Optics EXPRESS

Dual-plane wavefront sensing using a vision transformer

EVAN O'ROURKE AND KEVIN O'KEEFFE[*]

Faculty of Science and Engineering, Department of Physics, Swansea University, Singleton Park, Swansea, SA2 8PP, United Kingdom
[*]*K.Okeeffe@Swansea.ac.uk*

Abstract: Image-based wavefront sensing using deep-learning allows Zernike coefficients to be estimated directly from intensity measurements. To date, the majority of experiments have focused on using convolutional neural networks to estimate coefficients. Here, we demonstrate a dual-plane wavefront sensor trained using a vision transformer model and compare its performance to that of the widely used convolutional neural network (CNN) architecture. Both results of experiment and simulation indicate that the vision transform can outperform the CNN where image data is significantly downsampled, due to the former's ability to more accurately predict high-order Zernike coefficients.

1. Introduction

Rapid and precise measurement of wavefront aberrations is critical for many applications, including astronomy [1], ophthalmology [2], and microscopy [3,4]. Wavefront sensing covers a broad range of techniques in which the shape and/or phase of a wavefront in an optical system are measured with respect to a reference wavefront. Typically, a wavefront sensor (WFS) such as a Shack-Hartmann sensor [5], or a pyramid sensor [6,7] is used to measure the gradient of an impinging wavefront. A wide variety of interferometric systems have also been developed to extract wavefront information, such as the Fizeau and shearing interferometers [8]. Alternatively, wavefront information can be acquired from measured spatial intensity distributions using iterative phase retrieval algorithms [9–11], however, these methods are often slow to converge, limiting their use in applications where real-time wavefront correction is required. Combined with adaptive optic elements such as a spatial light modulator (SLM) or deformable mirror, wavefront sensing allows for the correction of aberrations that can otherwise severely limit the performance of imaging systems. Interferometry and conventional wavefront sensors offer reliable methods for recovering wavefront information. However, these approaches can increase the cost and complexity of systems, and in the case of interferometry, place stringent requirements on stability.

Recently, there has been immense interest in using deep learning methods to extract wavefront information directly from spatial intensity distributions. Such deep learning wavefront sensing (DLWFS) takes advantage of the ability of machine learning algorithms, such as convolutional neural networks (CNN), to map non-linear relationships between measured spatial intensity profiles and the Zernike coefficients describing the wavefront, offering the prospect of rapid and robust wavefront sensing without the need for a conventional WFS or interferometer. Significant progress has been made with numerous models and applications demonstrated. Models have been trained to enable improved imaging in scattering media [12–14], retrieve Zernike coefficients from atmospheric turbulence [15,16], and correct for aberrations in the focal plane [17–23]. Once trained, DLWFS has also been demonstrated to allow faster wavefront retrieval than traditional algorithms [24–26].

#586748 https://doi.org/10.1364/OE.586748
Journal © 2026 Received 3 Dec 2025; revised 18 Jan 2026; accepted 21 Jan 2026; published 12 Feb 2026

To date, many DLWFS studies have focussed on the use of intensity distributions measured at a single plane, typically the back focal plane of a lens. However, due to the symmetry present in even-order Zernike polynomials, it is not possible to unambiguously determine the sign of certain coefficients solely from a measurement of the intensity [22,27]. This ambiguity presents a challenge for DLWFS based on measurements at a single plane, as learning models often fail to converge in this case. A solution to this was proposed by Siddik et al. [22] where a modified set of Zernike polynomials, containing only the absolute value of even polynomial coefficients, was used for training. This method avoids convergence failure but has limited practical use as negative value coefficients may be present in real systems. Phase diversity has also been demonstrated as a means to mitigate sign ambiguity but requires an annular entrance pupil [28].

A more robust method to avoid sign ambiguity is to record the spatial intensity distribution at two longitudinally-separated planes, one of which is typically the focal plane. Such dual-plane sensing has been shown to provide sufficient data for a CNN to predict even and odd coefficients, without the need to change the beam or sample [19,20,23,24,29–31], at the expense of a slight increase in the complexity of the system.

Until recently, the majority of DLWFS investigations have focused on the use of CNNs. For pattern recognition tasks CNNs offer many advantages including translational invariance, data efficiency, low prediction times and excellent feature recognition. However, vision transformer (ViT) models have recently gained significant attention. Unlike CNNs, which use a kernel to perform pixel-wise operations on input data, ViTs break images into patches and process each one, giving more weight to information-rich patches. ViTs have demonstrated superior performance compared to CNNs in numerous image-based tasks such as diagnostic medical imaging and classification problems [32]. They have also been used in a host of optical experiments including compressive sensing [33], phase-retrieval [34,35], plenoptic sensors [36], and wavefront sensing [20,37]. An in-depth study of the use of ViT models for the correction of aberrations induced by atmospheric turbulence was carried out by Liu et al. [38], with the ViT in that study outperforming several popular CNNs while also being more robust to noise, demonstrating the potential for ViTs to replace CNN models. However, a drawback of ViTs is that training times tend to be longer than for CNNs due to their higher computational complexity, dependence on the patch size, and requirement for larger datasets.

In this paper we demonstrate dual-plane DLWFS trained using a ViT and compare its performance to a CNN for both experimental and simulated data. The performance of the DLWFS is investigated for different sizes of training datasets, as well as the effect of down-sampling input data. The CNN is observed to outperform the ViT, until the data is downsampled significantly in which case the ViT outperforms the CNN due to its ability to more accurately predict higher order Zernike coefficients. These results indicate that ViTs may be advantageous for DLWFS applications where low-resolution images are available, or where data is deliberately downsampled in order to boost performance, such as training and inference times.

2. Experiment setup and data collection

The experimental setup for the dual-plane DLWFS is illustrated in Fig. 1. The beam of a helium-neon laser (wavelength 633 nm) was expanded and collimated using a telescope consisting of two lenses of focal length 200 mm and −50 mm placed 20 cm apart, resulting in a beam approximately 3.5 mm in diameter. The collimation of the beam after the telescope was verified using a Shack-Harmann sensor. The beam was then reflected from a SLM (Hammamatsu model X13138-02) at an angle of incidence of approximately 10 degrees and passed through a lens of focal length 500 mm, with the SLM located in the front focal plane of the lens. To perform dual-plane sensing the spatial profile of the laser was recorded at two separate planes. To achieve this a 1 mm thick 50:50 beamsplitter was used to split the beam after the lens. One camera (UI-3880CP-C-HQ Rev.2), located at the back focal plane of the lens, recorded the transmitted portion of the beam. A second camera (UI-1240SE-M-GL), located approximately 7 cm past the back focal plane, recorded the reflected part of the beam. A 500 μm diameter pinhole, located close to the focus of the reflected beam, was used to block unwanted diffraction from the SLM. The intensity of the laser was controlled using the combination of a half-wave plate and polarizing cube beamsplitter, as well as a selection of neutral density filters. The polarizing cube beamsplitter also ensured the horizontal polarization required for optimum phase modulation using the SLM.

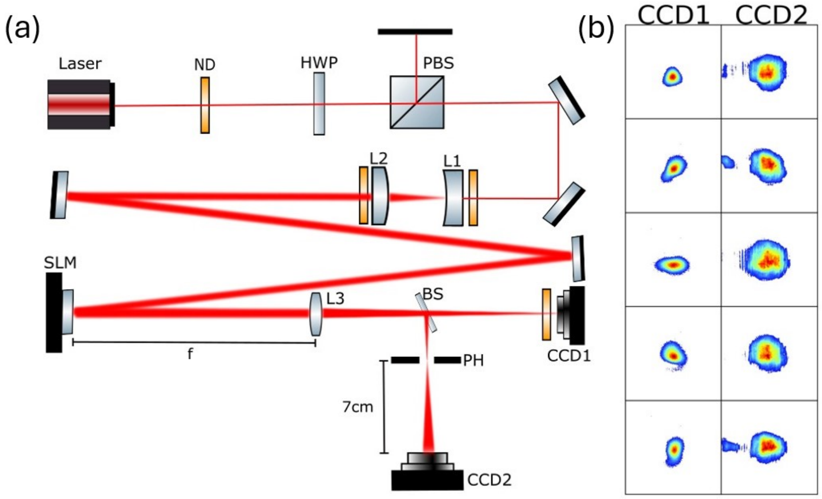

Fig. 1. (a) Experiment setup for collection of training data. ND, neutral density filter; HWP, half-waveplate; PBS, polarising beam splitter; L1-L3, lenses; SLM, spatial light modulator; BS, beam splitter; CCD1-CCD2, charged-coupled device cameras; PH, pinhole. (b) Random selection of measured intensity profiles.

Training data was acquired by altering the phase of the beam using the SLM. A common way to describe the phase $\phi(x, y)$ of a wavefront is to use a set of orthogonal functions called Zernike polynomials. These polynomials are defined across a unit circle and each term represents a different aberration [39–41]. Each polynomial mode Z_j is weighted by a coefficient a_j with the Noll indexing scheme [41]:

$$\phi(x, y) = \sum_j a_j Z_j(\rho, \theta). \tag{1}$$

The Zernike polynomials are defined by:

$$\begin{aligned} Z_j(\rho, \theta) &= Z_n^m(\rho, \theta) \\ &= \begin{cases} \sqrt{2n+1} R_n^m(\rho) \cos m\theta & \text{if } m \neq 0 \text{ and } j \text{ is even,} \\ \sqrt{2n+1} R_n^m(\rho) \sin m\theta & \text{if } m \neq 0 \text{ and } j \text{ is odd,} \\ \sqrt{n+1} R_n^m(\rho) & \text{if } m = 0 \end{cases} \end{aligned} \tag{2}$$

where n, m and j are the index orders, ρ and θ are polar coordinates and $R(\rho)$ is a radial function. Using the Noll indexing scheme, a dataset of 100, 000 phase maps, each with different Zernike coefficients, was generated. It has previously been shown that the inclusion of very high-order Zernike modes can artificially increase the performance in DLWFS, as CNNs tend to lower the contributions from higher-order modes [27]. To mitigate this the first 18 Zernike coefficients (excluding piston, tip and tilt) were used with each coefficient randomly generated as a number between -1 and 1, which was then divided by the radial order of its respective coefficient. This gives an approximation of a power spectral density profile associated with typical polishing errors in optics [42].

A constant manufacturer-supplied calibration mask was applied to the SLM and a Shack-Hartmann sensor was used to verify that the majority of aberrations were corrected for. A blazed grating was also added to each mask to separate modulated light into the first diffraction order. Each of the phase masks was applied to the SLM, with the Zernike polynomials occupying a circular regions at the centre of the mask, and the images from both cameras were recorded. The camera in the focal plane had a 2.4 μm pixel size and recorded a 200×200 pixel region of interest. The second camera had 5.3 μm pixel size and, due to the diffraction of the beam past the focal plane, recorded a 400×400 pixel region of interest. With all $200,000$ ($100,000$ for each plane) images collected, each focal-plane image was concatenated with its respective out-of-focal-plane image (now downsampled to $(200, 200, 2)$ using bicubic interpolation), creating a final dataset of $100,000$ concatenated images of shape $(200, 200, 2)$.

3. Models and training

The dataset was used to train a ViT and compare its performance to a standard CNN. There are an abundance of CNNs available but the Residual Neural Network (ResNet) architectures are still a popular choice for many applications, performing on a par with other models [43] and have been shown to work well for DLWFS applications [19,20,23,29]. In this work ResNet-34 [44] was used as it has been demonstrated to perform well for a wide range of applications while requiring less training time than its larger counterparts such as ResNet-50 or ResNet-101. An illustration of how a CNN processes input data can be seen in Fig. 2(a). A kernel representing a matrix of numbers with learnable weights is scanned across the input image, performing convolutions to generate a feature map, with a single layer in the network employing multiple kernels. Stacking several such layers allows the network to extract progressively more abstract features. The size of the kernel and the number of pixels it moves across at a time are user defined, typically changing per layer throughout the CNN model as the convolutions reduce the image spatially.

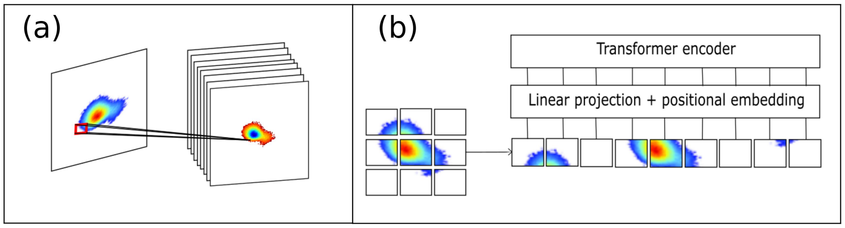

Fig. 2. Comparison of input image processing by (a) CNN and (b) ViT models.

In this experiment a total of 16 residual layers were used in the model, each containing three convolutional layers. A dense layer of 1000 with a dropout rate of 0.2 is used before the final output layer of 18 neurons with linear activation representing the Zernike coefficients. These parameters were chosen based on initial tests to find the best results in terms of accuracy and training efficiency. The mean squared error between the predicted and actual coefficients was used as the loss function. The model was trained and tested on different dataset sizes and different image sizes with root mean squared error (RMSE) used to measure the performance of the model. Min-max normalisation was carried out on the training images before the model was run as it has been shown to improve ResNet performance [33]. Prior to training data sets were first random-shuffled. Each model was trained for 40 epochs with a batch size of 32 and a train-validation-test split of $80 : 10 : 10$. The Adam optimizer with a learning rate of 0.0001 was used. During training the model validation loss (MSE between actual and predicted Zernike coefficients) is calculated at each epoch to evaluate the model performance on unseen data. This learning rate was reduced by a factor of 0.5 if the validation loss began to plateau with a minimum learning rate of 1×10^{-7} permitted. Early stopping was implemented if this loss did not improve for 3 consecutive epochs. Training was carried out using Tensorflow 2.17.0 on Python 3.10.12 with a NVIDIA TESLA P100 GPU.

As shown in Fig. 2(b), the ViT model treats input data differently than a CNN by separating input images into patches that embed position and feature information. In contrast to CNNs a ViT models the global relationship between all the image patches using self-attention. Self-attention was first designed to weigh the importance of different words relative to one another in natural language processing but has since been applied in the transformer encoder of ViTs to measure which patches contain the most information [45]. A mechanism called multi-head attention is used to run self-attention in parallel so numerous features can be extracted simultaneously [46].

In this work, a projection dimension of 128 was used in the patch encoding, 6 heads were used in the multi-head attention and 6 transformer layers were used. These values were selected based on initial tests to find the lowest training time that did not significantly sacrifice accuracy. To process the data from the transformer block, a multi-layer perceptron with two input layers is used with swish activation function and units of 1024 and 512. A final output layer of 18 is used with linear activation to represent the 18 Zernike coefficients. Training time and model performance were taken into account when deciding the patch sizes for different input data as training time and memory consumption increase with patch size. A patch size of 10 used for the (200,200,2) dataset. For the downsampled datasets discussed later patch sizes of 5, 4, 2 and 2 were used for the (100,100,2), (64,64,2), (32,32,2) and (16,16,2) datasets, respectively. Additionally, min-max normalization was removed as it has been shown to influence how weight values are calculated in the multi-head attention layer, degrading performance [33]. The remaining training parameters were kept the same as those in the ResNet model.

4. Results and discussion

In order to test the training time and model performance as a function of dataset size, subsets of size 50,000, 25,000 and 10,000 of the original dataset of 100,000 concatenated images were selected. The performance of both models can be seen in Fig. 3. Each model was run 5 times for each dataset, with the resulting error (standard deviation) indicated by the shaded regions in Fig. 3. As expected, the performance of both models, in terms of the returned RMSE value, improves with the size of the dataset used for training. However, in almost all cases the CNN outperforms the ViT. This is in contrast to recent results using a ViT to analyze the simulated spot patterns of a Shack Hartman wavefront sensor [33]. In that case the ViT was shown to outperform a ResNet model in all scenarios studied. However, in that work the ViT was pretrained and only simulated data was considered. In our experiment we have not pretrained the ViT as the benefit of pretraining has been demonstrated to diminish with increased training data [38]. It has also previously been reported that ViTs require significantly longer time to train than CNNs due to the computational expense of the self attention mechanism [38]. However, in Fig. 3 it is seen that the training time for the ViT and CNN is almost identical for each dataset tested.

The results shown in Fig. 3 suggest that for dual-plane DLWFS there is no benefit in using a ViT rather than the CNN architectures typically employed for these applications. However, previous work has shown that the performance of DLWFS using CNNs is sensitive to the resolution of the recorded intensity distributions [21]. To investigate this for the case of a ViT each of the images in the measured $100,000$ concatenated dataset was downsampled from $(200, 200, 2)$ to $(100, 100, 2)$, $(64, 64, 2)$, $(32, 32, 2)$, and $(16, 16, 2)$. An example of the effect of downsampling for the in-focus image is shown in Fig. 4. For each downsampled dataset the ViT and CNN models were retrained using the same parameters described previously for the $(200, 200, 2)$ images.

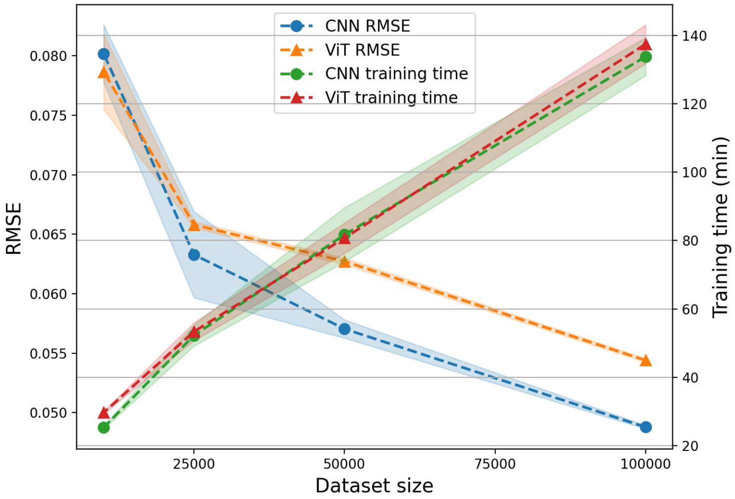

Fig. 3. RMSE and training time versus dataset size for the (200, 200, 2) input data using ViT and CNN.

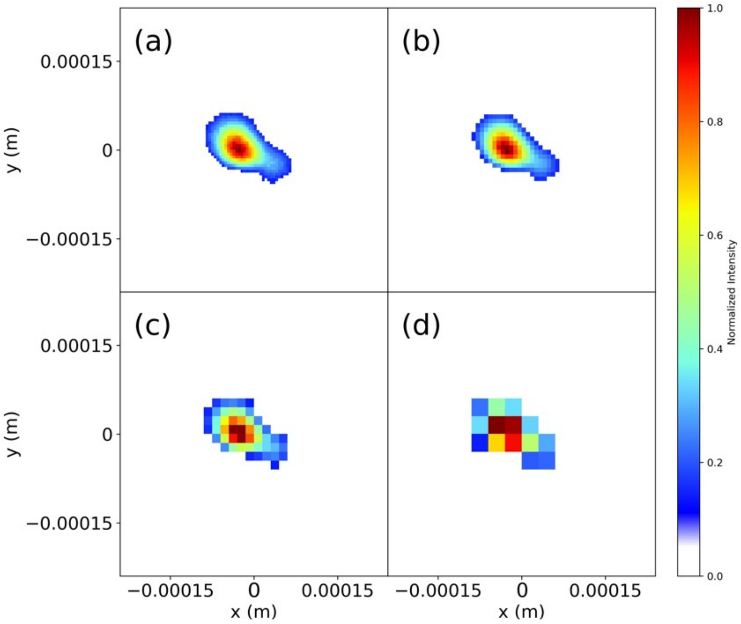

Fig. 4. Downsampled in-focus image for resolutions (a) (100,100), (b) (64,64), (c) (32,32) and (d) (16,16).

The performance of both models for different degrees of downsampling is shown in Fig. 5(a). As might be expected, for both models the RMSE is seen to increase as the images are downsampled, with the RMSE in both models increasing noticeably below (64, 64, 2). However, for the (32, 32, 2) and (16, 16, 2) downsampled datasets the ViT outperforms the CNN, exhibiting a lower RMSE. To see if this is a robust feature, the performance of the ViT and CNN models was also compared for dataset sizes 10,000, 25,000 and 50,000. As can be seen in the inset of Fig. 5(a) the ViT outperforms the CNN in all cases when downsampled below (64, 64, 2). The exception to this is for the 10,000 datatset size, which is too small for reliable ViT training, as indicated by the significantly larger RMSE and errors.

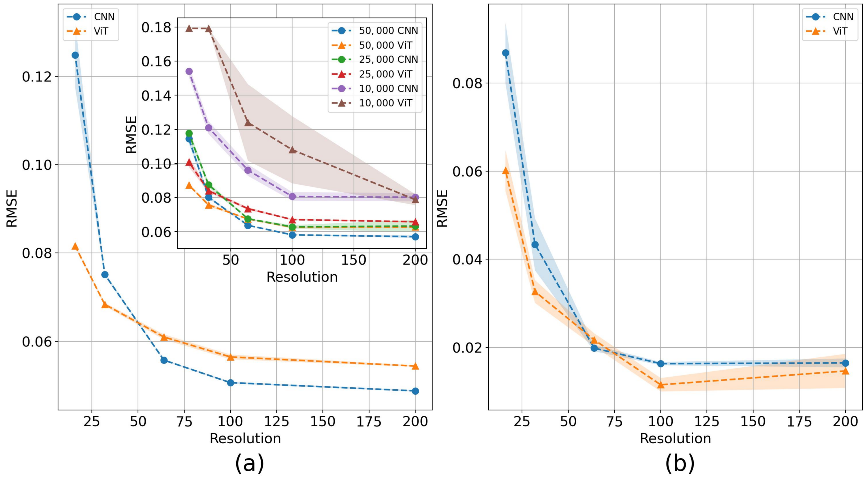

Fig. 5. Comparison of model performance against downsampling for (a) experimental and (b) simulated data.

4.1. Simulation method

The performance of both ViTs and CNNs has previously been demonstrated to depend sensitively on noise [38]. The images gathered in the dual-plane DLWFS presented here are subject to noise in the form of camera noise, as well as unwanted back-reflections and scatter. The effect of this unwanted light on the recorded data can be seen most clearly in the out-of-focal-plane intensity

Table 1. Zernike coefficients used to create the simulated and experimental beams shown in Fig. 6.

Aberration	**Coefficient**	Aberration	**Coefficient**
Defocus	-0.49	Secondary 45 Degree Astigmatism	0.24
45-Degree Astigmatism	-0.12	X-Tetrafoil	-0.24
0-Degree Astigmatism	-0.37	Y-Tetrafoil	-0.03
Y-Coma	0.33	Secondary X-Coma	-0.14
X-Coma	0.18	Secondary Y-Coma	-0.00
Y-trefoil	0.08	Secondary X-Trefoil	-0.14
X-trefoil	-0.02	Secondary Y-Trefoil	-0.15
Spherical	0.22	X-Pentafoil	-0.12
Secondary 0-Degree Astigmatism	-0.08	Y-Pentafoil	-0.05

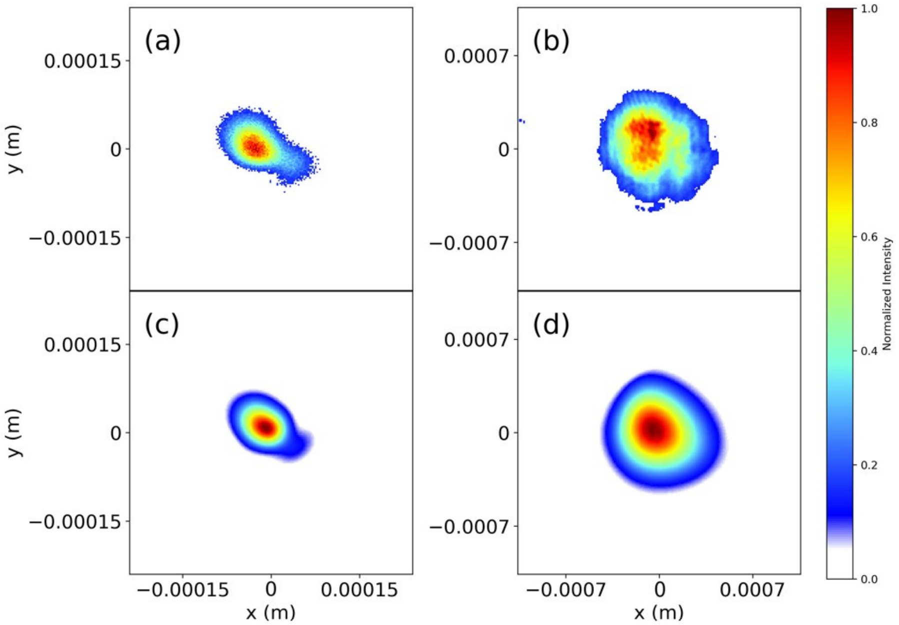

Fig. 6. (a) Measured and (c) simulated in-focus and (b) measured and (d) simulated out-of-focus intensity profiles for a beam with Zernike coefficients shown in Table 1.

distributions, an example of which is shown in Fig. 6(b). In order to assess whether the difference in performance between the ViT and CNN models was due to noise, a simulation of the DLWFS was developed.

The same procedure as described for the experimental data was used to generate a random set of the first 18 Zernike polynomials (excluding piston, tip and tilt). We note that the Zernike coefficients used to generate the simulated data are not identical to those used to generate the experimental data, allowing the models to be tested on two distinct but related data sets, one with and one without noise. These were then used as aberrations on a beam with Gaussian transverse spatial profile and wavelength 633 nm located at the front focal plane of a lens of focal length 500 mm. The aberrated beam was then propagated to the back focal plane via a Fourier transform, such that the propagation grid size matched the finite size of the camera pixels and the resulting normalized intensity distribution was recorded. The field in the out-of-focus-plane was then calculated by forward propagating the field from the back focal plane a distance 7 cm using the angular spectrum method. The field was then resampled such that the grid size matched the size of the camera pixels at this plane, and the normalized intensity distribution was recorded. The resulting focal-plane and out-of-focal-plane images were then concatenated to make an image of shape (200,200,2). This procedure was repeated until a dataset of 100,000 noise-free concatenated images was obtained.

Figure 6 shows an example of the simulated and experimental in-focus and out-of-focus intensity distributions. For both experimental and simulated beams, the Zernike coefficients shown in Table 1 were used, with very good agreement observed between the simulated and experimental intensity distributions.

Following the same procedure as for the experimental data, the complete simulated dataset was then downsampled from (200, 200, 2) to (100, 100, 2), (64, 64, 2), (32, 32, 2), and (16, 16, 2) to produce 5 simulated datasets each consisting of 100,000 noise-free concatenated images. Both ViT and CNN models were then trained on each of these datatsets 5 times. The resulting performance of the models is shown in Fig. 5(b) for different levels of downsampling. As expected, due to the absence of noise, the RMSE returned by both models using the simulated datasets is lower than for the corresponding experimental datasets. It is also seen, consistent with the behaviour observed in the experimental data, that the ViT outperforms the CNN when images are downsampled below $(64, 64, 2)$. The results of these simulations demonstrate that the ability of the ViT to outperform the CNN model when trained using lower resolution images is not due to a difference in the noise sensitivities of the models, rather a feature of their fundamentally different architectures.

So far, only the total RMSE across all Zernike modes has been considered when comparing the performance of both models. To examine the data in more detail the RMSE per Zernike mode was also investigated. Fig. 7 shows the RMSE per Zernike mode for each of the experimental and simulated datasets using both models. From the results for the experimental datasets $(200, 200, 2)$, $(100, 100, 2)$ and $(64, 64, 2)$, shown in Fig. 7(a), (c) and (e), respectively, it is seen that the CNN performs better than the ViT across all the Zernike modes considered, although this outperformance diminishes with increased downsampling. It is also seen that both models exhibit similar variation of the RMSE across all of the Zernike modes considered.

However, when the experimental datasets are downsampled to $(32, 32, 2)$ a different behaviour emerges, with the ViT achieving a lower RMSE that the CNN for the two highest order Zernike modes considered (20 and 21). For orders lower than this the same behaviour as observed previously is seen, with the CNN providing a slightly better performance than the ViT and both models showing similar variation across Zernike modes. This effect becomes more significant when downsampling is increased to $(16, 16, 2)$. In this case the ViT achieves a significantly lower RMSE than the CNN for all Zernike modes greater than 11. Below this order the RMSE value and variation across modes is similar for both models.

For the simulated datasets the behaviour is initially more complex. For the $(200, 200, 2)$, $(100, 100, 2)$ and $(64, 64, 2)$ datatsets, shown in Fig. 7(b), (d) and (f), respectively, the ViT yields a lower RMSE than the CNN for several Zernike modes, despite the total RMSE being greater than for the CNN, as shown in Fig. 5. However, as the data is downsampled to $(32, 32, 2)$ and $(16, 16, 2)$ it is clear that for the higher order modes the ViT achieves a significantly lower RMSE than the CNN, mirroring the behaviour observed in the experimental dataset. Both the experimental and simulation results indicate that the ViT's ability to outperform the CNN, when data is downsampled significantly, stems from the former's ability to more accurately predict the contributions from the higher-order Zernike polynomials.

The training time for both models, for both experimental and simulated data, was also investigated for different levels of downsampling, and the results are summarized in Table 2. In each case, the full dataset of 100,000 concatenated images was used. It can be seen that there is considerable variation in training time between the models, with the ViT taking up to approximately twice a long to train as the CNN for the experimental data, and up to approximately four times as long for the simulated data, depending on the level of downsampling. However, the training time of the ViT depends heavily on patch size, and it is possible that this could be reduced by optimizing this parameter. The inference time, taken as the average per-image inference time measured over the 10,000-image test set, is also shown. We note that these times are for the simulation of the uniform intensity beam discussed below. For both the ViT and CNN the inference time decreases with image size, with similar values from both models.

The results presented so far have used a beam with a Gaussian transverse profile and wavefront described by Zernike polynomials. However, the Zernike polynomials are only orthogonal over a unit circle, such that weighting by a Gaussian results in the loss of orthogonality and can result in cross-coupling of aberrations [47]. To assess the impact of this, 100,000 concatenated images were simulated for the case of a circular beam at the SLM with a uniform top-hat intensity profile. In this case the orthogonality of the Zernike polynomials is maintained. The same data

Optics EXPRESS

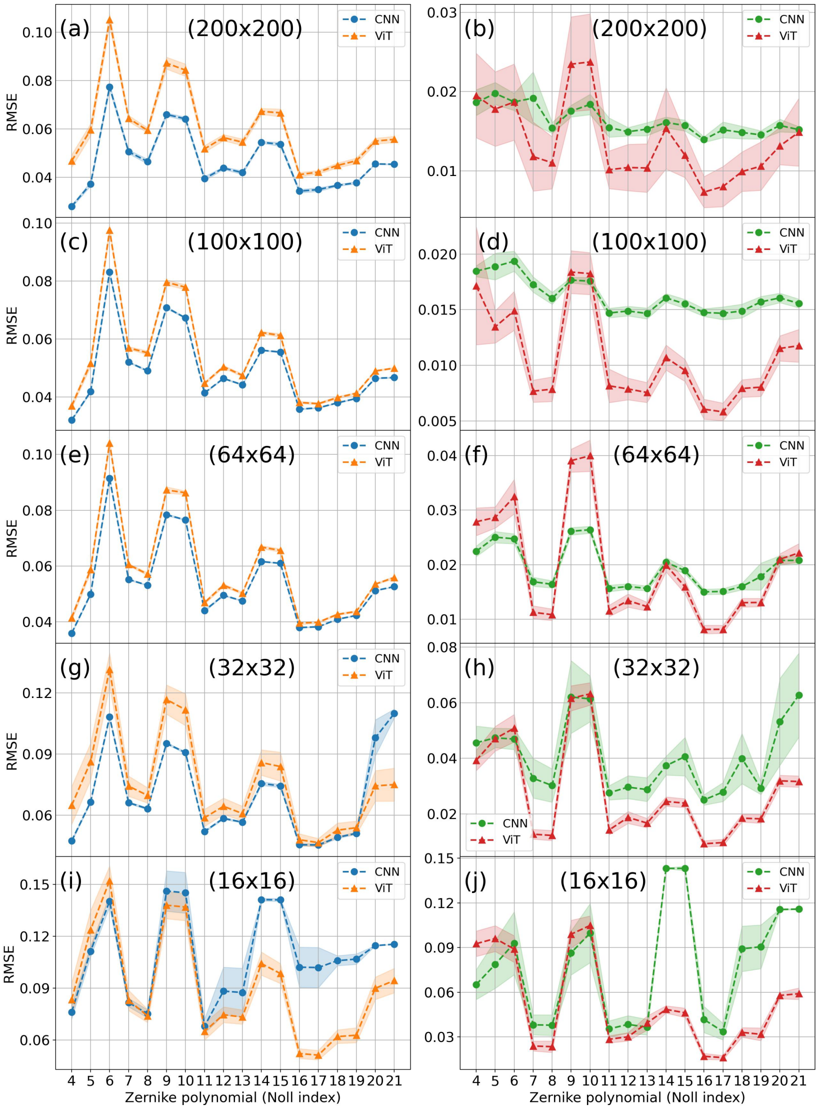

Fig. 7. RMSE per Zernike mode for CNN (circles) and ViT (triangles) models trained on 100,000 concatenated images for experimental data (left column) and simulated data (right column) at resolutions: (a) and (b) (200, 200, 2), (c) and (d) (100, 100, 2), (e) and (f) (64, 64, 2), (g) and (h), (32, 32, 2), (i) and (j) (16, 16, 2).

Table 2. Training times for experimental and simulated data and average per-image inference time for simulated data.

Model	Image Size	Patch Size	Training Time	Training Time	Inference Time
			Experiment	Simulation	Simulation
	(Pixels)	(Pixels)	(Minutes)	(Minutes)	(ms)
ViT	(200, 200)	(10, 10)	137 ± 6	150 ± 30	1.5
	(100, 100)	(5, 5)	143 ± 8	160 ± 10	1.4
	(64, 64)	(4, 4)	77 ± 3	93 ± 9	0.87
	(32, 32)	(2, 2)	75 ± 4	100 ± 10	0.85
	(16, 16)	(2, 2)	47 ± 3	23 ± 4	0.75
CNN	(200, 200)		134 ± 6	130 ± 20	1.3
	(100, 100)		69 ± 4	55 ± 4	0.96
	(64, 64)		70 ± 2	55 ± 5	0.67
	(32, 32)		49 ± 2	28 ± 4	0.84
	(16, 16)		24 ± 7	24 ± 7	0.83

generation and training, validation and testing procedures as previously described were followed, with the exception that each model was run once at each resolution, rather than 5 times, as only marginal variations between runs were observed previously, as shown in Fig. 5(b).

The performance of both models, in terms of the total RMSE as a function of image resolution, is shown in Fig. 8(a). The CNN exhibits very similar behaviour to the case of Gaussian illumination (Fig. 5(b)) with a steep increase in RMSE being observed for images downsampled below $(64, 64, 2)$. However, in contrast to the case of Gaussian illumination, the performance of the ViT is now seen to be much more robust to decreasing image resolution, with only small changes in RMSE observed when images are downsampled below $(64, 64, 2)$. In Fig. 8(b) the RMSE per Zernike mode at each resolution is shown for both models. Consistent with the results shown for the case of Gaussian illumination (Fig. 7), the ViT can outperform the CNN by its ability to more accurately predict the contributions from higher-order Zernike polynomials.

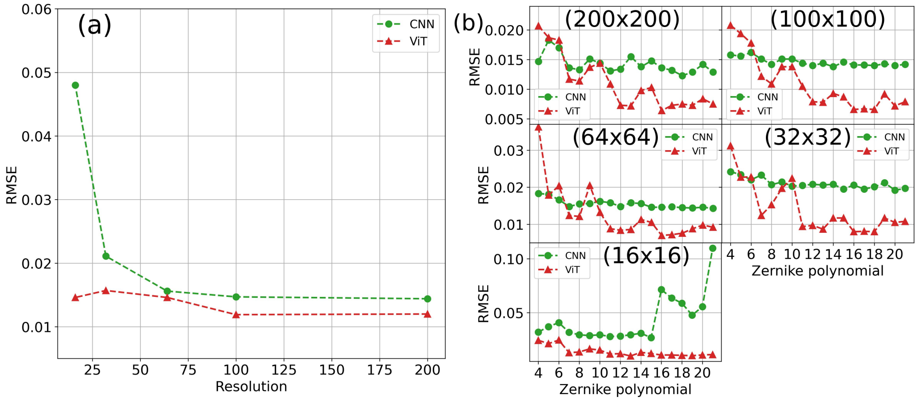

Fig. 8. (a) Comparison of model performance against downsampling for simulated data for the case of a uniformly illuminated pupil. (b) RMSE per Zernike mode for CNN and ViT models at each image resolution tested.

These results indicate that, although the ViT can outperform the more widely used CNN, for optimal performance an appropriate basis should be considered. For example, Zernike-Gauss polynomials may be more appropriate for the case of apodized pupils [47], or Zernike-like Laguerre Gaussian polynomials for elliptical geometries [48]. A detailed study of the impact of basis set on ViT performance will be the subject of a future experiment.

5. Conclusion

The development of machine learning models provides new opportunities for accelerating aberration measurement and correction in DLWFS applications, however, it is not yet clear which machine learning model is best suited for a given application. In this paper the performance of a dual-plane DLWFS employing the recently developed ViT architecture has been compared with the CNN architecture typically used in DLWFS applications. Without downsampling, the CNN outperformed the ViT in most of the scenarios investigated. However, for significant levels of downsampling the ViT could outperform the CNN due to the former's ability to more accurately predict the contributions from higher order Zernike polynomials. This behaviour was also observed in the simulation, demonstrating that this feature is intrinsic to the models rather than a consequence of noise sensitivity. A challenge of this setup, common with other DLWFS methods, is that it has been trained for a specific experimental geometry. Changes to focussing conditions, or significant misalignments, would likely require retraining. However, our results indicate that ViTs may have an advantage over CNNs in cases where data is significantly downsampled, for example, to boost training and inference times [21]. In this work the optimization of the ViT parameters was not investigated in detail, suggesting that further improvements could be achieved through hyperparameter tuning and use of appropriate basis sets. Future work will also focus on pretraining the ViT to further enhance performance, as well as the use of recently developed hybrid ViT-CNN [37], which leverage the best features of both models.

Funding. Engineering and Physical Sciences Research Council (EP/W524694/1).

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data and code underlying the results presented in this paper are available on Zenodo at [49].

References

1. J. W. Hardy, *Adaptive optics for astronomical telescopes*, vol. 16 (Oxford university press, 1998).
2. E. Brunner, J. Shatokhina, M. F. Shirazi, *et al.*, "Retinal adaptive optics imaging with a pyramid wavefront sensor," Biomed. Opt. Express **12**(10), 5969–5990 (2021).
3. J. Scrimgeour and J. E. Curtis, "Aberration correction in wide-field fluorescence microscopy by segmented-pupil image interferometry," Opt. Express **20**(13), 14534–14541 (2012).
4. M. Booth, "Adaptive optical microscopy: The ongoing quest for a perfect image," Light: Sci. Appl. **3**(4), e165 (2014).
5. B. C. Platt and R. Shack, "History and principles of shack-hartmann wavefront sensing," J Refract Surg **17**(5), S573 (2001).
6. H. I. Campbell and A. H. Greenaway, "Wavefront Sensing: From Historical Roots to the State-of-the-Art," in *EAS Publications Series, vol. 22 of EAS Publications Series* M. Carbillet, A. Ferrari, and C. Aime, eds. (2006), pp. 165–185.
7. R. Ragazzoni, "Pupil plane wavefront sensing with an oscillating prism," J. Mod. Opt. **43**(2), 289–293 (1996).
8. D. Malacara, *Optical Shop Testing* (Wiley-Interscience, 2007), 3rd ed.
9. R. W. Gerchberg, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik **35**, 237–246 (1972).
10. J. R. Fienup, "Reconstruction of an object from the modulus of its fourier transform," Opt. Lett. **3**(1), 27–29 (1978).
11. A. Give'on, N. J. Kasdin, R. J. Vanderbei, *et al.*, "Stochastic optimal phase retrieval algorithm for high-contrast imaging," in *Astronomical Adaptive Optics Systems and Applications, vol. 5169 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* R. K. Tyson and M. Lloyd-Hart, eds. (2003), pp. 276–287.
12. Y. Li, Y. Xue, and L. Tian, "Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media," Optica **5**(10), 1181–1190 (2018).
13. Y. Sun, J. Shi, L. Sun, *et al.*, "Image reconstruction through dynamic scattering media based on deep learning," Opt. Express **27**(11), 16032–16046 (2019).
14. Q. Li, J. Zhao, Y. Zhang, *et al.*, "Imaging reconstruction through strongly scattering media by using convolutional neural networks," Opt. Commun. **477**, 126341 (2020).
15. J. You, J. Gu, Y. Du, *et al.*, "Atmospheric turbulence aberration correction based on deep learning wavefront sensing," Sensors **23**(22), 9159 (2023).

16. Q. Tian, C. Lu, B. Liu, *et al.*, "Dnn-based aberration correction in a wavefront sensorless adaptive optics system," Opt. Express **27**(8), 10765–10776 (2019).
17. Y. Jin, Y. Zhang, L. Hu, *et al.*, "Machine learning guided rapid focusing with sensor-less aberration corrections," Opt. Express **26**(23), 30162–30171 (2018).
18. Y. Nishizaki, M. Valdivia, R. Horisaki, *et al.*, "Deep learning wavefront sensing," Opt. Express **27**(1), 240–251 (2019).
19. P.-O. Vanberg, "Machine learning for image-based wavefront sensing," Master's thesis, University of Liége (2019).
20. M. Liu, D. N. Lopez, and G. C. Spalding, "Experimental implementation of wavefront sensorless real-time adaptive optics aberration correction control loop with a neural network," in *Emerging Topics in Artificial Intelligence 2020*, vol. 11469 G. Volpe, J. B. Pereira, D. Brunner, and A. Ozcan, eds., International Society for Optics and Photonics (SPIE, 2020), p. 114691S.
21. E. Vera, F. Guzmán, and C. Weinberger, "Boosting the deep learning wavefront sensor for real-time applications [invited]," Appl. Opt. **60**(10), B119–B124 (2021).
22. A. B. Siddik, S. Sandoval, D. Voelz, *et al.*, "Deep learning estimation of modified zernike coefficients and recovery of point spread functions in turbulence," Opt. Express **31**(14), 22903–22913 (2023).
23. Y. E. Kok, A. Bentley, A. Parkes, *et al.*, "Direct zernike coefficient prediction from point spread functions and extended images using deep learning," arXiv (2024).
24. J. B. Shohani, M. Hajimahmoodzadeh, and H. Fallah, "Using a deep learning algorithm in image-based wavefront sensing: determining the optimum number of zernike terms," Opt. Continuum **2**(3), 632–645 (2023).
25. B. Chen, Y. Zhou, J. Jia, *et al.*, "Adaptive optical closed-loop control on the basis of hyperparametric optimization of convolutional neural networks," Appl. Sci. **13**(15), 8589 (2023).
26. T. E. Andersen, M. Owner-Petersen, and A. Enmark, "Image-based wavefront sensing for astronomy using neural networks," J. Astron. Telesc. Instrum. Syst. **6**(03), 1 (2020).
27. C. Lu, Q. Tian, L. Zhu, *et al.*, "Mitigating the ambiguity problem in the cnn-based wavefront correction," Opt. Lett. **47**(13), 3251–3254 (2022).
28. M. Quesnel, G. Orban de Xivry, G. Louppe, *et al.*, "A deep learning approach for focal-plane wavefront sensing using vortex phase diversity," Astronomy Astrophysics **668**, A36 (2022).
29. L. Möckl, P. N. Petrov, and W. E. Moerner, "Accurate phase retrieval of complex 3d point spread functions with deep residual neural networks," Appl. Phys. Lett. **115**(25), 251106 (2019).
30. H. Ma, H. Liu, Y. Qiao, *et al.*, "Numerical study of adaptive optics compensation based on convolutional neural networks," Opt. Commun. **433**, 283–289 (2019).
31. Y. Wu, Y. Guo, H. Bao, *et al.*, "Sub-millisecond phase retrieval for phase-diversity wavefront sensor," Sensors **20**(17), 4877 (2020).
32. J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," Appl. Sci. **13**(9), 5521 (2023).
33. Q. Zhang, H. Zuo, X. Cui, *et al.*, "Automatic compressive sensing of shack–hartmann sensors based on the vision transformer," Photonics **11**(11), 998 (2024).
34. K. Hu, D. Sun, and Y. Zhao, "Enhanced single-frame interferometry via hybrid conv-transformer architecture for ultra-precise phase retrieval," Opt. Express **32**(17), 30226–30241 (2024).
35. Z. Zhao, M. Zhou, Y. Du, *et al.*, "Robust phase unwrapping algorithm based on zernike polynomial fitting and swin-transformer network," Meas. Sci. Technol. **33**, 055002 (2022).
36. H. Chen, H. Zhang, Y. He, *et al.*, "Direct wavefront sensing with a plenoptic sensor based on deep learning," Opt. Express **31**(6), 10320–10332 (2023).
37. H. Kou, J. Gu, J. You, *et al.*, "Single-shot wavefront sensing in focal plane imaging using transformer networks," Optics **6**(1), 11 (2025).
38. X. Liu, W. Luo, P. Hu, *et al.*, "Transformer-based wavefront sensing for atmospheric turbulence aberration correction," Appl. Opt. **64**(10), 2451–2463 (2025).
39. K. Niu and C. Tian, "Zernike polynomials and their applications," J. Opt. **24**(12), 123001 (2022).
40. V. Lakshminarayanan and A. Fleck, "Zernike polynomials: a guide," J. Mod. Opt. **58**(7), 545–561 (2011).
41. R. J. Noll, "Zernike polynomials and atmospheric turbulence*," J. Opt. Soc. Am. **66**(3), 207–211 (1976).
42. M. Lamb, D. R. Andersen, J.-P. Véran, *et al.*, "Non-common path aberration corrections for current and future AO systems," in *Adaptive Optics Systems IV*, vol. 9148 E. Marchetti, L. M. Close, and J.-P. Véran, eds., International Society for Optics and Photonics (SPIE, 2014), p. 914857.
43. X. Du, Y. Sun, Y. Song, *et al.*, "A comparative study of different cnn models and transfer learning effect for underwater object classification in side-scan sonar images," Remote Sens. **15**(3), 593 (2023).
44. K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," arXiv (2015).
45. A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," arXiv (2023).
46. A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations* (2021).
47. V. N. Mahajan, "Zernike circle polynomials and optical aberrations of systems with circular pupils," Appl. Opt. **33**(34), 8121 (1994).
48. B. D. Strycker, "Zernike-like laguerre–gaussian orthonormal polynomials for optical field reconstruction," Opt. Lett. **47**(23), 6137–6140 (2022).
49. E. O Rourke and K. O Keeffe, "Dual-plane wavefront sensing using a vision transformer," Zenodo (2025), doi.org/10.5281/zenodo.18268629.