AI-driven Classification and Imputation Techniques for High-contrast Imaging in Astronomy

Fangyi Cao ♠ Bin Ren ♢ Weixin Yao♠
♠fcao017@ucr.edu ♢bin.ren@oca.eu
♠weixin.yao@ucr.edu

Introduction

The rise in computational power and the advent of advanced AI models offer a promising solution to these challenges. AI has shown remarkable success in handling vast datasets across various fields, yet its potential in astronomy, specifically for direct exoplanet imaging, remains largely untapped. This project aims to leverage AI to uncover new exoplanets and enhance the detection of circumstellar disks. Even a single new protoplanet discovery could transform our understanding of planet formation. Moreover, improving H-band disk detection by a factor of 50 will provide crucial insights into the dust properties of planet-forming and debris disks.

This project seeks to develop a streamlined AI-driven approach to analyze both existing and future datasets of VLT/SPHERE, moving beyond the current labor-intensive techniques. By integrating AI into exoplanet imaging, we aim to pave the way for new discoveries and a deeper understanding of exoplanetary systems.

1 State of the Art

Several methodologies are available in the post-processing procedure of the observations. Under Reference Differential Imaging (RDI), statistical methods such as Principle Component Analysis (PCA[1]) and Non-negative Matrix Factorization ([2]), are generalized for data imputation. As the recent AO improvements have enhanced image quality, the lack of revolutionary methods severely hinders the analysis of high-contrast imaging data. We herein propose to develop deep semi-supervised learning techiniques for high-contrast imaging data analysis.

Research Goals

- Automate binary classification of reference and target star from GPI and SPHERE.
- Imputation of coronagraphic data, with a focus on avoiding self- and over-subtraction to image new planets and characterize circumstellar disks.
- Using meta data to recover the point spread functions (PSFs), thus boosting the observation efficiency from 50% (RDI imaging when reference star exposures are needed) to 100% by removing the need for reference stars.

2 Methodology and Preliminary Analysis

2.1 The classification Model: labeling a reference star, or a target?

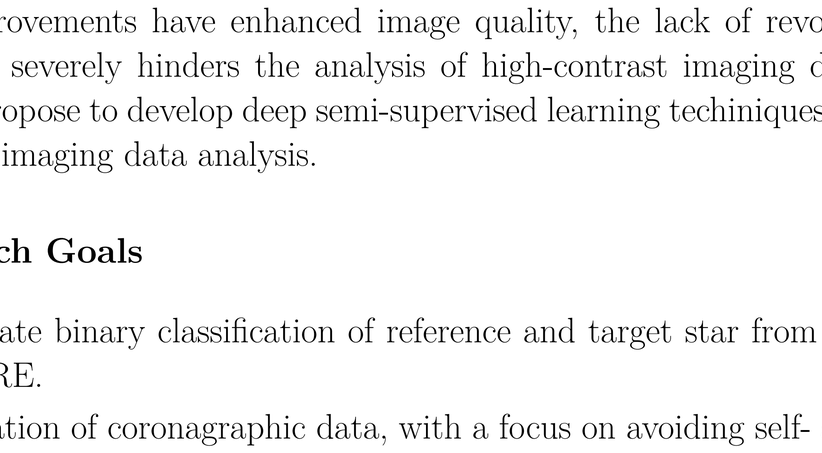

Figure 1: Top: The Semi-supervised CNN model employing a pseudo loss technique for the polarized image classification discussed in section 2.1. Bottom: The Semi-supervised SOS-VAE model for the imputation of reference image discussed in section 2.2.

Our VLT/SPHERE dataset includes 288 labeled and 2439 unlabeled circumstellar images in polarized light. We use 10% of the labeled images for validation and the rest, along with all unlabeled images, for training. Our CNN architecture, inspired by the holistically-nested edge detection model, focuses on edge response extraction for binary classification. Using pseudo-labeling, we train our model for 250 epochs with PyTorch's Adam optimizer (batch size: 16, learning rate: 0.01). Our framework achieves an 87.6% classification accuracy on the validation set, outperforming the baseline PCA and logistic classifier methods ($< 80\%$ accuracy).

2.2 The Imputation Model: stellar signals on exoplanetary objects.

With pairs of labelled target and reference images from section **2.1**, I will adopt the Variational Autoencoders (VAEs) approach for the reduction and imputation of our images, specifically adopting the Structured Observation Space VAE (SOS-VAE)approach. Unlike standard VAEs, which use pixel-wise independent distributions, SOS-VAE incorporates spatial dependencies with a fully populated covariance matrix. This method aligns with the modified Rician distribution expected from AO system speckles. I will mask non-central regions of reference images to isolate labeled speckles, treating the central region with potential target objects as unlabeled data. Our semi-supervised SOS-VAE model, starting with 5 latent variables, aims to improve disk image recovery. Performance will be compared using PCA, DIKL, and random forest imputation models.

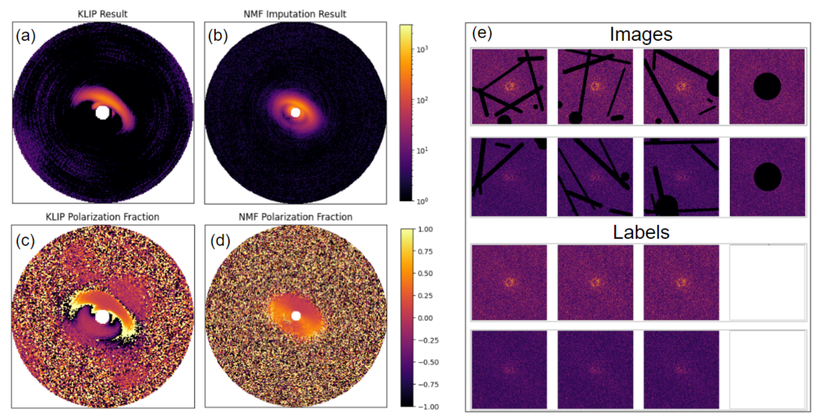

Figure 2: (a) KLIP imputation with unrealistic negative values indicating overfitting. (b) NMF imputation showing scalability issues. (c) KLIP polarization fraction calculated from (a). (d) NMF polarization fraction derived from (b). (e) Example training data with random masks and the corresponding labels for the SOS-VAE model , which is aimed to outperform both NMF and KLIP. The absence of a label is denoted by a blank.

3 Conclusions

This project will deliver significant advancements in exoplanet imaging through high-quality circumstellar disk images, polarization fraction maps, and identification of direct imaging candidate planets. By publicly releasing our deep learning-based pipeline, we will facilitate transfer learning for future datasets, addressing the scarcity of exoplanet images.

Our work aligns with NASA's mission objectives to enhance coronagraph contrast and efficiency, supporting initiatives in planetary data archiving and discovery data analysis. The developed tools and methods will streamline the discovery and characterization of exoplanets, contributing valuable resources to the astronomical community and advancing our understanding of planetary formation and evolution.

References

[1] B. Ren, L. Pueyo, C. Chen, É. Choquet, J. H. Debes, G. Duchêne, F. Ménard, and M. D. Perrin, "Using Data Imputation for Signal Separation in High-contrast Imaging," , vol. 892, p. 74, Apr. 2020.

[2] B. B. Ren, "Karhunen-Loève data imputation in high-contrast imaging," , vol. 679, p. A18, Nov. 2023.