Deep Optics for Optomechanical Control Policy Design

Justin Fletcher
University of Hawai'i at Mānoa
jfletch@hawaii.edu

Abstract

An emerging class of Fizeau optical telescopes have the potential to upend prior cost scaling models, substantially improving the angular resolution and contrast attainable by ground-based astronomical instruments. However, this design introduces a challenging visual control problem that must be solved to compensate for wavefront aberrations induced by the flexible substructure it employs. We subvert this problem with a deep optics approach to policy design and image recovery that exploits, rather than corrects, aberrations to obtain domain-specific object recovery performance exceeding that of more costly filled aperture designs.

1. Introduction

The answers to many fundamental questions about our universe lie hidden within the diffraction limit of modern optical telescopes [30]. Light diffracts at the edge of the telescope's aperture, dispersing the inbound wavefront and mixing intensity information from different angular regions. The innermost peak of the resulting pattern of light circumscribes the angular extent that the sensor can disambiguate [19]. Larger apertures (i.e., optical baselines) have smaller point-spread functions (PSFs) and better angular resolution [22], but this comes at a price; cost increases with aperture diameter, placing a practical upper limit on resolving power [9]. To build the instruments needed to confirm the existence of exoplanetary life [2] and to solve practical problems posed by the expanding scope of human activity in space [12, 18], a new approach to telescope design is needed.

Distributed-aperture telescopes scale more effectively than traditional designs but introduce difficult actuation design challenges. In this work, we propose a solution to one such challenge: inter-aperture phase errors introduced by structural flexibility in a leading distributed aperture telescope design. We formulate the manipulation of optical diffraction to correct for these errors as an optomechanical metasurface actuation control problem and develop a deep optics approach that yields performant policies and policy-aligned task models. Our approach compensates

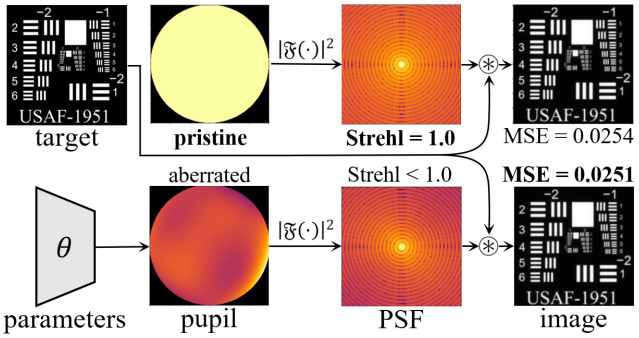

Figure 1. An optomechanical surface can be adapted to the object of observation to improve recovery. Here, we adapt an actuated aperture to a target, which adds wavefront error (i.e., decreases Strehl) but improves recovery (bold is better). The improvement is small, but it serves as an existence proof. How might we learn actuations that improve recovery on a broader target domain?

for phase differences between apertures, while exploiting residual phase differences to improve image recovery performance when observing a chosen class of imaging target.

The geometry and actuation of a sensor's aperture determines the spatial frequencies (i.e., features) it captures in the images it forms. If the target of observation is known, one may apodize (i.e., shape) the aperture to maximize sensitivity to that target [16] as shown Figure 1. If the target is not known, one can only seek to maximize recovery of all possible images by minimizing aberrations. From this perspective, the removal of aberrations, which is the goal of primary aperture design and the purpose of adaptive optics [21], may be viewed as a naive but powerful strategy for image recovery. That strategy works equally well on all imaging targets because it incorporates no information about those targets. It does not, however, constitute a *fundamental* limit on angular resolution [34]. Instead, it represents the best one can do *without foreknowledge of the target*. We are driven to ask: given some information about the object (e.g., a dataset representing the object distribution), how might we use that information to enhance imaging?

Our key insight is that object domain information may be incorporated into an optical system by jointly learning a sequential aperture actuation plan and image recovery

model using task gradients propagated through a differentiable proxy model of the system. In this way, domain- and task-specific features come to be represented by the weights of the recovery model and the sequence of parameters specifying the control plan, which are learned end-to-end. We apply this insight to a challenging control problem that, once solved, unlocks a technology development pathway that has the potential to change the cost-size relationship for large telescopes.

Contributions. We 1) extend the study of deep optics to distributed aperture (i.e., Fizeau) image formation, including atmospheric effects and active, high-order actuation; 2) provide evidence supporting the hypothesis that deformable mirrors alone can act as optical feature extractors; and 3) demonstrate the efficacy of deep optics when applied to the construction of jointly learned optomechanical metasurface control policies and recovery algorithms under realistic atmospheric and imaging noise conditions.

2. Related Work

Large telescopes. The emerging ExoLife Finder (ELF) class of Fizeau telescope designs employ a tensegrity-suspended annulus of subapertures to reduce the moving mass of the primary aperture [10] without sacrificing optical baseline (i.e., outer diameter). Reducing the mass needed to support additional aperture diameter improves cost scaling, which expands the set of astronomical objects (e.g., exoplanets) that one can practically afford to detect [11]. Unfortunately, this design also necessitates solutions to difficult wavefront error correction caused by the spatially distributed subapertures shown in Figure 2. Further, like all ground-based telescopes, ELF must contend with wavefront errors caused by atmospheric turbulence.

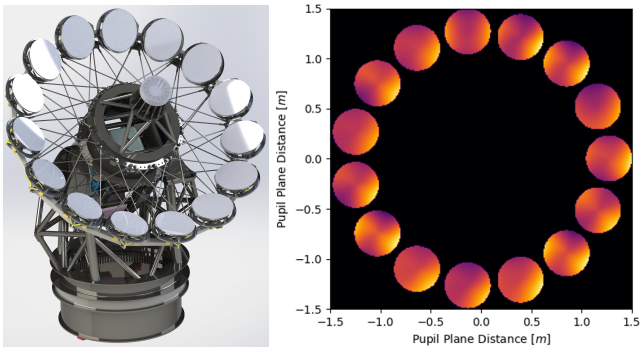

Figure 2. The ExoLife Finder distributed aperture telescope design (left) includes a tensegrity-supported annulus of primary apertures which we model as an articulated aperture (right).

Rather than viewing this as an issue to correct, we follow our key insight and treat it as just one of many complications in the design of control policies for domain-specific image reconstruction. One variant of the ELF design includes a secondary deformable mirror (DM) that enables actuation of the aperture. This, in turn, allows us to influence the interference of light on the focal plane to achieve domain- and task-specific objectives. Viewed in this way, we see that DMs can be thought of as dynamic (i.e., sequentially controllable) optomechanical metasurfaces. While this work focuses on ELF, the approach is applicable to any dynamic optomechanical metasurfaces used to image distant objects.

Deep Optics. The joint optimization of optical element design parameters and learned image processing is an area of active research known as deep optics [4, 14, 17]. WISH is a deep optics approach that achieves simultaneous high resolution and high quality wavefront phase measurement that has been applied to achieve high quality image reconstruction in the telescopic imaging regime [32]. In traditional adaptive optics (AO) astronomical imaging, wavefront information is used to remove aberrations using a deformable mirror inserted along the optical path [3]. While a combination of WISH and AO may be used to mitigate aberrations, this approach incorporates neither downstream image processing task nor object domain information. Further, the division between wavefront estimation and wavefront correction represents a silo between these two tasks; significant progress has been made in deep optics by identifying and removing silos in the image formation process [5]. Like WISH and WISHED [31], our proposed approach includes a dynamic optical element along the path. Unlike these techniques, the state of the dynamic element in our work (i.e., the DM actuations) are learned, providing a means by which to extend end-to-end learning from the downstream task model to the sequence of commands that determine the optical properties of the instrument. Following Tseng et. al. , we frame our approach as an optical metasurface design problem [24], but extend this formulation to include dynamic surfaces.

Optomechanical adaptation. The features preserved by an aperture are determined by its modulation transfer functions (MTF). An MTF that responds to all frequencies equally will maximize recovery of any object, but this is not attainable in practice because the range of an MTF is constrained by the geometry of the aperture from which it is derived (see Section 3). The MTF can, however, be adjusted within those constraints through aperture actuation. The MTF that results in the least aberrated image is the MTF that allocates its actuation capacity (e.g., DM stroke) to those the spatial frequencies that are most common in the object, while neglecting those that are less common. Extending this reasoning from a single target object to a target object *domain* (i.e., a class of object planes), we observe that the most performant constrained MTF is one that responds to the *expectation* of the spatial frequency distribution of that

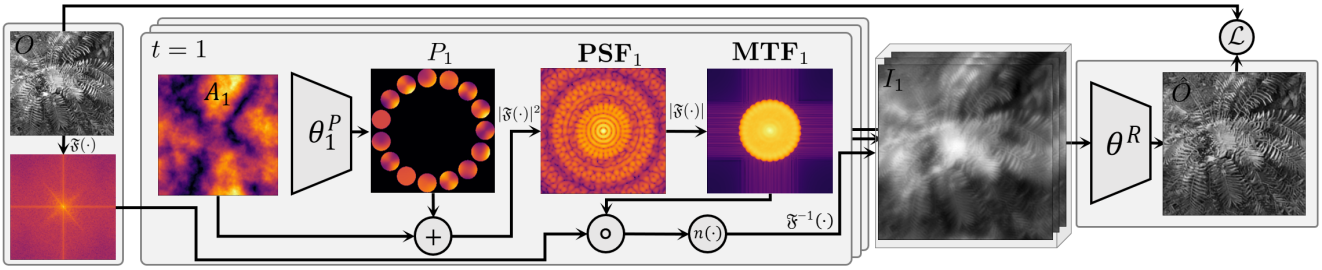

Figure 3. Overview of the end-to-end model illustrated using a single object example (left). The aperture actuation function, parameterized by θ^P, is added to the masked atmospheric phase (A_1). This aberrated aperture (P_1) is used to produce an MTF with which the object spectrum is multiplied to form an image spectrum. The inverse Fourier transform of this spectrum yields the image for a single image of the ensemble (center left). Each element of the ensemble corresponds to the formation of a single image; across these images, the parameterized aperture function, atmosphere, and noise vary. The ensemble is generated simultaneously and then is stacked (center right). The ensemble is combined by the image recovery model, parameterized by θ^R to produce an object estimate for which the recovery loss is computed (right).

domain.

From this perspective, we can see that a task-adapted control policy need not *necessarily* eliminate all wavefront errors introduced by the ELF substructure (i.e., differential motion) and atmospheric turbulence to enable performant image recovery. Naturally occurring differences in inter-aperture phase strictly increase the number of reachable aperture actuation states. The constraints on the MTF are only reduced by this feature of the system. This additional freedom must be compensated for, which is the wavefront control problem that motivates this work, but that freedom confers the benefit of greater adaptability. Thus, by adopting a learned approach to actuation and recovery, we *subvert* the original problem.

3. Differentiable Optomechanical Simulation

Our objective is to jointly train a DM actuation plan and image recovery model that, when used together, produce accurate recovery estimates when used to image an object domain represented by a dataset. Following our key insight, we adapt recent work in differentiable simulation of optical systems [6] to serve as the deep optics physical layer (i.e., proxy) representing an ELF-class optical system. This differentiable model produces images, which serve as the input domain of a trainable image recovery solution (e.g., frame stacking, deep neural networks) that corresponds to the deep optics digital layer of our architecture. We train the full deep optical model end-to-end by propagating task loss gradients through the recovery algorithm to the physical layer parameters by way of the deep optical proxy model.

Optical proxy model overview. Fourier optics provides an efficient computational model of optical image formation [22]: an image, I, is the dot product of the MTF and the power spectrum of an object, O. The MTF, in turn, is the real component of the Fourier transform of the PSF, which is the squared amplitude of the spectrum of the pupil function. In our work, T images are taken in rapid succession, such that O is constant while the pupil function changes due to structural instability, atmospheric effects, and planned articulations. Noise is introduced during transduction of light into electrons on the focal plane, and is modeled by a combination of Gaussian and Poisson noise following [24], a sample of which is denoted n_t. In summary, an image collected at time t is modeled as

$$I_t = \mathfrak{F}^{-1}\left(\mathbf{MTF}_t \circ \mathfrak{F}(O)\right) + n_t \tag{1}$$

where

$$\mathbf{MTF}_t = |\mathfrak{F}(\mathbf{PSF}_t)| = \left|\mathfrak{F}\left(|\mathfrak{F}(P_t)|^2\right)\right| \tag{2}$$

in which $\mathbf{PSF}_t$ and P_t are the PSF and pupil function at time t, respectively, and $\mathfrak{F}$ is the Fourier transform operator.

Clearly, the relationship between P_t and the object determines the formed image, but what determines P_t? We answer this question in Sec 3.2 in terms of a parameterization, θ_t^P, that specifies the articulations of P_t. Collectively, we refer to the sequence of these parameterized articulations as a *plan*, $\theta^P = \{\theta_0^P, \theta_1^P, \ldots, \theta_T^P\}$.

For each object, the execution of a plan produces an *ensemble* of images, $\mathbf{I} = \{I_0, I_1, \ldots, I_T\}$. An image recovery model, f, maps $\mathbf{I}$ to an object estimate, $\hat{O}$, and is parameterized by θ^R. We represent the domain of objects to which our model is adapted using a dataset, $\mathbf{O}$. Task performance is then estimated by a task loss, $\mathcal{L}$. We seek

$$\underset{\theta^P,\, \theta^R}{\operatorname{argmin}} \;\; \mathbb{E}_{\mathbf{O}}[\mathcal{L}(O, \hat{O})], \tag{3}$$

where $\hat{O} = f(\mathbf{I}, \theta^R)$, which we pursue via mini-batch stochastic gradient descent.

In this section, we propose an approach to this problem beginning with the object domain, then moving through the optical system model to the recovery model, and ending in the task loss. We conclude the section with a discussion of the policy and model update process.

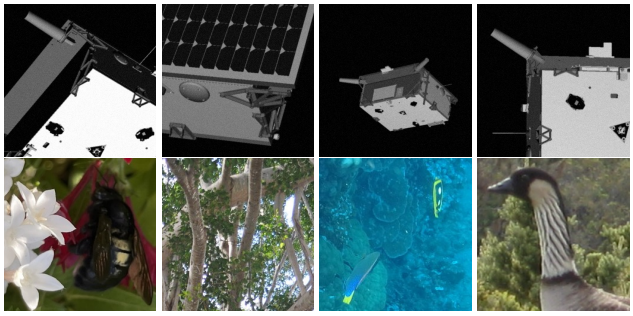

Figure 4. Chips from the SPEED+ (top) and iNaturalist (bottom) datasets, which we use as object planes. These chips are cropped to the sizes used in this work.

3.1. Object Domain

Image formation begins with the object of observation. An *object plane* encodes the wavefront incident upon the aperture of the telescope, originating from the object of observation, as a 2-dimensional raster[1] of normalized luminosity values over a spatial extent.

We must adapt our plan and recovery model to a domain of possible object planes because the object of observation is not known *a priori*. For example, even if the identity of the target is known, it may be rotated or illuminated in a way not previously observed. We represent the distribution of potential object planes using a dataset comprising samples from that distribution. Examples from these datasets are shown in Figure 4.

We use SPEED+ and iNaturalist to represent our observation target domains. Examples from these domains are encoded as single-channel rasters in which each cell (i.e., pixel) ranges from 0 to 1, representing a wavefront comprising monochromatic 1 μm light. Examples of these datasets are shown in Figure 4.

SPEED + (Synthetic). The SPEED+ dataset comprises images of a small satellite and is a leading dataset for satellite pose estimation research [15]. We use SPEED+ to approximate an imaging task to which a ground-based distributed aperture telescope would be applied: spatially-extended imaging of anthropogenic satellites in low-Earth orbit [28]. The physical parameters that define the proxy model (e.g., focal length, aperture extent) are chosen so as to correspond to imaging an 1 m object at a distance of $1,000$ km which is the approximate distance to low-Earth orbit. We filter the SPEED+ dataset to include only the synthetic image subset, and from that subset we remove all images with a simulated background. The synthetic subset is useful because it excludes imaging artifacts that may otherwise complicate an optical bench demonstration of the techniques described in this work; the inclusion of iNaturalist is intended to ensure that our approach generalizes to real data. Our SPEED+ comprises $50,000$ images of a single satellite in different poses and illumination conditions. All images are center-cropped to 512×512 pixels which is representative of modern focal plane arrays. We partition the dataset into $30,000$ training examples, and set aside $10,000$ for validation and $10,000$ for testing. This dataset corresponds to training against a sample of satellite images and then generalizing to unseen poses and illumination.

iNaturalist. The iNaturalist comprises images of species in their natural environment annotated with the taxa to which those species correspond [26]. We use the mini variant of the 2021 iNaturalist dataset [25] for this study. Although smaller than the full dataset, the variant we adopt contains $500,000$ training images which we believe to be sufficient for training behavior we wish to observe; the full dataset validation partitions of $100,000$ images is used. All iNaturalist images are center-cropped to 256×256, as this provides for more computationally efficient experimentation without substantial loss of spatial frequency information. The iNaturalist dataset does not correspond exactly to the physical imaging scenario intended for distributed aperture telescopes, because the object planes are neither of a consistent length-scale nor are their contents (i.e., plants and animals) found at $1,000$ km above the surface of the Earth. Nevertheless, this is what our proxy model simulates. We include iNaturalist because the semantics of a dataset do not directly influence learning under our approach. Thus, we may use iNaturalist to enable comparison to a larger body of machine learning work and to evaluate learning performance on a richer and more diverse dataset.

3.2. Distributed aperture telescope proxy models

Before we can compute and apply task gradients, we must first form an image of the object. This image must model imaging through a pupil function that is, in turn, conditioned upon the parameterized articulation plan. For notational convenience, we introduce our model for a single subaperture first, then generalize to an annulus, and group into an ensemble. We model the pupil function of each subaperture as a radial J-term Zernike polynomial,

$$Z(\rho, \varphi \mid \theta) = \sum_{j=0}^{J} \theta_j^P Z_n^m(\rho, \varphi), \tag{4}$$

in which our parameterization of the aperture actuation (i.e., the Zernike coefficients) is denoted θ^P. The radial Zernike function is given by

$$Z_n^m(\rho, \varphi) = \begin{cases} R_n^m(\rho)\,\texttt{cos}(m\varphi) & m = 0 \\ R_n^m(\rho)\,\texttt{sin}(m\varphi) & m \neq 0 \end{cases} \tag{5}$$

[1] While this raster conforms the common usage of the term word "image," we will reserve that term to refer to an image formed by an optical system on a focal plane to prevent confusion.

where the radial polynomial, R_n^m, is

$$R_n^m(\rho) = \sum_{k=0}^{\frac{n-m}{2}} \frac{(-1)^k (n-k)!}{k!(\frac{n+m}{2}-k)!(\frac{n-m}{2}-k)!} \rho^{n-2k}. \quad (6)$$

The OSA/ANSI single index polynomial scheme [23] defines m and n for any value of j, completing our model of an actuated aperture. Each value of j corresponds to a single differentiable expression, which we implement directly.

Aperture composition. To construct an ELF-like multi-aperture telescope pupil function, we arrange N individual pupil functions into an annulus on a spatial grid, as illustrated in Figure 2. Construction of this annulus is quite involved and the details are not essential for our objectives. We closely follow the geometric annulus construction algorithm introduced in [6], but denote the combined aperture in shorthand as

$$P\left(u, v \mid \theta^P\right) = \sum_{n=0}^{N} Z\left(u, v \mid \theta_n^P\right) \texttt{mask}\left(u, v, r_s\right), \quad (7)$$

where θ_n^P is the parameterization of the articulations planned for sub-aperture n, r_s is the radius of a subaperture, and `mask` is the circular mask function. The actuated pupil function constructed by this method encodes aperture deflections in units of radians of phase.

Aberrations. Our work is motivated by the need to correct for structural aberrations in a distributed aperture telescope system. To model this effect, we introduce a small amount of piston noise in each subaperture by randomly generating the initial value of parameters corresponding to the Zernike coefficients, θ^P, in each exposure. This simulates the physical challenge that a learned plan will need to overcome when used to control a DM on a distributed aperture telescope: the initial perturbations caused by structural flexibility will not be known.

Additionally, all ground-based optical telescopes must contend with wavefront errors caused by atmospheric turbulence. These changes in the arrival time of different wavefront regions result in undesirable image aberrations. We use the von Karman atmosphere model [20] to generate uncorrelated phase screens. This complicates training because the recovery models must learn to perform recovery in the presence of phase noise that is only partially correctable by the DM. All phase screens used in this work are generated with an outer scale parameter of $2,000\ m$, an inner scale parameter of $1\ m$, and an Fried parameter, r_0, of $20cm$. The Fried parameter measures the wavefront aberration caused by the atmosphere at a given wavelength; in our work, we use $1\ \mu m$. An example phase screen is labeled A in Figure 3.

Both forms of aberration represent the most challenging assumptions (i.e., spatially and temporally uncorrelated noise) for their respective phenomena. Through this choice, we pose the hardest possible version of our task in an attempt to measure the worst-case performance.

Image formation. We now have a complete model of the image formation process from object illumination through the focal plane. The last step on the path to an image is the transduction of light into electrical signals by a camera. To account for this process, we model read noise is modeled as a Gaussian random variable and shot noise as a Poisson random variable [21]. We differentiate through these sources of noise using the reparameterization trick [8] and score-gradient trick [29], respectively.

Ensemble imaging. Our image formation approach models a single image conditioned upon a single DM actuation. Our goal is to execute several actuations successively to emphasize different object features, thereby increasing the diversity of features available to the recovery model. To model this feature, we simulate an ensemble image formation simultaneously and stack the output. This stack is known as an ensemble, and it represents several exposures formed from a single object. We don't impose a time scale on this ensemble, but in practice the inter-exposure period will be defined by the effective actuation frequency of the DM, which is expected to be $200Hz$.

3.3. Image Recovery

The output of our optical proxy model is an ensemble of images. Each image is a different representation of the object plane features extracted during the image formation process based on DM actuations. To build an estimate of the original object of observation we must combine these disparate representations. We consider two approaches to image recovery, each of which is differentiable.

Frame stacking. The simplest method of image recovery is known as frame stacking. As the name implies, this method consists of adding a sequence of frames to increase contrast. We modify this method by weighting each image by a learnable parameter, collectively denoted θ^R. The resulting object estimation algorithm is $\hat{O} = \Sigma_{t=0}^{T} \theta_t^R I_t$.

Learned recovery. We adopt the learned decoding model described in [24] to enable comparison of reconstruction approaches across different diffraction regimes, but borrow modifications of that model architecture from [6] to simplify model capacity tuning. This model comprises multi-scale feature extractor followed by a feature fusion network, in which every convolutional block is parameterized by a filter

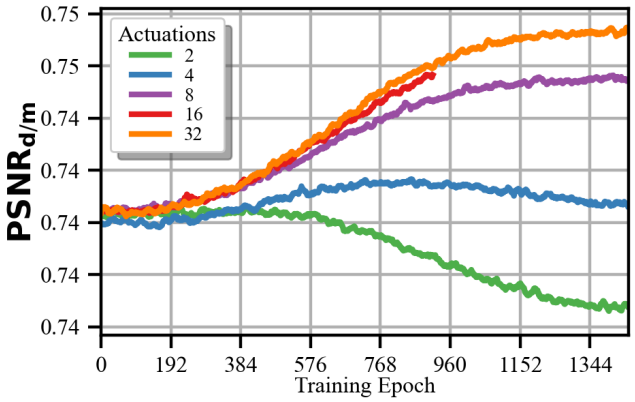

Figure 5. Adapting a sequence of deformable mirror actuations to a single target enables improved recovery, even without sophisticated image combination techniques. Recovery also improves with increasing sequence length, which suggests that deformable mirrors are able to extract features that are relevant to an object.

scale, which dictates the number of filters used in the model. In Section 4.3 we compare the recovery performance of small, medium, and large instances of this model, which correspond to filter scales of 4, 8, and 16, respectively. While hallucination is possible in any learned recovery approach, we take care to evaluate model generalization performance on held-out examples from a large, visually distinct dataset to ensure that overfitting is punished with decreased metric performance.

3.4. Policy design

The preceding sections describe a model that takes a set of Zernike coefficients, produces a set of focal plane images, and recovers an object plane estimate. There remains the task of choosing the Zernike coefficients for each subaperture in each exposure from which this model proceeds. Together these choices comprise an open-loop policy (i.e., plan) for the DM. Once selected, these coefficients can be provided directly to the mirror controller for actuation.

Loss. We use the mini-batch mean absolute error (MAE),

$$\mathcal{L}_{\texttt{MAE}}(B, O) = \frac{1}{|B|} \sum_{\hat{O} \in B} \left| \sum \hat{O} - O \right|, \tag{8}$$

over a mini-batch of object estimates, B, as our loss function. The MAE loss is effective for image restoration tasks [33], which are closely analogous to image recovery. We compute the gradients of our loss with respect to the parameters of both our recovery model and optical proxy model using the TensorFlow autodifferentiation [1] framework and apply the gradients using an adaptive momentum optimizer [7].

Metrics. Image restoration and super-resolution task performance is often measured using the peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [27], and mean squared error (MSE) between the object and object estimate. However, these metrics measure only the recovery performance of the model in isolation. They do not tell us how a recovered image compares to the image formed by a perfectly phased, filled (i.e., monolithic) aperture telescope of the same diameter under equivalent conditions. The application that motivates this work is the development of alternatives to traditional filled aperture imaging, so this is essential when judging the potential utility of our work. As such, we report each distributed aperture metric as a quotient of that metric and the same metric achieved through filled aperture imaging.

We construct our metrics such that an increase in metric value corresponds to an increase distributed aperture performance relative to filled aperture performance. Thus, we report MSE as

$$\mathbf{MSE_{m/d}} = \frac{\mathbf{MSE_m}}{\mathbf{MSE_d}},$$

while SSIM and PSNR are denoted

$$\mathbf{SSIM_{d/m}} = \frac{\mathbf{SSIM_d}}{\mathbf{SSIM_m}} \quad \text{and} \quad \mathbf{PSNR_{d/m}} = \frac{\mathbf{PSNR_d}}{\mathbf{PSNR_m}},$$

respectively. We say that a model achieves *recovery parity* in a metric when the value of that metric is 1.0. Recovery parity indicates that a plan and recovery model enable a distributed aperture telescope to match the performance of a filled aperture telescope of the same size. This satisfies our original goal because we have produced an actuation plan that enables a distributed aperture telescope to perform at least as well as a more costly monolithic design.

4. Experiments and Results

We now turn to the central question of this work: to what extent can jointly learned metasurface actuations and recovery models compensate for the challenges introduced by distributed aperture telescopes? Our answer to this question is divided into three experiments, which are presented in order of increasing difficulty and generality. First, we evaluate the ability of a DM plan *alone* to extract useful representations from a known object when trained using our approach in Section 3.4. Then, in Section 4.2 we address the challenge of learning a plan and a recovery model end-to-end, such that the resulting solution generalizes to unseen data of the same class of objects. Finally, Section 4.3 describes the extent to which our approach is able to produce performant models that generalize well when applied to a dataset that includes unseen target types in the presence of atmospheric turbulence.

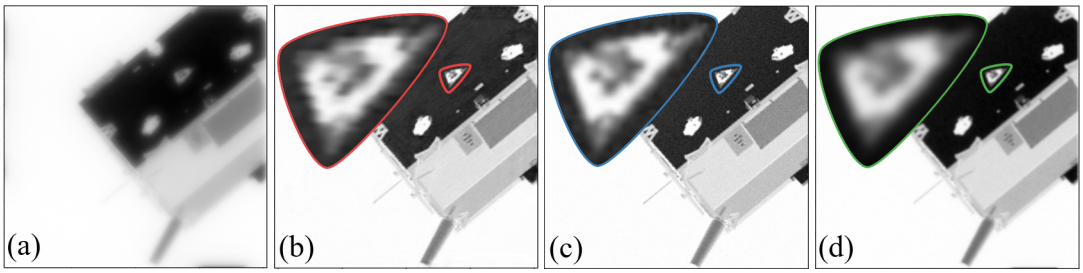

Figure 6. Qualitative image recovery performance improves during model training. We illustrate the progression from (a) a raw distributed aperture image to (b) an object estimate from a recovery model trained for 128 epochs, and provide comparison with (c) the object of observation (i.e., ground truth) from the validation partition and (d) an aberration-free monolithic aperture recovery of that same object. Insets show a detailed view of a triangular feature.

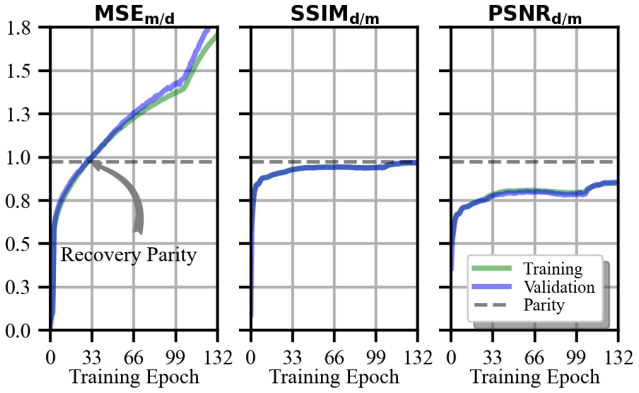

Figure 7. MSE, SSIM, and PSNR increase during training for both the training and validation partitions of the SPEED+ dataset.

4.1. Optical metasurfaces are feature extractors

Physically, when we actuate a DM to extract certain spatial frequency features, we manipulate the interference of light on the focal plane such that those spatial frequencies that are not relevant to the chosen task are suppressed, while those that are relevant are retained. This phenomenon is modeled using Fourier optics, in which the MTF of the pupil plane emphasizes some spatial frequencies in the object spectrum while neglecting others. By modeling the system in this way, it becomes clear that the DM is a *physically realized* feature extractor acting upon the object wavefront. The computation that extracts features from the wavefront is done by optical diffraction, much like diffraction neural networks [13]. Our approach involves iterating through many DM actuations in rapid succession (i.e., faster than the object changes) and generating a stack of learned representations. This leads us to ask: to what extent can these physical feature extractors *alone* adapt to a chosen object domain and task?

We explore this question by simulating the imaging of a known target with a distributed aperture telescope while using only naive frame-stacking for image recovery. This corresponds to training, validating, and testing our end-to-end model on a single image. We apply this approach to the Air Force Test Target shown in Figure 1.

Figure 5 illustrates that DM actuations alone can be used to realize feature extraction for improved image recovery. Additionally, we find that both the maximum PSNR achieved and the rate of improvement during training increase as the number of unique parameterized pupil functions increases. These observations support the intuition that rapid, sequential actuations of imaging metasurfaces can be used to capture a set of learned representations. However, we observe that increasing the sequence length alone is insufficient to reach recovery parity and provides diminishing returns after approximately 16 exposures. Furthermore, we report that this approach fails in the presence of any realistic amount of atmospheric turbulence.

4.2. Jointly learned control policy and recovery

When frame stacking is used for image recovery, we observe diminishing returns as we add additional DM actuations. This is unsurprising because the recovered image is only a linear combination of the extracted feature maps. In effect, this recovery model acts as a decoder for the encoding produced by the image formation process. As such, a naive image recovery process limits the utility of more sophisticated feature representations. To further improve recovery performance, learned image recovery is needed.

We jointly train our DM actuation plan and the small convolutional image recovery model described in Section 3.3 on the $30,000$ image SPEED+ dataset training partition, without turbulence, and measure metric performance on both the validation and training partitions separately. We report a recovery advantage in $\mathbf{MSE_{m/d}}$ after approximately 30 epochs, as illustrated in Figure 7. We also achieve $\mathbf{SSIM_{d/M}}$ recovery parity at approximately 130 epochs.

Qualitatively, we confirm that the recovery quality of object estimates produced using end-to-end trained actuation plans and recovery models exceeds that of images formed by a filled aperture of the same size. This is consistent with the

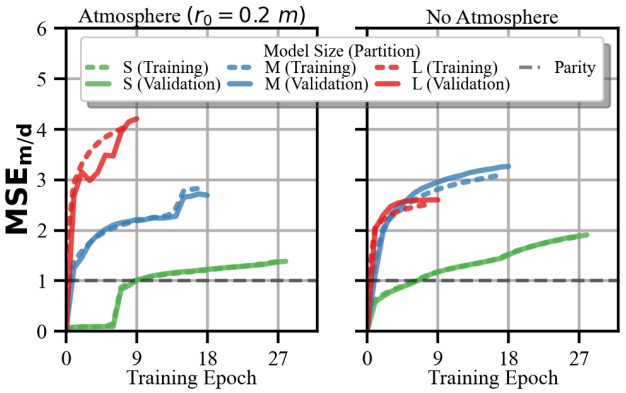

Figure 8. Larger recovery models generalize more effectively to unseen examples and have less difficulty with atmospheric turbulence. One epoch is one pass through the `iNaturalist` dataset.

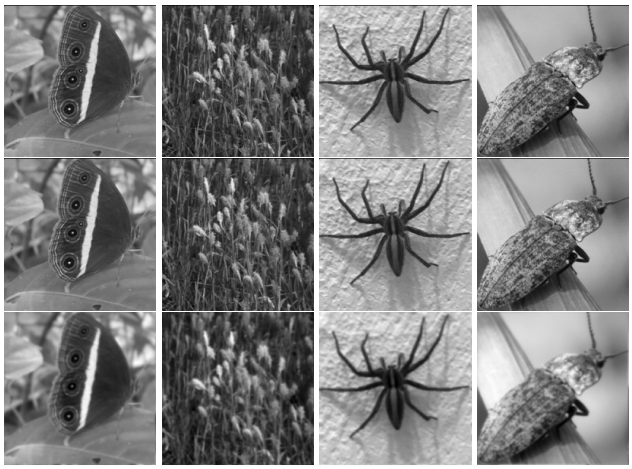

Figure 9. Recovered object estimates (top) for ground truth `iNaturalist` object planes (middle) compare favorably with filled aperture images (bottom) qualitatively.

quantitative performance measured by our metrics. Figure 6 illustrates recovery of high spatial frequency features that are lost during imaging by a filled aperture telescope.

4.3. Generalization

The preceding experiment evaluated our learned recovery and actuation approach against a dataset that represents the intended application domain for this technology. Unfortunately, those results tell us little about the extent to which our approach generalizes to observation targets that are outside the training data distribution because `SPEED+` contains only a single object viewed under different conditions. For a more realistic challenge, we train on `iNaturalist` both with and without an atmosphere.

We fix the size of the ensemble of the models at 8 in this experiment while varying the model filter scale between 4 (small), 8 (medium), and 16 (large) to explore the relationship between capacity and generalization. By turning on and off our atmosphere simulation, we can also assess the degree to which the learned plan and recovery model are able to compensate for aberrations. Training and validation performance are shown in Figure 8 and validation set performance is summarized in Table 1. Each model was trained for 72 GPU hours on an NVIDIA A100 with batch sizes of 256, 128, and 64 for the small, medium, and large models, respectively.

Table 1. Maximum validation set recovery metrics for several model sizes with and without atmospheric turbulence.

ATMOSPHERE	MODEL	$\text{MSE}_{m/d}$	$\text{SSIM}_{d/m}$	$\text{PSNR}_{d/m}$
$r_0 = 0.2\ m$	4	1.39	0.97	0.77
$r_0 = 0.2\ m$	8	2.72	1.04	0.89
$r_0 = 0.2\ m$	16	**4.21**	**1.07**	**0.95**
`None`	4	1.91	0.99	0.81
`None`	8	**3.27**	1.05	0.90
`None`	16	2.60	**1.07**	**0.93**

We report improved recovery performance, with clear recovery advantages as measured by SSIM and MSE. We also observe near-parity in PSNR. Generalization performance is observed to increase almost uniformly with model size, as expect. One experiment is observed to break this trend; we attribute this deviation to a bad training initialization, and include it, rather than re-running the experiment, to provide an accurate representation of training stability. These results provide evidence that distributed aperture telescopes equipped with learned wavefront control solutions will achieve performance that is comparable to more costly filled aperture designs. Figure 9 enables qualitative assessment of the recovered images.

5. Conclusion

Distributed aperture telescopes may open a new frontier in astronomy, but methods to correct the wavefront errors inherent to their design must first be developed. We propose a deep optics approach that incorporates both task and domain information, and show that it achieves image recovery that compares favorably to imaging with more costly filled aperture designs, even in the presence of atmospheric turbulence. Avenues for future work include specialized task losses for objectives such as direct exoplanet imaging and biosignature detection, as well as adaptations for other optical information processing objectives. Reformulating this problem as a visuomotor sequential decision making task may also prove useful.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and Michael Isard. Tensorflow: A system for large-scale machine learning. In *12th ${$USENIX$}$ Symposium on Operating Systems Design and Implementation (${$OSDI$}$ 16)*, pages 265–283, 2016.

[2] Anthony J. Beasley, Mark Dickinson, Eric J. Murphy, Sidney Wolff, and Michael H. Wong. Astro2020 APC White Paper Multiwavelength Astrophysics in the Era of the ngVLA and the US ELT Program. 2020.

[3] Jacques M. Beckers. Adaptive optics for astronomy: Principles, performance, and applications. *Annual review of astronomy and astrophysics*, 31(1):13–62, 1993.

[4] Julie Chang and Gordon Wetzstein. Deep Optics for Monocular Depth Estimation and 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10193–10202, 2019.

[5] Steven Diamond, Vincent Sitzmann, Frank Julca-Aguilar, Stephen Boyd, Gordon Wetzstein, and Felix Heide. Dirty Pixels: Towards End-to-end Image Processing and Perception. *ACM Transactions on Graphics*, 40(3):23:1–23:15, May 2021.

[6] Justin Fletcher and Peter Sadowski. Towards jointly learned control policies and image recovery for distributed aperture telescopes. In *Sensors and Systems for Space Applications XV*, volume 12121, pages 75–86. SPIE, June 2022.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, Jan. 2017.

[8] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[9] Jeff Kuhn, Jean-Fabien Capsal, Ian Cunnyngham, Maud Langlois, Kevin Lewis, Nicolas Lodieu, Gil Moretto, Rafael Rebolo, Joe Ritter, Ryan Swindle, Ye Zhou, Stuart Jefferies, Kritsadi Thetprapi, David Audigier, and Justin Fletcher. The small-ELF project: Toward an ultra-large coronagraphic optical receiver. In *Ground-Based and Airborne Telescopes IX*, volume 12182, pages 161–184. SPIE, Aug. 2022.

[10] Jeffrey Kuhn, Gil Moretto, Svetlana Berdyugina, Maud Langlois, Jean-Fabien Capsal, Mike Gedig, and Kritsadi Thetpraphi. *The Exo-Life Finder (ELF) Telescope: Design and Beam Synthesis Concepts*. July 2018.

[11] J. R. Kuhn, S. V. Berdyugina, M. Langlois, G. Moretto, E. Thiébaut, C. Harlingten, and D. Halliday. Looking beyond 30m-class telescopes: The Colossus project. In *Ground-Based and Airborne Telescopes V*, volume 9145, pages 533–540. SPIE, July 2014.

[12] Trent Kyono, Jacob Lucas, Michael Werth, Brandoch Calef, Ian McQuaid, and Justin Fletcher. Machine learning for quality assessment of ground-based optical images of satellites. *Optical Engineering*, 59(5):051403, Jan. 2020.

[13] Xing Lin, Yair Rivenson, Nezih T. Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. All-optical machine learning using diffractive deep neural networks. *Science*, 361(6406):1004–1008, 2018.

[14] Christopher A. Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep Optics for Single-Shot High-Dynamic-Range Imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1375–1385, 2020.

[15] Tae Ha Park, Marcus Märtens, Gurvan Lecuyer, Dario Izzo, and Simone D'Amico. SPEED+: Next-Generation Dataset for Spacecraft Pose Estimation across Domain Gap. *arXiv:2110.03101 [cs]*, Dec. 2021.

[16] Martin Paúr, Bohumil Stoklasa, Jai Grover, Andrej Krzic, Luis L. Sánchez-Soto, Zdeněk Hradil, and Jaroslav Řeháček. Tempering Rayleigh's curse with PSF shaping. *Optica*, 5(10):1177–1180, Oct. 2018.

[17] Yifan (Evan) Peng, Ashok Veeraraghavan, Wolfgang Heidrich, and Gordon Wetzstein. Deep optics: Joint design of optics and image recovery algorithms for domain specific cameras. In *ACM SIGGRAPH 2020 Courses*, SIGGRAPH '20, pages 1–133, New York, NY, USA, Aug. 2020. Association for Computing Machinery.

[18] Matthew Phelps, J. Zachary Gazak, Thomas Swindle, Justin Fletcher, and Ian Mcquaid. Inferring Space Object Orientation with Spectroscopy and Convolutional Networks. *AMOS (Sept. 2021)*, 2021.

[19] Rayleigh. Investigations in optics, with special reference to the spectroscope. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 8(49):261–274, Oct. 1879.

[20] Jason Daniel Schmidt. Numerical simulation of optical wave propagation: With examples in MATLAB. SPIE, 2010.

[21] Daniel J. Schroeder. *Astronomical Optics*. Elsevier, 1999.

[22] Edward G. Steward. *Fourier Optics: An Introduction*. Courier Corporation, 2004.

[23] Larry N. Thibos, Raymond A. Applegate, James T. Schwiegerling, and Robert Webb. Standards for reporting the optical aberrations of eyes, 2002.

[24] Ethan Tseng, Shane Colburn, James Whitehead, Luocheng Huang, Seung-Hwan Baek, Arka Majumdar, and Felix Heide. Neural nano-optics for high-quality thin lens imaging. *Nature Communications*, 12(1):6493, Nov. 2021.

[25] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12884–12893, 2021.

[26] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[27] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[28] Michael Werth, Brandoch Calef, Kevin Roe, and Amanda Conti. LUCID: Accelerating Image Reconstructions of LEO Satellites Using GPUs. In *2020 IEEE Aerospace Conference*, pages 1–11, Mar. 2020.

[29] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992.

[30] Michael H. Wong, Karen J. Meech, Mark Dickinson, Thomas Greathouse, Richard J. Cartwright, Nancy Chanover, and Matthew S. Tiscareno. Transformative Planetary Science with the US ELT Program. *arXiv:2009.08029 [astro-ph]*, Sept. 2020.

[31] Yicheng Wu, Fengqiang Li, Florian Willomitzer, Ashok Veeraraghavan, and Oliver Cossairt. WISHED: Wavefront imaging sensor with high resolution and depth ranging. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, Apr. 2020.

[32] Yicheng Wu, Manoj Kumar Sharma, and Ashok Veeraraghavan. WISH: Wavefront imaging sensor with high resolution. *Light: Science & Applications*, 8(1):44, May 2019.

[33] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.

[34] Sisi Zhou and Liang Jiang. Modern description of Rayleigh's criterion. *Physical Review A*, 2019.