layer. Additionally, the coarse localizer's predicted bounding box coordinates and object center coordinates were concatenated onto the first fully connected layer along with approximate scale

(ground truth scale +/- 2.5% drawn from a uniform random distribution). The intent was to provide the model with as much relevant information as possible to improve accuracy. In real-world application the approximate scale would be known, and any uncertainty would result from inexact slant range (computed from the two-line element set), sensor IFOV, and CAD model geometry. The Scale-Invariant representation for Translation Estimation (SITE) from [3, 8] was adapted for this application to accommodate orthographic projection and sensor-target range values ($Z$ world coordinate) approximately $10^6$ times larger than the $X$ and $Y$ world coordinate offsets from the sensor's optical axis (image center), and the 6D representation was used for ground truth rotation labels. The final training stage for the 6DOF pose model used the disentangled 6D pose loss from [3].

## 5. RESULTS AND CONCLUSION

Passing the full, uncropped test set first through the coarse localizer, then the dynamic zoom algorithm, and finally the 6DOF pose model, we demonstrated a mean rotational error of 13.2° (2.7° median), a mean 2-axis translation error of 31 cm (18 cm median), and a mean slant range error of 1.4% or 13.7 km (1.2% or 11.8 km median). Translation error can also be expressed angularly as a mean of 300 nrad (170 nrad median) or in pixel space (relative to the uncropped images) as a mean of 2.1 pixels (1.2 pixels median). The significantly lower median errors compared to the mean errors were due to long-tailed error distributions, as shown in Fig. 6. These distributions likely resulted from the random image degradations used in the augmented dataset, which occasionally combined to produce poor image quality, resulting in large pose prediction errors. Pose estimates are shown overlaid on test images in Fig. 7. When testing the second stage (6DOF pose model) independently, RoI-cropped input images produced by decoupled random draws from truncated normal distributions (as described previously), improved performance to a mean rotational error of 11.3° (2.7° median), a mean 2-axis translation error of 24 cm (18 cm median), and a mean slant range error of 1.3% or 13.2 km (1.2% or 11.8 km median). Angular translation error improved to a mean of 230 nrad (170 nrad median) and pixel space translation error to a mean of 1.7 pixels (1.2 pixels median).

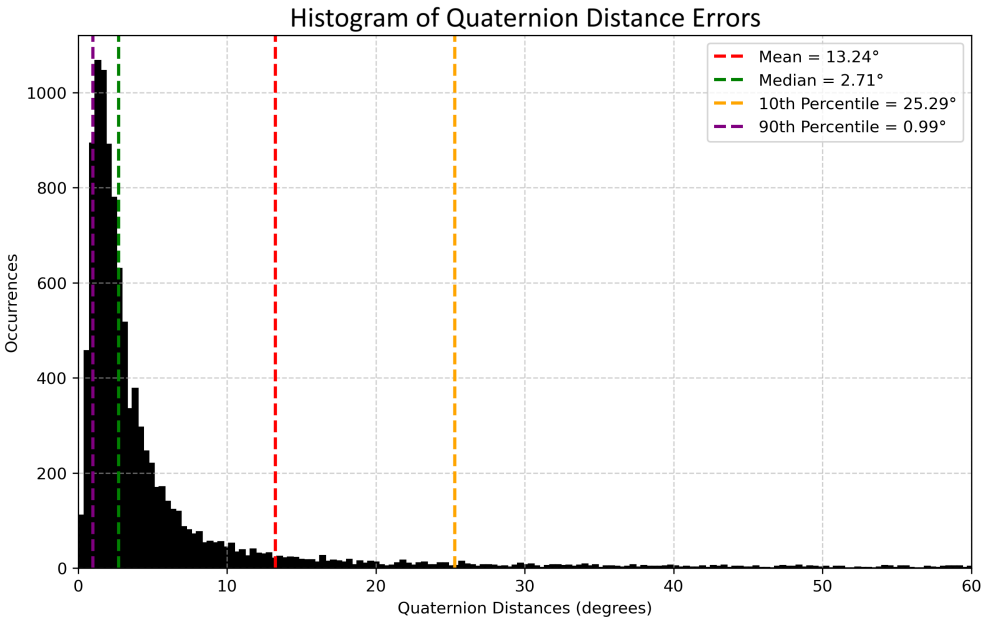


Fig. 6: Histogram of quaternion distance rotational errors for the full 11,000-image 6DOF test set with predictions produced by the coarse localizer and 6DOF pose model. As shown, the median error is significantly lower than the mean error.

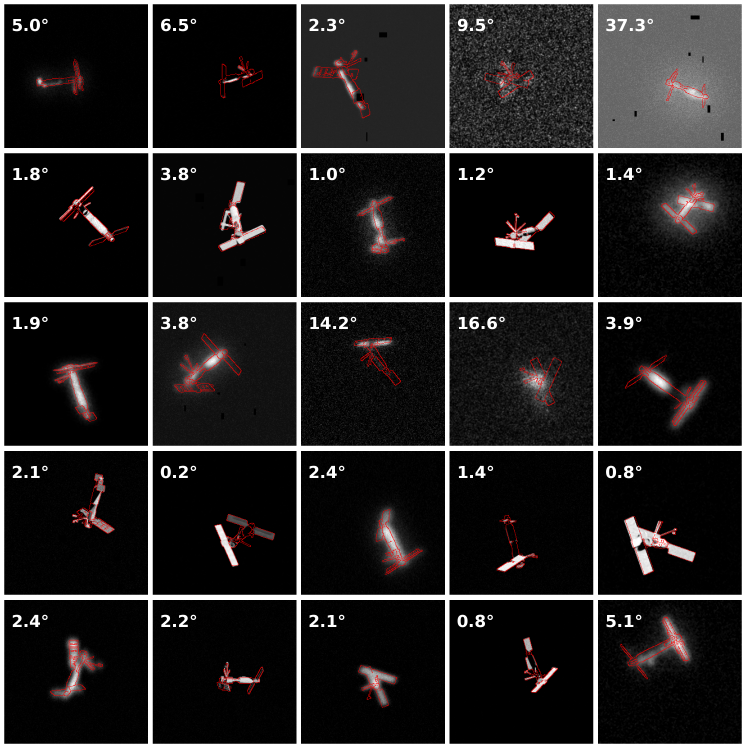


Fig. 7: Pose predictions from the model (coarse localizer, 6DOF pose model) are shown as red edges overlaid on the first 25 grayscale test images. The quaternion distance rotation error in degrees for each test image is shown at the upper left corner of each image.

The improvement in mean errors, and the lack thereof in median errors, suggest that the decoupled dynamic zoom is unrealistic for low-quality images – such that randomly perturbed localization is more accurate than the localization predicted by the coarse localizer model. By necessity, our training and testing datasets contain imagery of much lower quality than the data used in [3, 8, 2]. This difference in training and test data results in poor RoI crops (from coarse localizer inference) and therefore increased pose error for low-quality images. It is possible that a coupled training and testing process, where coarse localizer predictions are used to create the RoI crop data for training the 6DOF pose model, could improve mean performance. Alternatively, localization for training could be perturbed based on image quality rather than entirely randomly, or the RoI crop size could be purposefully expanded, as suggested by [3], to ensure ”the area containing the object is approximately half the RoI,” reducing the likelihood of cropping out parts of the object even with poor localizations. All performance metrics above are provided for individually tested images, with no temporal processing. When time-series data were available, a windowed Kalman filter was applied to the coarse localizer output prior to dynamic zoom and another windowed Kalman filter was applied to the 6DOF pose model output, reducing error. Normalized localization and pose parameters are shown before and after Kalman filter application in Fig. 8.

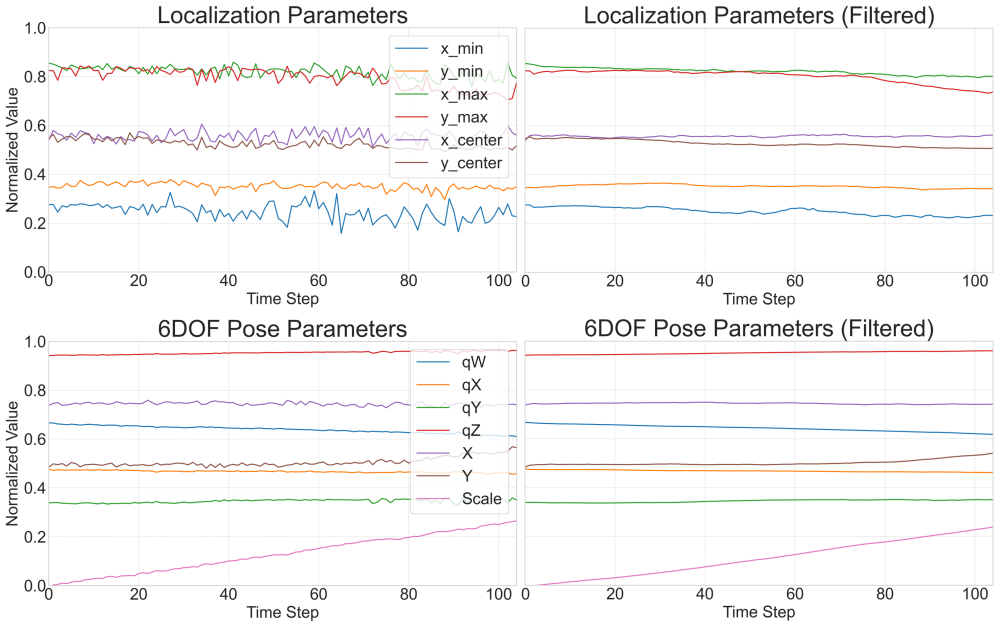


Fig. 8: Localization and pose parameters before and after filtering by a temporal, windowed Kalman filter. The top left section shows the localization parameters (bounding box lower right and upper left coordinates, and object center coordinates) as a function of video frame (timestep) before filtering, while the top right section displays these parameters after filtering. The bottom left section presents the normalized pose parameters (*WXYZ* quaternion rotation, *X* coordinate, *Y* coordinate, and image orthographic scale) before filtering, and the bottom right section shows these parameters after filtering. The Kalman filter was applied in a manner simulating real-time operation, with the filtered values at each timestep calculated using only the data from previous timesteps. All values have been normalized to a range of 0 to 1 for visualization.

The full model (coarse localizer, 6DOF pose model, and temporal Kalman filter) achieved a mean rotational error of 4° (3° median, 11° maximum) on 105 frames of real video of Seasat captured by a ground-based telescope, successfully bridging the Sim2Real domain gap. A single frame with pose prediction overlaid is displayed in Fig. 9. Without Kalman filtering the model achieved a mean rotational error of 5° (4° median, 57° maximum). Ground truth rotational pose labels for the real imagery were manually estimated by overlaying the CAD model on the real images and adjusting pose parameters until the model and image aligned. Manual pose estimation is an inexact process, therefore the ground truth pose labels contain significant uncertainty. Given the lack of real, accurately labeled data we are currently producing a new, fully labeled synthetic test set using Digital Imaging and Remote Sensing Image Generation (DIRSIG) with a perturbed satellite CAD model (to account for potential inaccuracies in satellite CAD models) to render object planes and High Contrast Imaging for Python (HCIPy) to apply physically realistic noise and AO blur [10]. The aim is to demonstrate and quantitatively evaluate model performance across a simulated domain gap (given changes in the CAD model, rendering software, and image degradation process) for a range of atmospheric conditions.

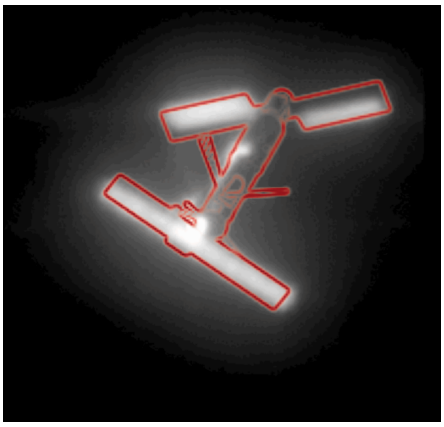

Fig. 9: Pose prediction from the full model (coarse localizer, 6DOF pose model, and temporal Kalman filter) overlaid as red edges on a real image of Seasat captured by a ground-based telescope. Pose estimate accuracy was visually similar for the other 104 frames.

## 6. ACKNOWLEDGEMENTS

The authors acknowledge Research Computing at the Rochester Institute of Technology for providing computational resources and support that have contributed to the research results reported in this publication [11]. DISTRIBUTION A. Approved for public release: distribution is unlimited. Public Affairs release approval #AFRL-2024-3237.

## REFERENCES

[1] J. Lucas, T. Kyono, M. Werth, N. Gagnier, Z. Endsley, J. Fletcher, and I. M. AFRL, "Estimating satellite orientation through turbulence with deep learning," *Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS) Conference Proceedings*, 2020.

[2] T. H. Park, M. Märtens, M. Jawaid, Z. Wang, B. Chen, T. J. Chin, D. Izzo, and S. D'Amico, "Satellite pose estimation competition 2021: Results and analyses," *Acta Astronautica*, vol. 204, 2023.

[3] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," *Proceedings of the IEEE/CVF Conference on Computer Vision*, 2021.

[4] T. H. Park and S. D'Amico, "Robust multi-task learning and online refinement for spacecraft pose estimation across domain gap," *Advances in Space Research*, 2023.

[5] B. O. Community, "Blender - a 3d modeling and rendering package," 2024.

[6] Y. Zhou, C. Barnes, J. Lu, A. Research, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[7] M. Phelps, J. Z. Gazak, T. Swindle, J. Fletcher, and I. Mcquaid, "Inferring space object orientation with spectroscopy and convolutional networks," *Advanced Maui Optical and Space Surveillance Technologies Conference (AMOS) Conference Proceedings*, 2021.

[8] Z. Li, G. Wang, and X. Ji, "Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7677–7686, 2019.

[9] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information (Switzerland)*, vol. 11, 2020.

[10] E. H. Por, S. Y. Haffert, V. M. Radhakrishnan, D. S. Doelman, M. van Kooten, and S. Bos, "High contrast imaging for python (hcipy): an open-source adaptive optics and coronagraph simulator," p. 152, SPIE-Intl Soc Optical Eng, 7 2018.

[11] Rochester Institute of Technology, "Research computing services," 2019.