Wide Field Camera 3 Database Visualization: Understanding an Astronomical Archive using Dimensionality Reduction

THE UNIVERSITY OF CHICAGO DATA SCIENCE INSTITUTE
2024 AI+Science Summer School
July 15-19 2024

Frederick Dauphin (fdauphin@stsci.edu, GitHub:FDauphin) and the WFC3 Team
Space Telescope Science Institute

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

Introduction

The Wide Field Camera 3 (WFC3) onboard the Hubble Space Telescope (HST) is one of the premier instruments in astronomy, being responsible for countless discoveries in astrophysics. Since its installation in May 2009, WFC3 has captured a diverse set of over 300K observations, including globular clusters, galaxies, and nebulae. As future telescopes begin to dramatically increase astronomy's data volume, it is critical that we optimize data exploration using machine learning because traditional methods will be impractical. Since the archive holds a deep and rich complexity, WFC3 is an optimal sandbox for exploring astronomical imaging data using unsupervised learning.

Dataset and Processing

WFC3 consists of detectors UVIS for ultraviolet and visible bands, and IR for infrared bands [1,2]. We limited our data to high quality, direct external images of extrasolar objects with nominal readouts. We made training, validation, and test sets split by date to ensure there were no temporal shifts. For UVIS, we used ~91K images (78K/5K/8K). For IR, we used ~81K images (75K/3K/3K). We then reduced our data by converting all exposure units to electrons, min-max clipping pixels to one electron and the 99.9th percentile, log-scaling, resizing to 128x128, and rescaling the pixels to [0,1].

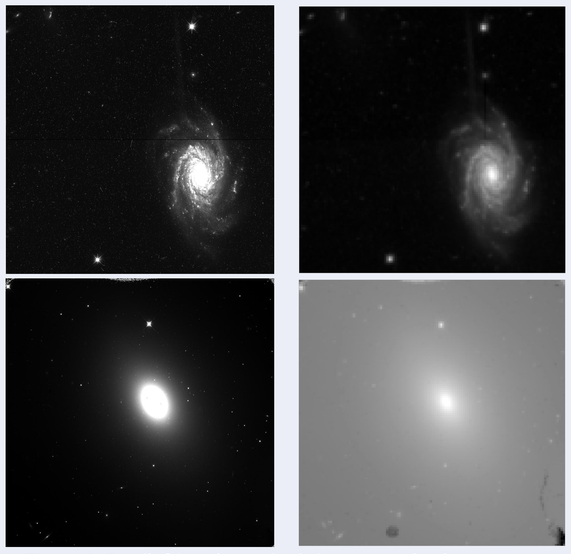

Figure 1. Pre (left) and post (right) processed galaxy images taken by UVIS (top; iedh24hrq) and IR (bottom; iedy60e0q).

Methods

For intermediate dimensionality reduction, we used principal component analysis (PCA) [3] and a convolutional autoencoder (ConvAE) [4] with 8 latent features. PCA uses singular value decomposition to find the axes of most variation, and projects the data onto those axes. ConvAE uses convolutional layers to encode features to a latent representation and decode a latent representation to features. For the final dimensionality reduction, we used uniform manifold approximation and projection (UMAP) [5], which builds neighbor graphs in high dimension (8) and projects that graph to a low dimension (2). This final embedding should be a feature rich representation of our data.

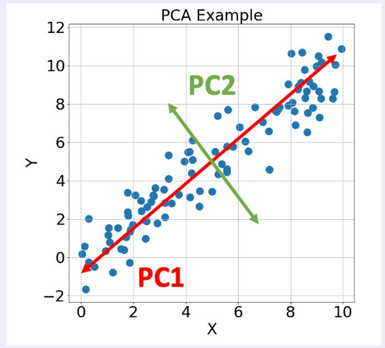

Figure 2. PCA example using linear data. The most variation is along PC1.

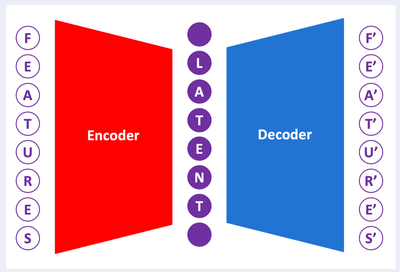

Figure 3. Autoencoder overview. Features are encoded to the latent space and decoded to the feature space.

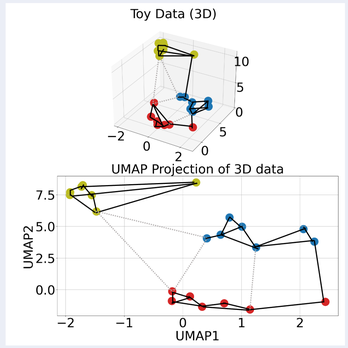

Figure 4. UMAP example. Graphs are built in 3D and projected in 2D.

UVIS and IR UMAP Projections

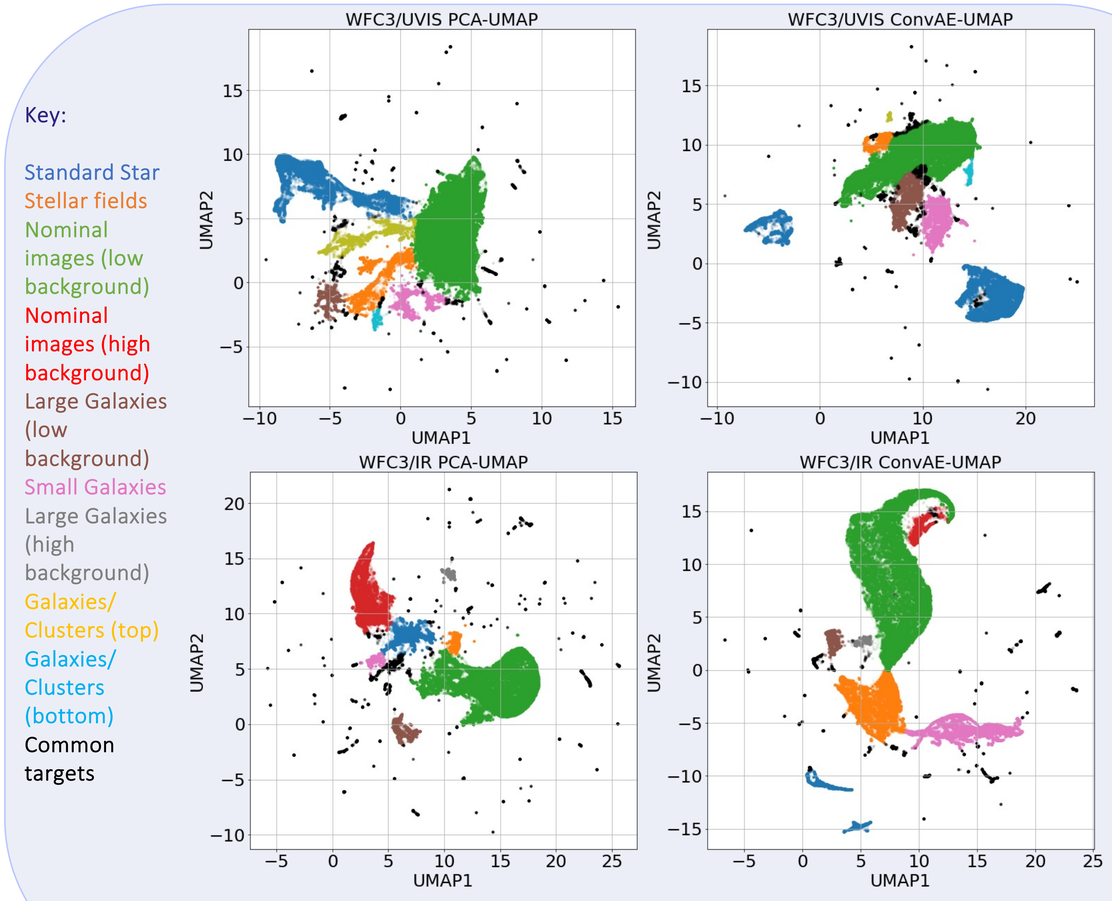

Figure 5. UMAP projections for UVIS and IR using PCA and ConvAE latent features. A color-coated key for the labels is on the left. Labels were determined using DBSCAN with additional manual labeling, and act as a high level description of the objects found within each cluster.

Evaluation

The 2D PCA-UMAP and ConvAE-UMAP reductions for both UVIS and IR were complex, exemplifying the diversity of the images. The main, large clusters consisted of typical observations of sparse star fields with low background images separating from high background images. Some subclusters emerged, containing standard star observations, galaxies, and position-based observations. The "anomalous clusters" on the outskirts were dozens of images of a particular observation. This result was a slight weakness since these anomalies didn't reduce to a main or subcluster. Although PCA-UMAP sufficiently reduced the data, ConvAE provided more complexity due to the rich features learned in the latent space.

Conclusions and References

Since first light in May 2009, WFC3 has amassed an archive of over 300K observations. With future telescopes producing abundant datasets, utilizing unsupervised learning is a necessity to efficiently analyze astronomical data and understand the data's structure. We created a dataset of ~172K images (91K/UVIS; 81K/IR), and processed each image to standard [0,1] 128x128 images for efficient modeling. We then reduced the data to a 2D embedding using PCA, ConvAE, and UMAP. We were able to recover the global and local structure of the datasets with similar astronomical images clustered near each other. ConvAE and PCA qualitatively performed well for our purposes. For future work, we plan to investigate different methods, such as variational autoencoders and pretrained Inception embeddings [6,7].

[1] Marinelli, M. & Dressel, L., et al., 2024, "WFC3 Instrument Handbook, Version 16.0", (Baltimore: STScI).
[2] Pagul, A. & Rivera, I., et al., 2024, "WFC3 Data Handbook, Version 6.0", (Baltimore: STScI).
[3] Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." *Chemometrics and intelligent laboratory systems* 2, no. 1-3 (1987): 37:52.
[4] Kramer, Mark A. "Nonlinear principal component analysis using autoassociative neural networks." *AIChE journal 37,* no. 2 (1991): 233:243.
[5] McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426* (2018).
[6] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).
[7] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." *In Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 1-9. 2015.