Published in *Proceedings of the SPIE Photonics West 2025*, San Francisco, USA, 25-30 January 2025, which should be cited to refer to this work.
Link: https://spie.org/photonics-west/presentation/High-speed-AI-image-space-wavefront-sensing-using-embedded-computing/13373-25

High-speed AI image space wavefront sensing using embedded computing: achieving 1000 frames per second

Gaston Baudat[1], David Lavanchy[2], Guillaume Müller[3]

[1]Innovations Foresight, LLC., 4432 Mallard Point, Columbus Indiana, IN. 47201, USA
[1]Adjunct Research Professor, Wyant College of Optical Sciences, The University of Arizona, 1630 E. University Blvd., Tucson, Arizona 85721, USA
[2]School of Engineering and Management Vaud, HES-SO University of Applied Sciences and Arts Western Switzerland. Route de Cheseaux 1, Case postale 521 CH-1401 Yverdon-les-Bains, Switzerland
[3]School of Engineering and Management Vaud, HES-SO University of Applied Sciences and Arts Western Switzerland. Route de Cheseaux 1, Case postale 521 CH-1401 Yverdon-les-Bains, Switzerland

Abstract

Wavefront sensing is fundamental to numerous scientific and industrial applications in optics, particularly in adaptive optics (AO). This paper presents a high-speed image space AI-powered wavefront sensing system, based on a patented technology known as AI4Wave, that eliminates the need for dedicated wavefront sensors, such as Shack-Hartmann, relying solely on fast cameras and AI edge computing devices, specifically an NVIDIA Jetson embedded module in our implementation. This module could also be utilized in the future to directly control actuators, streamlining the entire AO process. The system captures defocused images of a point source or similar, which are processed by a feedforward artificial neural network (NN) trained exclusively on synthetic data [1]. This approach enables real-time phase retrieval, overcoming the local derivative limitations of Shack-Hartmann sensors and handling large wavefront or surface errors spanning multiple wavelengths. It also provides a unique capability for field-dependent wavefront measurement from a single image of multiple sources. Yu Wu et al. [2] proposed a sub-millisecond phase retrieval method that requires two simultaneous images for optimal performance and NN training on physical data. However, this approach necessitates retraining whenever system parameters, such as f/#, wavelength, or pixel size, are altered. In contrast, AI4Wave employs synthetic, normalized data that is agnostic to most optical layout changes, enabling robust and versatile training and validation. Combined with off-the-shelf NVIDIA hardware, it offers a ready-to-use solution for industrial deployment. By integrating optimized algorithms within modern AI acceleration frameworks, such as NVIDIA TensorRT, and leveraging advanced AI edge computers, the system achieves processing speeds of 1000 frames per second or more, effectively providing sub-millisecond wavefront sensing capability. The minimal hardware setup, consisting only of a camera and an NVIDIA module, supports various optical layouts and surface measurement setups. The proposed approach is entirely deterministic. AI4Wave phase retrieval leverages feedforward neural network computation, eliminating the need for iterations. This approach ensures consistent results and computation times, free from real-time optimization challenges or issues related to local minima. Preliminary tests demonstrate high accuracy and repeatability, with exposure times as short as 24 µs, effectively capturing environmental perturbations and vibrations. This work provides an industrial-grade, reliable foundation for high-speed, high-fidelity wavefront sensing in applications such as AO and similar systems. It also includes precise offline benchmarking of wavefront errors, based on millions of samples to ensure deterministic dependable performance.

1. INTRODUCTION

Image space wavefront sensing (IS-WFS) techniques are methods for determining an optical wavefront directly from image data (irradiance), without the need for specialized wavefront sensors or interferometers. These techniques belong to the field of computational optics leveraging image processing algorithms to interpret the phase information encoded in images, offering a flexible and cost-effective alternative to traditional wavefront sensing methods.

Typical IS-WFS approaches include:

1. **Phase Retrieval:** Reconstructs the wavefront phase by iteratively refining its estimation to match the observed image, often using algorithms such as Gerchberg-Saxton (GS) or hybrid input-output methods. Typically, a single

image is sufficient since the nature of the source is known. Some level of a-priori phase modulation (typically defocus) is required to solve the problem without any ambiguity. The required initial condition impacts the performance.

2. **Phase Diversity**: Involves capturing multiple images with known phase modulations and jointly solving for the wavefront phase that explains all observed images. In this method, the nature of the source may be unknown, requiring several images under different modulations to achieve accurate reconstruction. The required initial condition impacts the performance.
3. **Curvature Sensing (CS)**: Although the methods described in 1 and 2 are non-parametric by nature, the CS approach is parametric, solving the irradiance transport equation. Roddier and Roddier [3] successfully applied this method in adaptive optics for astronomy. With known sources, typically plane waves, a single image suffices for phase retrieval, though two images (intra- and extra-focal) are often used to compensate for scintillation.

The main advantages of IS-WFS include its ability to integrate easily with existing imaging systems and the elimination of the need for additional wavefront sensing hardware. This makes IS-WFS particularly attractive for applications where space, weight, and cost constraints are critical, such as in space telescopes and portable optical devices. The absence of a reference beam and interferometry simplifies the system, enhancing robustness to vibrations, noise, and turbulence. Challenges remain, particularly in managing the computational load required for real-time processing. Single image WF sensing using phase retrieval algorithms from a known source typically relies on iterative error reduction techniques, such as the Gerchberg-Saxton (GS) algorithm. The GS algorithm converges slowly, often with plateaus where the error remains nearly constant over many iterations. Iterative joint maximum a posteriori estimation in phase diversity (multiple images) and Poisson solvers for CS typically produce approximations rather than exact solutions, with accuracy relying on initial conditions. Traditional IS-WFS approaches are inherently iterative, relying on nonlinear optimizations that often lead to suboptimal solutions (local minima) and unpredictable runtimes dependent on the initial guess. These limitations are critical for high-rate real-time wavefront sensing applications, such as adaptive optics, where rapid solution times are essential. Machine learning provides an alternative for IS-WFS, eliminating the need for time-consuming iterative processing at runtime. Typically, an artificial NN learns the inverse function mapping the image irradiance to the wavefront phase.

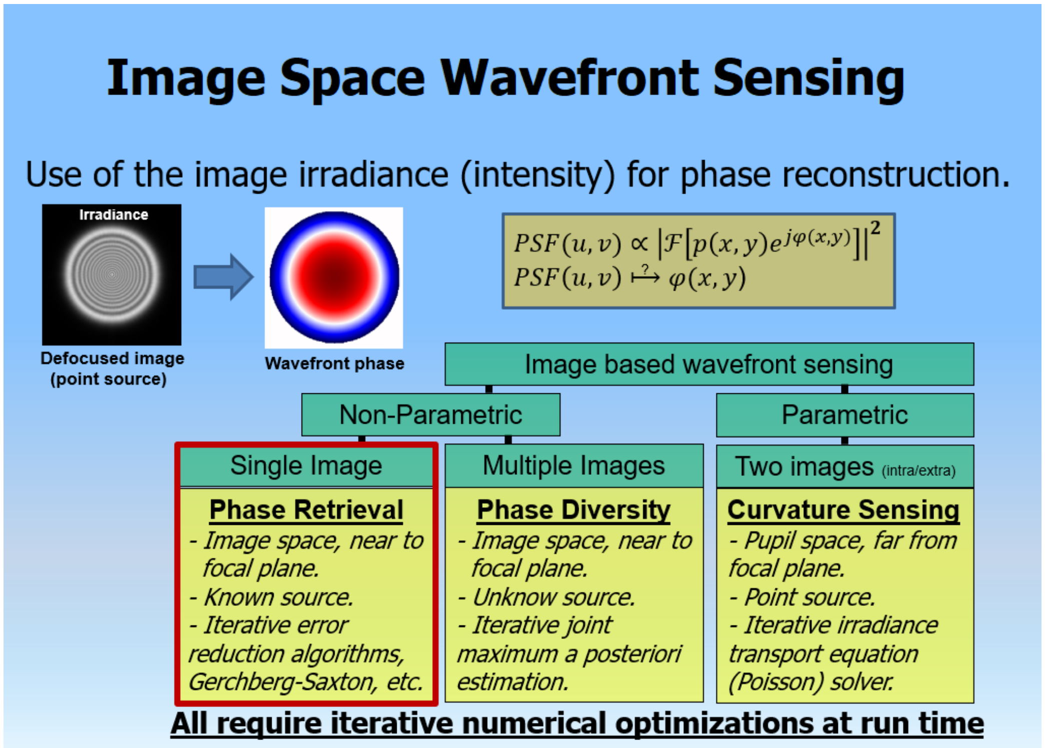


Image Space Wavefront Sensing

Use of the image irradiance (intensity) for phase reconstruction.

$$PSF(u,v) \propto \left|\mathcal{F}\left[p(x,y)e^{j\varphi(x,y)}\right]\right|^2$$
$$PSF(u,v) \overset{?}{\longmapsto} \varphi(x,y)$$

Image based wavefront sensing

Non-Parametric

Parametric

Single Image

Phase Retrieval

- *Image space, near to focal plane.*
- *Known source.*
- *Iterative error reduction algorithms, Gerchberg-Saxton, etc.*

Multiple Images

Phase Diversity

- *Image space, near to focal plane.*
- *Unknow source.*
- *Iterative joint maximum a posteriori estimation.*

Two images (intra/extra)

Curvature Sensing

- *Pupil space, far from focal plane.*
- *Point source.*
- *Iterative irradiance transport equation (Poisson) solver.*

All require iterative numerical optimizations at run time

Figure 1. Common approaches used in the context of image-space wavefront sensing

An IS-WFS approach using artificial intelligence was proposed by Baudat [1], enabling phase retrieval from a single defocused image, even under noisy conditions. This patented method, known as AI4Wave, offers four key advantages:

- Requires only a single image.
- Training and validation rely solely on synthetic (simulated) data.
- Fast, non-iterative and deterministic processing at runtime.
- Whole-field wavefront sensing.

The AI4Wave method relies on an artificial NN trained exclusively on synthetic data, consisting of millions of simulated images. This enables the creation of extensive training and validation datasets for accurately solving the inverse problem of mapping image irradiance to phase. Aberrated and defocused images are simulated using scalar diffraction theory, and the NN processes normalized images [1], ensuring independence from system parameters such as wavelength, aperture, focal length, and pixel size. This allows a single NN to function across diverse applications without retraining, as input images are normalized beforehand. Originally developed for telescope alignment and active optics in astronomy, AI4Wave has also been applied to optical metrology [4], [5]. The NN approximates a unique mathematical function mapping single defocused images to their corresponding wavefronts, typically represented as Zernike annular polynomial coefficients. During runtime, the NN's non-iterative feed-forward structure eliminates the need for iterations, optimizations, or convergence concerns, enabling fast, real-time phase retrieval with deterministic performance and timing. This method inherently supports whole-field wavefront sensing from a single defocused image including several sources (like with star field). While the scope of this paper focuses on monochromatic sources, extending the approach to polychromatic sources is straightforward by calculating diffraction at multiple wavelengths and combining results via quadrature summation. Monochromatic calculations remain relevant for filters with bandwidths around 100 nm. Baudat and Hayes demonstrated strong agreement between this technique and traditional interferometric methods [6]. Yu Wu et al. [2] proposed a sub-millisecond phase retrieval method for phase-diversity wavefront sensing, requiring simultaneous acquisition of two images for optimal performance. Their NN is trained on physical data acquired with a spatial light modulator (SLM) in a double-pass setup, necessitating retraining if system parameters like camera pixel size, f/#, or wavelength change. In contrast, AI4Wave relies entirely on synthetic, normalized data, making it independent of most changes in the optical layout. This approach enables the creation of large training and validation datasets without relying on physical systems, ensuring robust sampling of the inverse function and reliable validation. This, combined with easily available standard hardware like an off-the-shelf NVIDIA module, delivers a ready-to-use solution for deployment in real-world industrial applications.

2. METHODOLOGY

2.1 AI4Wave background and concept

The far field, intensity distribution of the monochromatic point spread function (PSF) is proportional to the square modulus of the two-dimensional Fourier transform of the complex pupil function:

$$\mathrm{PSF}(u,v) \propto |\mathcal{F}_{2D}\{\mathrm{P}(x,y)\}|^2 \tag{1}$$

The complex pupil function $\mathrm{P}(x,y)$ contains information about the shape of the pupil, the transmission function and optical phase in the pupil. In general, the complex pupil function is defined as follows:

$$\mathrm{P}(x,y) = \mathrm{p}(x,y)e^{j\varphi_p(x,y)} \tag{2}$$

where: $\mathrm{p}(x,y)$ = pupil amplitude transmission function,
$\varphi_p(x,y)$ = the pupil phase function = $2\pi W(x,y)/\lambda$, and
$W(x,y)$ = wavefront departure from the reference sphere.

With the required scaling factors, the radial PSF becomes:

$$\mathrm{PSF}(\mathrm{r}') = \frac{\mathrm{I}_0}{(\lambda\mathrm{f})^2}\left|\mathrm{T}(\rho)\right|^2_{\rho=\mathrm{r}'/\lambda\mathrm{f}} \tag{3}$$

where:

I_0 = the irradiance in power/area incident on the pupil,
λ = operating wavelength,
f = focal length,
r'= radial coordinate in the focal plane,
$\mathrm{T}(\rho) = \mathcal{F}\{\mathrm{P}(\mathrm{r})\}$ = Radial Fourier transform (Hankel transform) of the complex pupil function.

Using Equation (3), the monochromatic PSF for any pupil phase distribution can be readily computed. Baudat [1] demonstrated the normalization and sampling of input data for broad applicability across optical systems with the same NN. Radial annular Zernike polynomials form an infinite orthogonal basis of balanced aberrations. The number of terms and the polynomials retained in the subset for the Zernike expansion are typically tailored to specific applications. These polynomials are particularly well-suited (but not limited to) for circular apertures, including cases with a central obstruction, if present. In this paper, we focus on low-order polynomials commonly involved in alignment tasks and stress detection in optical surfaces. These include tilt, defocus, astigmatism, coma, trefoil, and primary spherical aberration, amounting to a total of 10 Zernike polynomials. Additional polynomials can be easily incorporated depending on application requirements. AI4Wave has been successfully applied to Zernike polynomials up to the 8th radial order in some contexts. Table 2 lists the 10 polynomials used for training the NN in this study. The subsequent experiments discussed in this paper use circular apertures without any obstruction, therefore $\epsilon = 0$. It should be noted that all phase information is contained within the defocused image of a point source, except for the piston term (z0 = 0), as absolute phase cannot be accessed without a reference beam and interference. This implies that not only higher-order Zernike polynomials can be included, but also that surface roughness information is inherently encoded in the image. Accessing such information is directly related to the spatial frequency range and signal-to-noise ratio (SNR) of the image.

Name	**Index**	**Radial Annular Polynomials** $0 < r \leq 1 \qquad 0 < \epsilon \leq 1 \qquad 0 \leq \theta \leq 2\pi$
Horizontal tilt (X)	Z_1	$\left(2r/\sqrt{1+\epsilon^2}\right)\cos(\theta)$
Vertical tilt (Y)	Z_2	$\left(2r/\sqrt{1+\epsilon^2}\right)\sin(\theta)$
Defocus	Z_3	$(2r^2 - 1 - \varepsilon^2)/\,(1-\varepsilon^2)$
Vertical astigmatism	Z_4	$\left(r^2/\sqrt{1+\epsilon^2+\epsilon^4}\right)\cos(2\theta)$
Oblique astigmatism	Z_5	$\left(r^2/\sqrt{1+\epsilon^2+\epsilon^4}\right)\sin(2\theta)$
Horizontal coma	Z_6	$[3r^3(1+\epsilon^2) - 2r(1+\epsilon^2+\epsilon^4)]/\left[(1-\epsilon^2)\sqrt{(1+\epsilon^2)(1+4\epsilon^2+\epsilon^4)}\right]\cos(\theta)$
Vertical coma	Z_7	$[3r^3(1+\epsilon^2) - 2r(1+\epsilon^2+\epsilon^4)]/\left[(1-\epsilon^2)\sqrt{(1+\epsilon^2)(1+4\epsilon^2+\epsilon^4)}\right]\sin(\theta)$
Primary spherical	Z_8	$(6r^4 - 6r^2(1+\epsilon^2) + 1 + 4\epsilon^2 + \epsilon^4)/\,(1-\epsilon^2)^2$
Oblique trefoil	Z_9	$\left[r^3/\sqrt{1+\epsilon^2+\epsilon^4+\epsilon^6}\right]\cos(3\theta)$
Vertical trefoil	Z_{10}	$\left[r^3/\sqrt{1+\epsilon^2+\epsilon^4+\epsilon^6}\right]\sin(3\theta)$

Table 1. Definition of the Z1 to Z10 annular Zernike terms using the Wyant-Creath numbering convention, following Mahajan's definitions. These are the basic common terms used for generating synthetic data and for training the AI system described here with $\varepsilon = 0$ for the experiments. In some context higher orders terms have been considered up to the Zernike radial order 8.

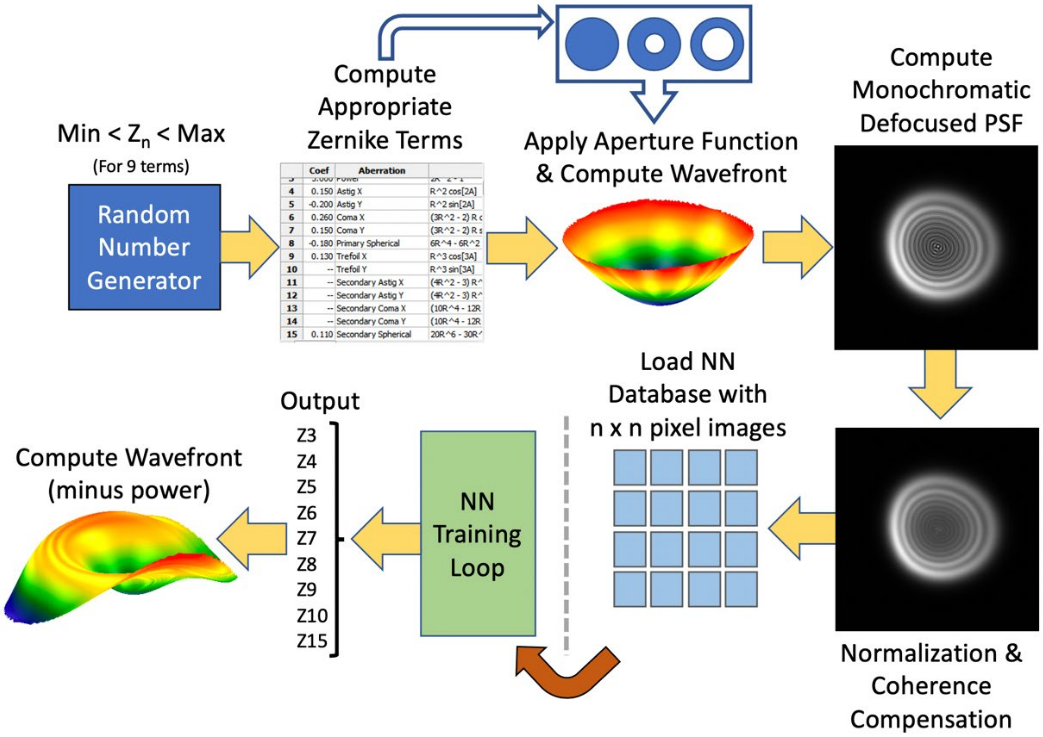
2.2 Data generation and training

The synthetic data, specific to a given optical imaging system, is computed using scalar diffraction theory. In this paper, we operate well within the far-field regime where the Fraunhofer approximation applies. Consequently, we use the Fast Fourier Transform (FFT) to model the defocused image. To train the artificial NN, defocused images are generated across predefined ranges of Zernike coefficients, tailored to the application and aberration budgets. Alternative strategies could involve directly sampling an aberrated wavefront, with the NN outputting the sampled wavefront itself rather than Zernike coefficients or other aberration terms. In this paper, we adopt a parametric approach, describing the wavefront error using radial annular Zernike polynomials.

The NN training phase employs three datasets:

a) Learning Dataset: Used to optimize NN synaptic weights, the largest datasets range from 500,000 to several million samples or more.
b) Validation Dataset: Monitors NN generalization performance during training to prevent overfitting, containing typically 10,000 to 50,000 samples.
c) Test Dataset: Assesses the model's post-training generalization capability, performance, bias and accuracy. All datasets are generated from identically distributed uniform random variable simulations, spanning the range of Zernike coefficients. Typically, from 100,000 to millions of samples.

Figure 2 below illustrates the basic steps involved in loading and training the neural network.


Min < Z_n < Max
(For 9 terms)
Random Number Generator
Compute Appropriate Zernike Terms
Apply Aperture Function & Compute Wavefront
Compute Monochromatic Defocused PSF
Normalization & Coherence Compensation
Load NN Database with n x n pixel images
NN Training Loop
Output
Z3
Z4
Z5
Z6
Z7
Z8
Z9
Z10
Z15
Compute Wavefront (minus power)

	Coef	Aberration	
3	3.000	Power	2R^2 - 1
4	0.150	Astig X	R^2 cos[2A]
5	-0.200	Astig Y	R^2 sin[2A]
6	0.260	Coma X	(3R^2 - 2) R c
7	0.150	Coma Y	(3R^2 - 2) R s
8	-0.180	Primary Spherical	6R^4 - 6R^2
9	0.130	Trefoil X	R^3 cos[3A]
10	--	Trefoil Y	R^3 sin[3A]
11	--	Secondary Astig X	(4R^2 - 3) R^
12	--	Secondary Astig Y	(4R^2 - 3) R^
13	--	Secondary Coma X	(10R^4 - 12R
14	--	Secondary Coma Y	(10R^4 - 12R
15	0.110	Secondary Spherical	20R^6 - 30R^

Figure 2. The basic steps for loading data and training the NN to estimate (infer) the Zernike coefficients

2.3 Uniqueness of phase retrieval

Phase retrieval faces a conjugate inversion ambiguity, as shown from the pupil function $\mathrm{P(x,y)}$ cross-correlation, the OTF:

$$\mathrm{OTF}(\zeta,\eta) = \mathrm{P(x,y)} \otimes \mathrm{P(x,y)} = \mathrm{P^*(-x,-y)} \otimes \mathrm{P^*(-x,-y)} \tag{4}$$

Where * denotes the complex conjugate operation.
There are 2 pupil functions $\mathrm{P(x,y)}$ leading to the same PSF:

$$\mathrm{P(x,y)} \quad = \mathrm{p(x,y)e^{j\varphi_p(x,y)}} \tag{5}$$

$$\mathrm{P^*(-x,-y)} = \mathrm{p(-x,-y)e^{-j\varphi_p(-x,-y)}} \text{ conjugate inversion} \tag{6}$$

For a circular pupil $\mathrm{p(x,y)}$, a real even function, the ambiguity lies in its phase $\varphi_\mathrm{p}(\mathrm{x,y})$. This can be addressed by introducing a known phase modulation, typically a defocus bias (Z3), though other modulations, such as spherical aberrations, may be used. Figure 3 illustrates a perfect PSF (left) and aberrated PSFs with +3 waves (center) and -3 waves (right) of primary spherical aberration (Z8). In the top row, with no phase modulation (z3 = 0), the aberrated PSFs are indistinguishable, indicating phase sign ambiguity. This is resolved in the bottom row by adding a 10-wave defocus (Z3) bias, making the aberrated PSFs distinct. A significant defocus bias ensures a known sign despite defocusing accuracy, field curvature, sensor tilts, and other errors. Unlike curvature sensing, this method does not require exceeding the caustic region. Defocus errors are treated as aberrations, and the NN is trained over a predefined range of defocus values.

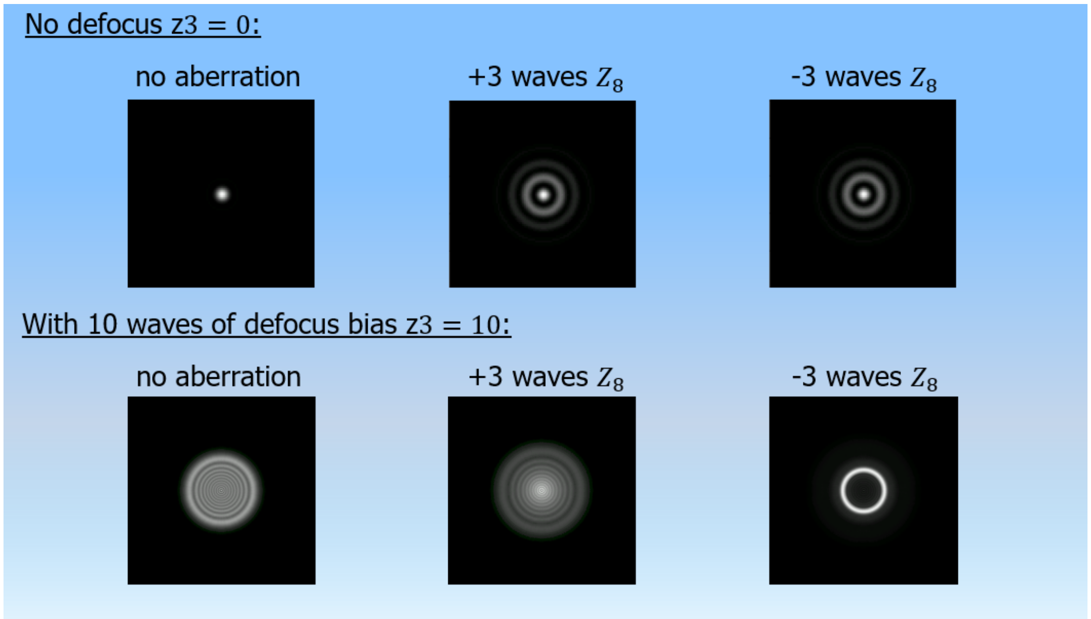


Figure 3. The top row corresponds to the in-focus situation (z3=0) with no phase modulation. The two aberrated PSFs, with respectively +3 (center) and -3 (right) waves of primary spherical (Z8), are identical, indicating that the phase sign has been lost. This ambiguity is resolved by adding 10 waves of defocus (Z3) bias, serving as the modulation, as observed in the bottom row.

3. EXPERIMENTS

3.1 Overview

Several experiments were conducted in the laboratory using both single-pass and double-pass configurations. A NN was trained using the AI4Wave approach outlined in section 2, employing up to one million synthetic samples (defocused images) to estimate the 10 Zernike polynomial coefficients described in Table 1. The defocused monochrome images used for training were 128×128 pixels in size. The defocus bias was set at 2.5 waves RMS, equivalent to 8.66 waves PV, assuming a circular aperture without obstructions. To account for pupil apodization caused by the finite size of a 5-micron pinhole used in the double-pass optical bench experiment, training samples were generated with variable illumination profiles. The illumination across the pupil was modeled as a truncated Gaussian profile, with a maximum drop of -1.08 dB at the pupil edge. In both experiments, we used a Basler acA720-520um USB3.0 camera with a resolution of 728×544 pixels, a pixel size of 6.9 microns squared, and a bit depth of 12 bits. The location of the defocused image of a point source was first determined, and then a cropped 128×128 pixels region of interest (ROI) image was sent to the NN. When the estimated tilt Zernike coefficients from the NN exceed a predefined threshold, a search operation is triggered to recenter the ROI to the source, ensuring continuous tracking. To accommodate for large real-time translations within the ROI, the tilt Zernike coefficients' training dynamic range was set to several waves. Since these tilt coefficients are very large and used solely for tracking purposes, they are excluded from the results. Only the remaining eight Zernike coefficients—defocus, astigmatism, coma, trefoil and primary spherical—are considered here as the relevant aberrations. All processing was conducted in near real-time using TensorRT via Python on an NVIDIA Jetson AGX Orin 32GB Developer Kit running Jetson Linux (based on Ubuntu), the target machine for standalone applications.

Figure 4 below shows the NVIDIA Jetson AGX Orin Series device and its functional block diagram. These devices offer up to 275 TOPS (Tera operations per second) for 8-bit operations (INT8). In our application, we found that a 16-bit floating-point format (FP16) provided a good tradeoff between speed, numerical accuracy and stability, delivering up to 137 TOPS. The Jetson platform is designed to handle multiple video streams and cameras in parallel, enabling its hardware resources to run numerous similar neural networks (NNs) simultaneously with comparable real-time performance. This level of parallelization allows the retrieval of multiple Zernike coefficient sets, for instance, from a single image featuring several sources, supporting real time field-dependent wavefront sensing for applications such as Multi-Object Adaptive Optics (MOAO). Alternatively, multiple cameras positioned at different field locations can operate in parallel, each capturing also several sources within its view. In addition, AO systems can address very large wavefront error dynamic ranges by employing multiple NN models, each optimized for different levels of aberration magnitude, running concurrently from a single image. The use of the AI4Wave approach on high-performance devices like the Jetson fully harnesses the potential of computational optics at sub millisecond range in a deterministic way.

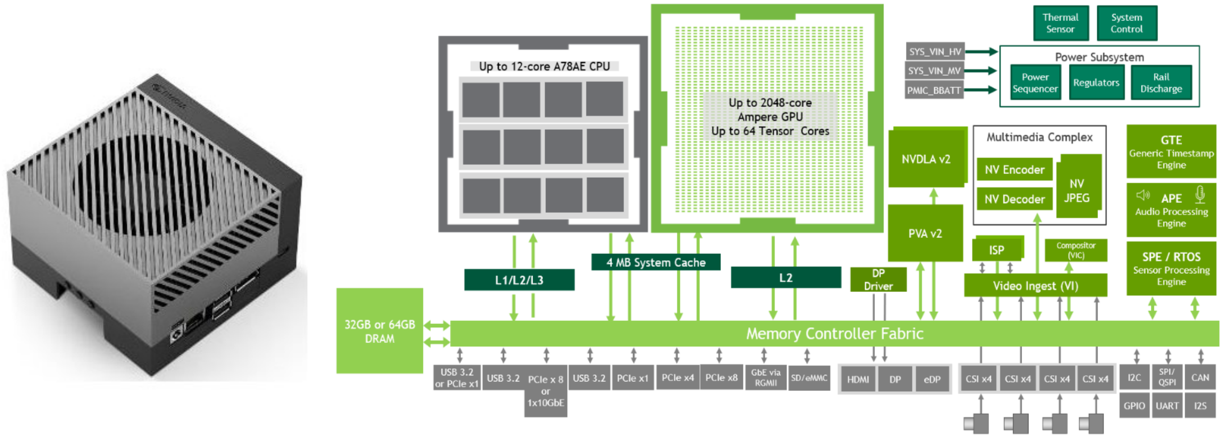
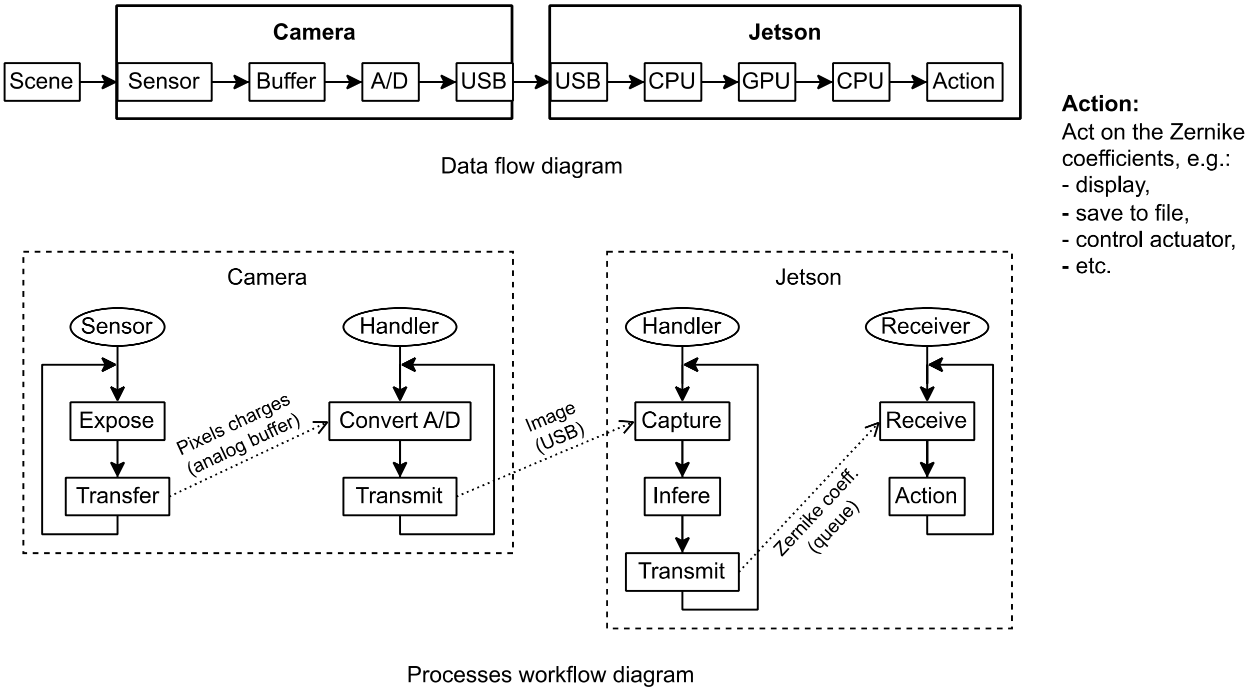


Figure 4. NVIDIA Jetson AGX Orin Series device and its functional block diagram

The next figure 5 illustrates the data flow and related process workflow diagrams. Experiments have shown that with the selected camera and the computational power required for running the NN and related processing, the bottleneck lies in the chosen camera sensor readout, even at the camera's shortest exposure time of 24 µs. At this stage, we were not yet constrained by the AI4Wave phase retrieval calculations. Therefore, by selecting a faster camera, the system's latency could be further improved, with the eventual limit being the NN inference time, essentially the time between submitting the image to the NN input and getting the estimated Zernike coefficients.



Data flow diagram

Processes workflow diagram

Action:
Act on the Zernike coefficients, e.g.:
- display,
- save to file,
- control actuator,
- etc.

Figure 5. Data flow and related processes workflow diagrams

The timeline in figure 6 below illustrates the fundamental timing. Since the camera latency is the bottleneck in the chain in this configuration, it operates in free-run image acquisition mode to maximize frame rate. It also natively supports overlapping the next exposure with the data readout from the previous one, as illustrated in figure 5 by a rapid transfer to an analog buffer, enabling parallelization of the time-consuming A/D conversion. It is set to a mode called “latest image only”, designed for real-time, where only a single, most recent acquisition is ever available to the computer in order to ensure a minimal latency. Real-time processing prevents frame loss, provided the Jetson is ready to read the next frame. For a ROI of 128×128 pixels at 12 bits, the camera readout time (which equals to the sampling period in this configuration) is specified as 918 µs, resulting in a maximum frame rate of 1089 fps and allowing a maximum image acquisition exposure time of approximately 800 µs. Longer exposure times, near the camera readout time, will extend the sampling period, reducing the frame rate accordingly. After each exposure is completed, there is a processing latency that includes the camera sensor data readout time, data transfer, and NN inference time before the Zernike coefficients are available for use (action). Finally, the sensor response time, defined as the interval from the start of the exposure to the availability of the Zernike coefficients, is the sum of the exposure time and the processing latency time. As discussed above, the lower limit for the latter is constrained by the NN inference time and with a faster camera it would be possible to approach this limit. Essentially, a new set of Zernike coefficients becomes available at each results interval, which equals to the sampling period, but with a processing latency time relative to the end of its corresponding exposure. In our implementation, this processing latency is primarily dictated by the camera readout time, which could be significantly reduced by selecting another one.

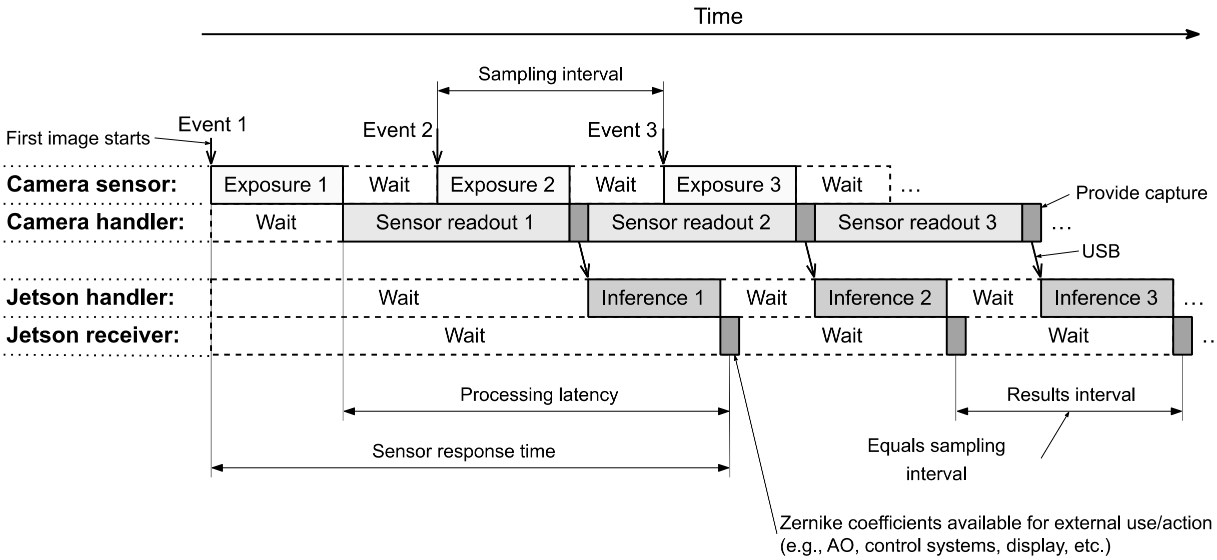


Figure 6. Timeline of the overall sensor application implementation. The camera operates on free-run image acquisition mode with overlapping acquisitions. In this timeline and our implementation, the NN inference and exposure times are shorter than the sensor readout time.

The synthetic data generation, as well as the inverse mathematical model (the NN) linking the irradiance to the Zernike coefficients estimations, were implemented in Python using TensorFlow. TensorFlow is an open-source framework developed by Google for machine learning and deep learning tasks, focusing on model development, training, and deployment across various platforms. The NN was trained with 32-bit floating-point precision (FP32) on a Windows OS virtual machine in the cloud. For integration into the Jetson, optimization was performed with different formats for use with TensorRT. TensorRT is a high-performance deep learning inference optimizer and runtime library by NVIDIA, designed to improve the speed and efficiency of NN inference on NVIDIA GPUs. The best trade-off between numerical accuracy, stability and computation time was achieved using a 16-bit floating-point format (FP16). The typical numerical errors across the Zernike coefficients and samples between the initial 32-bit TensorFlow precision and the optimized TensorRT 16-bit version averaged at or below 0.001 wave RMS of error, consistent with the standards of high-precision optical measurement.

3.2 Performances baseline

During software and system development, as well as to establish a baseline for system performance, we used simulated defocused source images with known aberrations (Zernike coefficients) without the camera attached to the Jetson. This approach enabled us to compare numerical accuracies across floating-point formats and optimization strategies chosen for integrating the NN on the Jetson using the trtexec utility software provided with TensorRT. It also provided a benchmarking baseline for NN inference time. Table 2 below summarizes the main results. The reference column represents the TensorFlow environment using the same numerical precision (FP32) as was applied during synthetic data generation and NN training in the cloud, with no NVIDIA optimizations. Computing time is not reported here as it is irrelevant in this context. The second and third columns correspond to optimized versions using TensorRT with FP32 and FP16 floating-point formats for runtime. All calculations in this comparison were performed on the Jetson. TensorRT optimization was applied with default parameters, except for specifying the floating-point format and using the highest optimization level during model building. This setting employs the most aggressive optimizations to maximize inference runtime performance on NVIDIA GPUs. Typically, 100,000 or more samples were used for the statistics.

Zernike coefficient [wave RMS]	TensorFlow FP32 mean	TensorFlow FP32 standard deviation	TensorRT FP32 mean	TensorRT FP32 standard deviation	TensorRT FP16 mean	TensorRT FP16 standard deviation
Defocus	0.0130	0.0110	0.0130	0.0110	0.0130	0.0110
Vertical astigmatism	0.0093	0.0067	0.0093	0.0067	0.0095	0.0069
Oblique astigmatism	0.0057	0.0046	0.0057	0.0046	0.0058	0.0046
Horizontal coma	0.0072	0.0068	0.0072	0.0068	0.0073	0.0069
Vertical coma	0.0061	0.0041	0.0061	0.0041	0.0066	0.0040
Oblique trefoil	0.0044	0.0032	0.0044	0.0032	0.0044	0.0032
Vertical trefoil	0.0074	0.0043	0.0074	0.0043	0.0073	0.0043
Primary spherical	0.0045	0.0033	0.0045	0.0033	0.0046	0.0033
Inference time [µs]	NA	NA	448	4.96	259	3.66

Table 2. Zernike coefficient error statistics (wave RMS) and NN inference time (µs) for TensorFlow FP32 floating point format (used for training) versus TensorRT FP32 and FP16 floating point formats (optimized versions for runtime)

Overall, the Zernike coefficient error biases (means) and standard deviations for the test images, unseen during NN training, are typically below 0.01 wave RMS. Differences between TensorFlow and TensorRT with FP32 are negligible, below 1e-6 wave RMS. With FP16, the differences remain at or below 0.001 wave RMS, making FP16 the selected format for subsequent experiments and tests. Other formats such as FP8 or INT8 were not fully tested and could still reduce the computation speed without affecting the accuracy. As noted, we used mainly default NVIDIA TensorRT optimization parameters. Adjusting these or refining the NN structure should potentially reduce inference time further, a direction worth exploring in future work. Finally, it is important to note that no camera was connected to the Jetson, and no data transfer occurred during this baseline test, making the recorded inference times the lowest possible boundaries.

3.3 Screen-based experiment

In this experiment, the Basler camera is connected to the Jetson, and the system operates at maximum speed using the workflow shown in Figure 5 and the optimized NN configuration (FP16) discussed in Section 3.1. To efficiently test hardware, data flow, and performance without an optical bench, a PC monitor displays computed defocused images with known aberrations. The camera is equipped with a Basler Lens C23-1224-5m-P (12 mm focal length, f/2.4) for reimaging the monitor's screen, maximizing SNR. Figure 7 illustrates the setup with the camera and its lens positioned near the screen displaying defocused aberrated images. This is a single pass test. Care was taken to avoid gamma compression and image processing artifacts. Camera gamma was disabled, and displayed images were pre-corrected to ensure linearity at the physical level (screen irradiance). This step is critical, as the IS-WFS used by AI4Wave is sensitive to irradiance levels. The defocused images are monochromatic calculations.

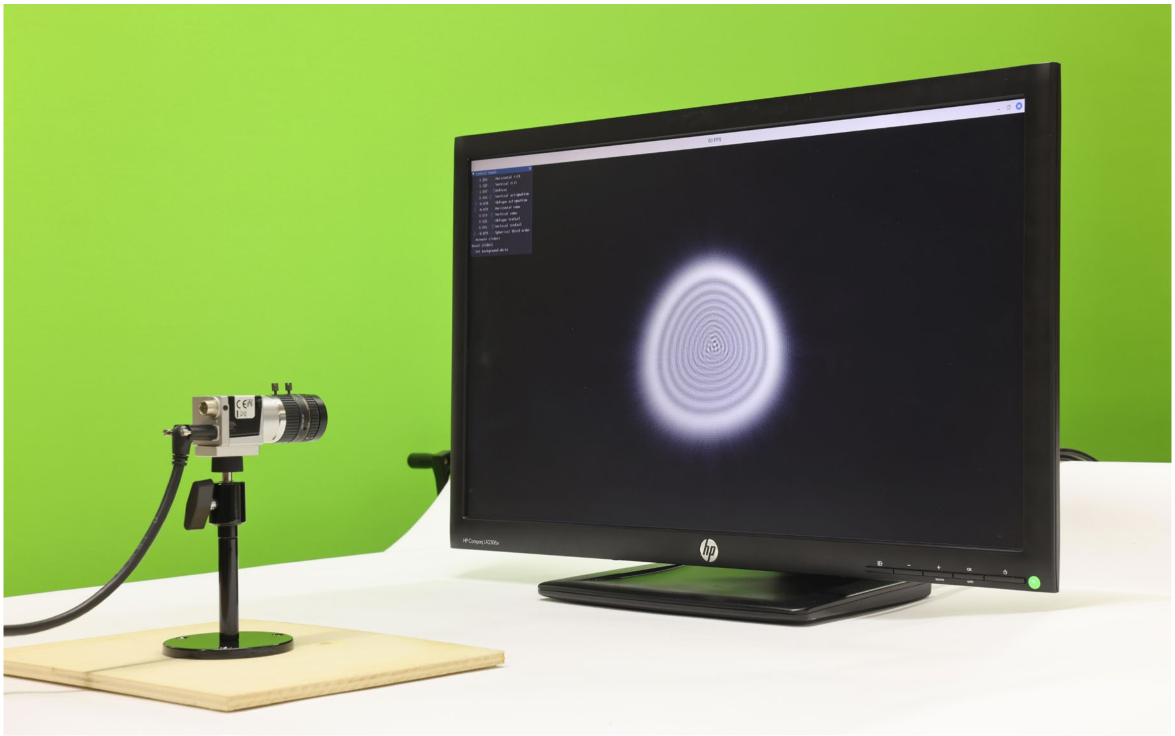


Figure 7. Screen-based experiment setup. The camera is connected to the Jetson via a USB3 cable, fitted with a lens, and positioned in front of a monitor to image the defocused source with known aberrations displayed on the monitor's screen

Table 3 summarizes the timing statistics for TensorRT FP16 floating-point format over 50,000 frames of 128×128 pixels at 12 bits with a 100 µs exposure. The sampling period, which defines the wavefront data rate, is primarily driven by processing latency for that exposure time, accounting for nearly 90% of it. Almost half of this latency is attributed to the camera's sensor readout and data transfer. A different camera choice should potentially reduce the sampling period by a few hundred microseconds. The NN inference time, which sets the lower bound for processing latency, constitutes about 30% of the sampling period and approximately 45% of the processing latency. The mean NN inference time of 370 µs is around 110 µs higher than the value reported in Section 3.2, Table 2 (without a camera), likely due to Python's real-time overhead and system management. These results demonstrate that a wavefront rate of 1kHz, or even higher, can be readily achieved using AI4Wave phase retrieval combined with a fast NN inference platform like the Jetson. The AI4Wave phase retrieval (IS_WFS) is fully deterministic, with runtime defined by NN feedforward computation, requiring no iterations or initial conditions, while the wavefront error accuracy is benchmarked offline with millions of samples, avoiding issues typically encountered by runtime non-linear optimization approaches.

TensorRT FP16 format, exposure time = 100 µs	mean	standard deviation
Frame rate, wavefront rate	1,091 Hz	NA
Sampling period	917 µs	129 µs
Processing latency	817 µs	NA
NN inference	370 µs	44 µs

Table 3. Basic timing statistics for TensorRT FP floating point format (50,000 frames).

Next table 4 presents the wavefront error statistics for 5,000 defocused source images with various known aberrations, re-imaged using the setup shown in Figure 7.

Zernike coefficient [wave RMS]	mean	standard deviation
Defocus	0.0210	0.0160
Vertical astigmatism	0.0089	0.0080
Oblique astigmatism	0.0082	0.0078
Horizontal coma	0.0058	0.0058
Vertical coma	0.0130	0.0070
Oblique trefoil	0.0041	0.0035
Vertical trefoil	0.0061	0.0040
Primary spherical	0.0034	0.0028

Table 4. Zernike coefficient error statistics (wave RMS) for 5,000 defocused sources, with known aberration, re-imaged by the camera and its lens opened at f/2.4

Overall, the mean and standard deviation results are very similar to those reported in the baseline test (section 3.2, table 2, no camera). Slight increases, below 0.01 wave RMS, are observed primarily in the defocus, oblique astigmatism, and vertical coma coefficients. These differences are expected and align with the lens's optical performance at f/2.4, as the camera objective lens was not calibrated off in the data. Also, some manual defocus error is likely present. However, the results remain well below the diffraction limit for a Strehl ratio of 0.8, which corresponds to 0.075 wave RMS. Below figure 8 shows an example, from left to right, the simulated image on the monitor, the camera-captured image, and the reconstructed image using Zernike coefficients inferred by the NN. The camera-captured image shows some contrast loss due to the camera objective lens's MTF limitations, but the NN provides a reliable estimation of the Zernike coefficients, as demonstrated by the reconstructed image. Noise and blur added to the synthetic samples used for NN training enhance robustness in the solution.

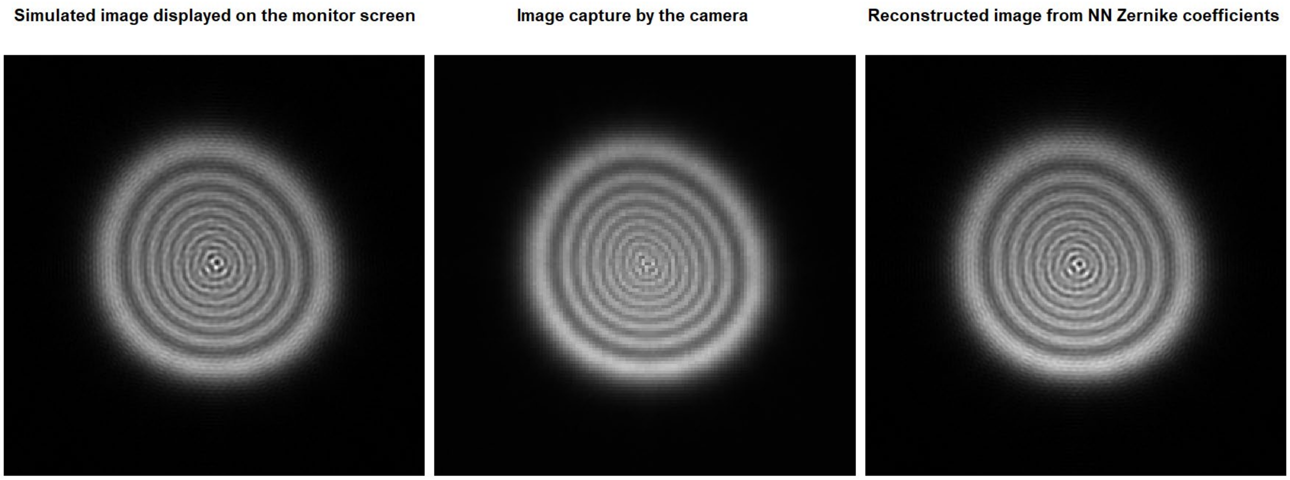

Simulated image displayed on the monitor screen
Image capture by the camera
Reconstructed image from NN Zernike coefficients

Figure 8. Example of an aberrated defocused source re-imaged by the camera (as shown in figure 7): from left to right, the simulated image on the monitor, the camera-captured image at the center, and the NN-reconstructed image using inferred Zernike coefficients.

3.4 Double pass optical bench experiment

In this experiment, a double-pass optical test bench shown in figure 9 measures the wavefront reflected by a flat mirror. A 10 mW 637 nm laser illuminates the F/12.5 pupil via a 5 µm pinhole. To address apodization from the pinhole, variable Gaussian illumination profiles were modeled, with a -1.08 dB edge attenuation. Light reflects through a 2 µm tick 50:50 pellicle to an achromat (25 mm dia., 300 mm FL, 24 mm aperture) and collimates for double-pass reflection. A Basler acA720-520um camera (728 × 544, 6.9 µm pixels, 12-bit) offset from the focal plane introduces defocus bias. Captured data feeds the NN to infer the 10 Zernike terms (table 1).

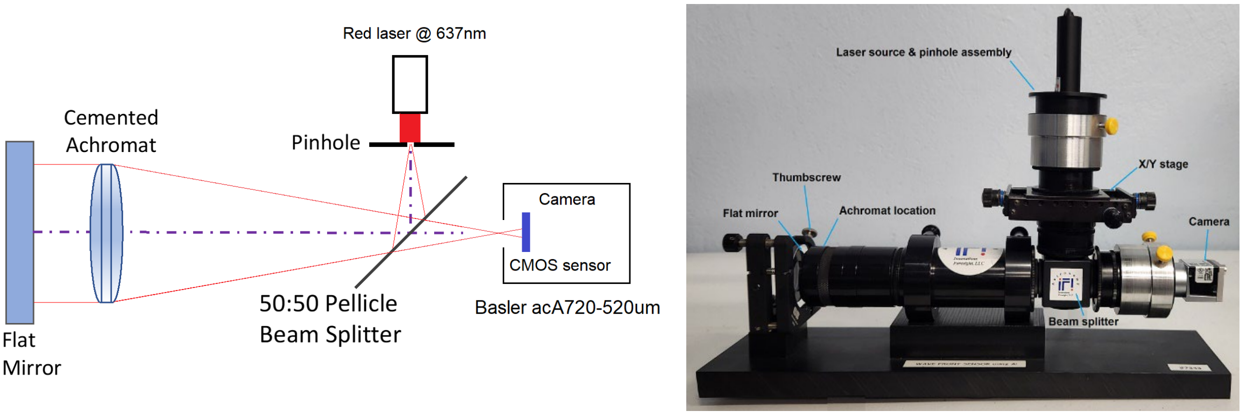


Figure 9. A double-pass optical test bench with a 10 mW 637 nm laser illuminates the f/12.5 pupil via a 5 µm pinhole. Light reflects through a 50:50 pellicle and achromat (24 mm aperture) before double-pass imaging onto a Basler acA720-520um camera, offset for the required defocus bias for AI4Wave

A thumbscrew located at 45° on the flat edge applies, on demand, mechanical stress, inducing some optical aberrations primarily in the form of astigmatism. The laser's power can be controlled from 0.1 mW to 10 mW, with a 3 mm beam diameter at the pinhole level. Consequently, the pinhole radiates approximately 30 nW at full power and ~3nW at 10% power (1 mW laser setting). In this double-pass configuration, less than 3% of the radiated power reaches the sensor, which at 637nm for 10% power is about 70 pW, or 220 photon per µs. Figure 10 below compares a raw image (left), captured with a 400 µs exposure time, 100% power setting (10 mW), and 24 dB camera analog gain, to a reconstructed image (right) generated using the 10 Zernike coefficients inferred by the NN from the raw image. No stress was applied to the mirror.

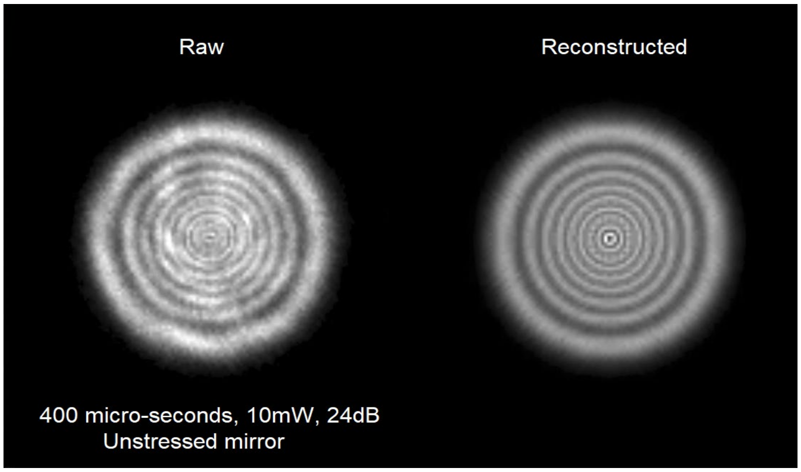


Figure 10. Comparison of raw vs reconstructed image, mirror unstressed

Eventually, the exposure time was reduced to the camera's minimum setting of 24 µs, with the laser power adjusted to 10% (equivalent to 1 mW). To introduce controlled distortions, a thumbscrew (seen on figure 9, next to the flat mirror) was used to apply stress to the mirror at a 45° angle, inducing mainly an astigmatism like distortion. This experimental setup led to a noticeable drop in the signal-to-noise ratio (SNR), primarily due to increased shot noise on the defocused image caused by the reduced exposure time and lower photon count. Despite this low SNR, as shown in Figure 11, the neural network (NN) demonstrated remarkable robustness, effectively reconstructing the raw image and accurately retrieving the wavefront information. This performance highlights the NN's ability to handle low-SNR scenarios, validating its reliability for demanding optical applications under low illumination conditions.

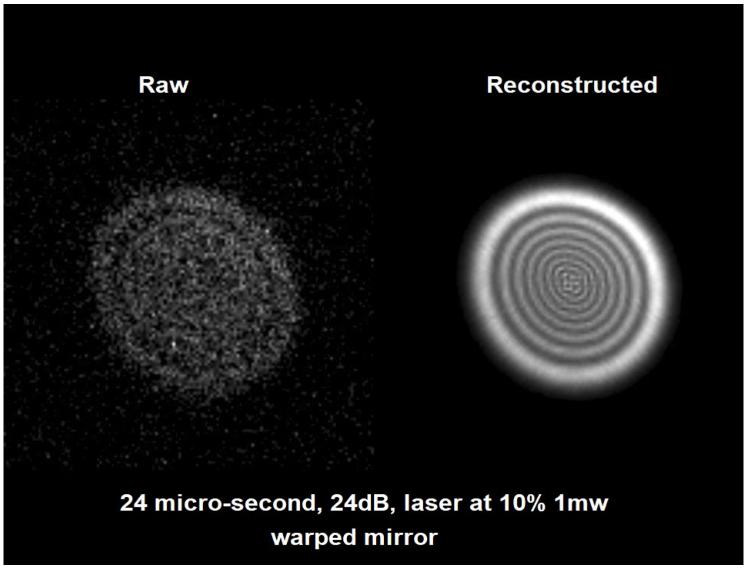
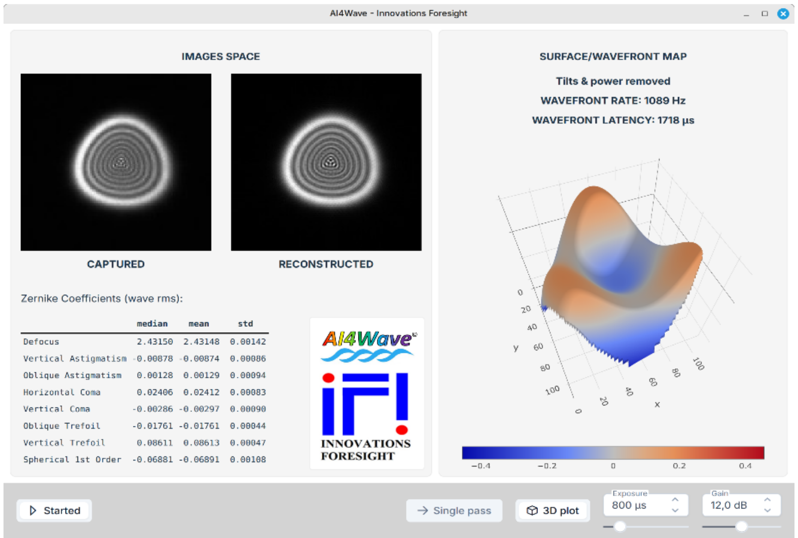


Figure 11. Comparison of raw vs reconstructed image at low intensity, laser set a 10%. Mirror wrapped using the side thumbscrew.

Figure 12 shows a screenshot, while Video 1 demonstrates the GUI on the Jetson platform, displaying live raw images, reconstructed images, Zernike coefficients, and a wavefront heat plot in real time.



Tilts & power removed

WAVEFRONT RATE: 1089 Hz

WAVEFRONT LATENCY: 1718 µs

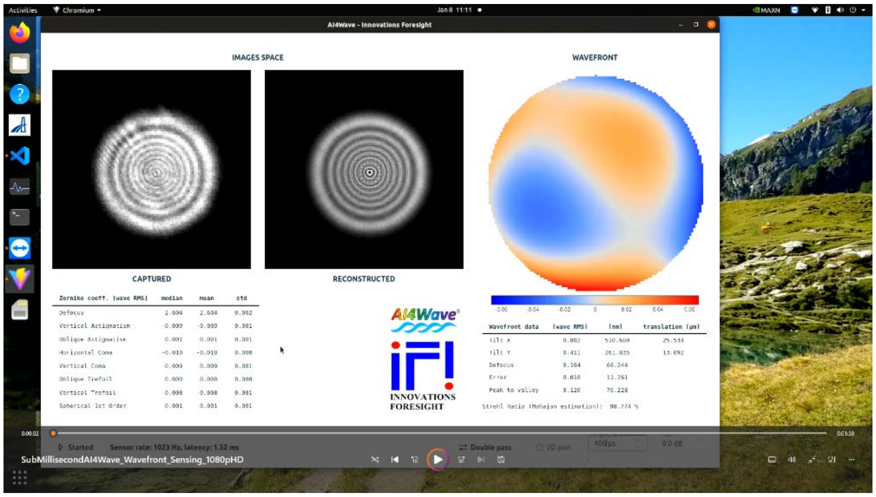
Zernike Coefficients (wave rms):

	median	mean	std
Defocus	2.43150	2.43148	0.00142
Vertical Astigmatism	-0.00878	-0.00874	0.00086
Oblique Astigmatism	0.00128	0.00129	0.00094
Horizontal Coma	0.02406	0.02412	0.00083
Vertical Coma	-0.00286	-0.00297	0.00090
Oblique Trefoil	-0.01761	-0.01761	0.00044
Vertical Trefoil	0.08611	0.08613	0.00047
Spherical 1st Order	-0.06881	-0.06891	0.00108

Exposure 800 µs

Gain 12,0 dB

Figure 12. Real-time GUI running on the Jetson platform



Zernike coeft. [wave RMS]	median	mean	std
Defocus	2.604	2.604	0.002
Vertical Astigmatism	-0.009	-0.009	0.001
Oblique Astigmatism	0.001	0.001	0.001
Horizontal Coma	-0.010	-0.010	0.000
Vertical Coma	0.009	0.009	0.001
Oblique Trefoil	0.009	0.008	0.000
Vertical Trefoil	-0.008	-0.008	0.001
Spherical 1st Order	0.001	0.001	0.001

Wavefront data	[wave RMS]	[nm]	translation [µm]
Tilt X	0.802	510.689	25.534
Tilt Y	0.411	261.855	13.092
Defocus	0.104	66.244	
Error	0.018	11.261	
Peak to valley	0.120	76.228	

Strehl Ratio (Mahajan estimation): 98.774 %

Started   Sensor rate: 1023 Hz, latency: 1.32 ms

SubMillisecondAI4Wave_Wavefront_Sensing_1080pHD

Video 1. Real-time GUI video on the Jetson platform at 1 kHz
http://dx.doi.org/doi.number.goes.here

4. CONCLUSIONS

This work has shown that the AI4Wave approach is a practical solution for high-speed image-space wavefront sensing systems. By relying entirely on synthetic, normalized data for training and leveraging optimized inference on off-the-shelf hardware like the NVIDIA Jetson module, the AI4Wave platform provides a deterministic, non-iterative solution for real-time phase retrieval (wavefront and surface) measurement directly from image space, without requiring dedicated wavefront sensing hardware. Unlike traditional methods that rely on physical data or iterative optimization, AI4Wave provides essentially layout-independent performance, allowing for rapid deployment across diverse optical configurations. Its ability to handle large wavefront errors, achieve sub-millisecond processing, and operate with minimal hardware makes it a compelling tool for AO, optical metrology, and other demanding applications. Experimental results confirm its accuracy, repeatability, and speed under various conditions, including low SNR scenarios. With computational efficiency, scalability, and the unique capability to measure field-dependent wavefronts from a single image, AI4Wave offers a simple, ready-to-deploy solution for fast wavefront sensing, overcoming most of the traditional limitations while providing versatility for industrial and scientific applications.

REFERENCES

[1] Gaston Baudat, "Low-cost wavefront sensing using artificial intelligence (AI) with synthetic data", Proc. SPIE 11354, Optical Sensing and Detection VI April 1, Strasbourg, France (2020).

[2] Wu, Yu & Guo, Youming & Bao, Hua & Rao, Changhui. "Sub-Millisecond Phase Retrieval for Phase-Diversity Wavefront Sensor". Sensors. 20. 4877. 10.3390/s20174877 (2020).

[3] Roddier Claude and Roddier Francois, "Wave-front reconstruction from defocused images and the testing of ground-based optical telescopes", Journal of Optical Society of America, vol. 10, no. 11, 2277-2287 (1993).

[4] Gaston Baudat, R. E. Parks, B. Anjakos, "A new approach to wavefront sensing: AI software with an autostigmatic microscope", Proc. SPIE 12672, Applied Optical Metrology V, San-Diego, USA (2023).

[5] Gaston Baudat, Robert E. Parks, "Aspherical surface measurement: a cost-effective and fast AI solution," Proc. SPIE 13134, Optical Manufacturing and Testing 2024, 1313403 (30 September 2024)

[6] Gaston Baudat and John Hayes, "A star-test wavefront sensor using neural network analysis", Proc. SPIE. 11490, Interferometry XX, 20 August 5, San-Diego, USA (2021).