Inverse problem on Imaging and Imaging System
The study of coherence, aberration, and optimization

Shao, Y.

DOI
10.4233/uuid:b486f090-362c-4567-b51e-71547abd871c

Publication date
2021

Document Version
Final published version

Citation (APA)
Shao, Y. (2021). *Inverse problem on Imaging and Imaging System: The study of coherence, aberration, and optimization*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:b486f090-362c-4567-b51e-71547abd871c

Inverse problem on Imaging and Imaging System

The study of coherence, aberration, and optimization

Inverse problem on Imaging and Imaging System

The study of coherence, aberration, and optimization

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op vrijdag, 25 juni 2021 om 10:00 uur

door

Yifeng SHAO

Master of Science in Applied Physics,
Technische Universiteit Delft, Nederland,
geboren te Wuhan, China.

Dit proefschrift is goedgekeurd door de promotoren.

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr. H.P. Urbach,	Technische Universiteit Delft, promotor
Dr. F. Bociort,	Technische Universiteit Delft, copromotor
Onafhankelijke leden:	
Prof.dr. A.P. Mosk	Universiteit Utrecht
Prof.dr.ir. A.J. den Boef	Vrij Universiteit Amsterdam & ASML Holding N.V.
Prof.dr. X.C. Yuan	Shenzhen Univerisity
Prof.dr. W.M.J.M. Coene	Technische Universiteit Delft & ASML Holding N.V.
Overige leden:	
Dr. M. Loktev	Kulicke & Soffa Liteq B.V.

Keywords: Phase retrieval, phase diveristy, aberration metrology, holography, diffractive imaging, spatial coherence measurement

Printed by: Ipskamp Printing

Front & Back: The cover uses a section of the famous painting ”starry night” by dutch artist Vincent Van Gogh. Scientists often research for the most accurate way for recording an object by removing the blur and the distortion. However, for artists these blur and distortion represent the beauty, and are often exaggerated for the interpretation of their artistic view. Like in this cover, the milky way is depicted as a distorted vortex with stars severely blurred, possibly due to spherical aberration.

ISBN 978-94-6421-399-7

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

4

Contents

Summary

In this thesis we try to provide novel solutions to key problems related to imaging and imaging system. Imaging is usually referred to as the technique for reproducing the information of the object. In optics, we usually refer the object information to the light field in the object plane due to the interaction of the illumination field and the object. Imaging technique allows the reproduced object information to be recorded by detectors such as human eye, photo-resist, or CCD/CMOS sensor.

In order to image an object, there must be light. We consider only unpolarized quasi-monochromatic light illuminating the object and we approximate the field in the object plane by the multiplication of the transmittance/reflectance of the object and the illumination field (the first Born approximation). However, the results in this thesis can also be generalized to situations beyond the above scope. Here we mainly focus on two aspects: the aberrations, the errors of the imaging system, and the spatial coherence of the light field.

In a typical imaging scenario, we consider the object plane field as a source consisting a series of point sources. The field generated by each point source propagates independently to the image plane through the imaging system. Ideally, these fields should all identically have the correct distribution and be centered at the correct location. However, this is not the case in presence of the aberrations. As a result, the image will become blurred.

In the image plane, different fields generated by different point sources interfere with each other and the intensity of the interference pattern is measured by a detector. The spatial coherence of the illumination field determines the ability of interference. In the complete coherent case, for example the object is illuminated by laser light, the field added together, while in the complete incoherent case, e.g. the object itself is an incoherent source, the intensity of the field added together.

In most situations, the object is illuminated by an incoherent source placed at certain distance such as that in the Köhler illumination and the illumination field is only spatially partially coherent. Disturbances during the measurement process like mechanical vibration and atmosphere turbulence also contribute to the degradation of spatial coherence.

Spatial coherence also plays an essential role in computational imaging. In this scenario, we measure diffraction patterns of the object field and we perform imaging by computationally reconstructing the object field. However, when the object field is spatially partially coherent, computing the diffraction pattern becomes rather complicated. Without modification to the algorithm, the reconstruction will be blurred, but even with modification, the algorithm can only handle limited degree of spatial partial coherence.

In chapter 2, we study a method for retrieving the aberrations from a series of blurred images. Conventional methods often characterize aberrations by measuring

the pupil wavefront and hence require spatially coherent illumination. Our method is particularly suitable for incoherent imaging and only requires the imaging system to be disturbed in a known fashion, e.g. by introducing known defocus.

While in chapter 2 we considers only shift-invariant aberrations, we propose a method to measure the shift-variant aberrations in chapter 3. Here the aberrations are functions of both the pupil and the field-of-view (FOV) coordinates. Our method collects data of the aberrations at a large number of FOV locations in parallel. As the data processing can also be parallelized, the proposed method is extremely efficient compared to existing methods. Potential applications are imaging systems with large FOV e.g. the lithography projection system for IC manufacturing in the semiconductor industry.

In chapter 2 and chapter 3, our problem of retrieving aberrations from some measurements is a typical inverse problem, which needs to be solved by employing iterative optimization technique. It usually requires formulating an error function that depends on the aberrations and deriving the gradient of the error function with respect to the aberrations.

In chapter 4 and chapter 5, we propose two non-iterative methods for measuring the spatial coherence of an arbitrary field. In particular we measure the correlation between the fields at a reference point and at a large number of sampling points. Both methods use specially designed masks and require measuring the far-field diffraction pattern of the transmitted/reflected light. The result is two-dimensional distribution that also serves the goal for computational imaging.

1

Introduction

1

In this thesis, we study the inverse problem on imaging and imaging system. We consider only quasi-monochromatic light and the scalar case of diffraction, in which the polarization effect is neglected. Our study covers a broad range of topics, including the spatial coherence of light, the aberration of the imaging system, and the optimization technique for solving the inverse problem. Through the thesis I tried to propose novel approaches for all problems. In the introduction chapter, we discuss about some fundamental concepts.

1.1. The coherence of light sources

Coherence is a fundamental property of light that describes the correlation between two electromagnetic waves. The degree of coherence is usually evaluated in terms of the visibility of the interference pattern formed by two waves. At optical frequencies, the interference pattern varies so fast that the variation cannot be captured by the eye or by the camera. Therefore, only the time-averaged intensity distribution of the interference pattern can be measured. The degree of coherence hence describes a statistical phenomenon that can only be observed in time-averaged measurements.

The degree of coherence depends on the phase difference between two waves. We define that two waves are coherent when the phase difference is stationary, which gives rise to the interference pattern (visibility = 1), and are incoherent when the phase difference is not stationary, which smears the interference pattern by time-averaging (visibility = 0). The intermediate state is referred to as partially coherent and is of great interest for theoretical studies and for practical applications.

We shall distinguish between the temporal coherence and the spatial coherence. The former is related to the spectral bandwidth of the source, while the later is related to the size of the source. To evaluate both coherences, we should observe the interference pattern formed by two electromagnetic waves at different times and at different locations, respectively.

Table. Relation between the source and the coherence of light

	Point source	Extended source
Monochromatic	Spatially coherent, Temporarily coherent	Spatially incoherent, Temporarily coherent
Chromatic	Spatially coherent Temporarily incoherent	Spatially incoherent, Temporarily incoherent

1.1.1. Temporal coherence

Consider a point source in the source plane which is placed on the axis of the setup of the Young's interference experiment as shown in Fig. 1.1. The point source emits a series of wave-packets due to spontaneous emission. Each wave-packet has a random initial phase and an amplitude that varies significantly in time. Such a wave-packet series can be written as:

$$E(\mathbf{r},t) = \sum_n A_n(\mathbf{r},t)\exp[-i(\omega t - \delta_n)], \tag{1.1}$$

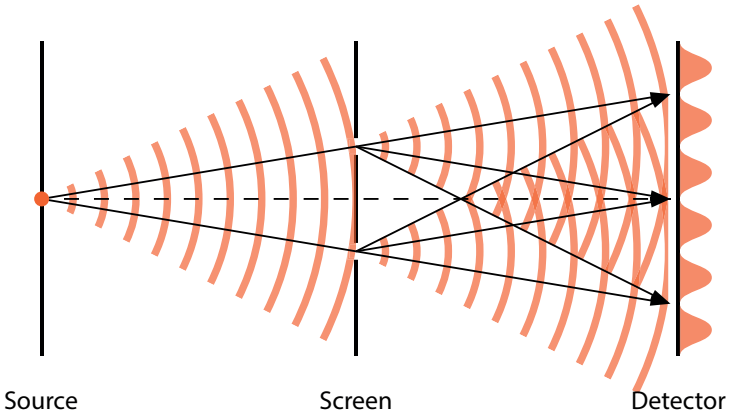

Figure 1.1: Setup of the Young's interference experiment. A chromatic point source is placed on the axis in the source plane. The electromagnetic wave emitted by the point source first propagates to a screen and then to a detector at certain distance. The arrows represent the direction of wave propagation, and the curves represent the wavefront.

where ω denotes the central frequency (also called the carrier frequency), and $A_n(\mathbf{r}, t)$ and δ_n denote the time-varying amplitude and the initial phase of the nth wave-packet, respectively. $E(\mathbf{r}, t)$ represents the electromagnetic wave emitted by the point source at time t and observed at $\mathbf{r}$ in the detector plane.

In Fig. 1.1, we place two pinholes symmetrically on both sides of the axis. We take the wave that passes through one pinhole as the reference, and we denote the time delay of the wave that passes through the other pinhole with respect to the reference as τ. So, for a given separation between the two pinholes, τ is a function of the detector plane location $\mathbf{r}$. The intensity distribution of the interference pattern is measured by integrating over time T, which can be expressed by

$$\begin{aligned} I(\mathbf{r}, \tau) &= \langle |E(\mathbf{r}, t) + E(\mathbf{r}, t+\tau)|^2 \rangle_T \\ &= \langle |E(\mathbf{r}, t)|^2 \rangle_T + \langle |E(\mathbf{r}, t+\tau)|^2 \rangle_T + 2\Re\{\langle E(\mathbf{r}, t) E(\mathbf{r}, t+\tau)^* \rangle_T\}, \end{aligned} \tag{1.2}$$

where $\langle \cdot \rangle_T$ denotes time-averaging and * denotes the complex conjugate. We define the cross-term in Eq. (1.2) as the mutual coherence function (MCF) for time delay τ:

$$J(\mathbf{r}, \tau) = \langle E(\mathbf{r}, t) E(\mathbf{r}, t+\tau)^* \rangle_T = \lim_{T \to \infty} \frac{1}{T} \int_T E(\mathbf{r}, t) E(\mathbf{r}, t+\tau)^* \mathrm{d}t. \tag{1.3}$$

The MCF describes the correlation between a wave and itself as a function of τ, which is independent of the integration time T as $T \to \infty$. Eq. (1.2) indicates that the MCF determines the visibility of the time-averaged interference pattern.

By substituting Eq. (1.1) into Eq. (1.2), we derive that the expression for the time-averaged interference pattern generated by a chromatic point source is given

1

by

$$I(\mathbf{r},\tau)=\frac{1}{T}\int_T\sum_{n=1}^{N}|A_n(\mathbf{r},t)|^2\mathrm{d}t+\frac{1}{T}\int_T\sum_{n=1}^{N}|A_n(\mathbf{r},t+\tau)|^2\mathrm{d}t+ 2\Re\left\{\frac{1}{T}\int_T\sum_{n=1}^{N}\sum_{n'=1}^{N}A_n(\mathbf{r},t)A_{n'}(\mathbf{r},t+\tau)^*\exp[-i(\delta_n-\delta_{n'})]\mathrm{d}t\right\}, \quad (1.4)$$

where N denotes the number of wave-packets captured by the detector during the integration time T. We have $N\to\infty$ because $T\to\infty$. Therefore, only wave-packets with the same initial phase ($\delta_n-\delta_{n'}=0$) can interfere, while wave-packets with different initial phase ($\delta_n-\delta_{n'}\neq 0$) cannot. The MCF in Eq. (1.4) thus becomes

$$J(\mathbf{r},\tau)\propto\int_T A(\mathbf{r},t)A(\mathbf{r},t+\tau)^*\mathrm{d}t, \quad (1.5)$$

which depends only on the time-varying amplitude of the wavepacket. Notice that here different wavepackets have the same time-varying amplitude $A(\mathbf{r},t)$.

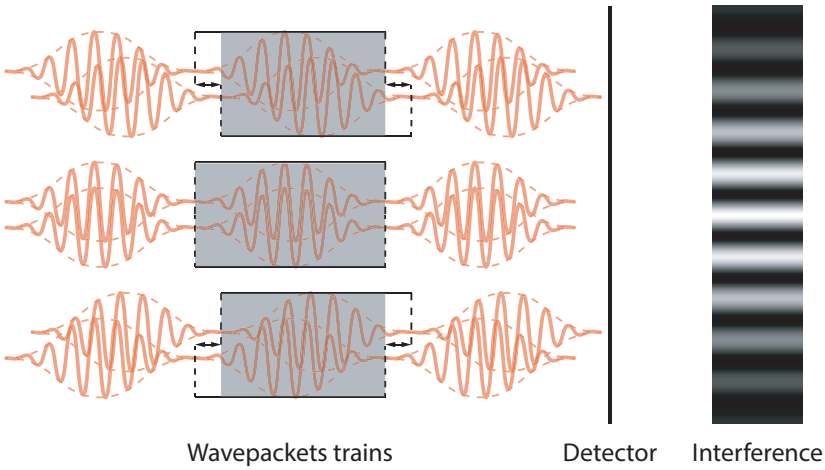

Figure 1.2: Illustration of the interference between two wave-packet trains in the Young's interference experiment. Each wave-packet train passes through a pinhole. The time delay, or the optical path difference, between the two wave-packets trains, depends on the detector plane location and determines the visibility of the interference. (Solid line: the time-varying oscillating amplitude. Dashed line: the profile of oscillation. Gray rectangle: overlap between the two wave-packts trains.)

$J(\mathbf{r},\tau)$ in Eq. (1.5) is given by the overlap between the two wave-packets. The overlap depends on the time delay τ and hence depends on the detector plane location $\mathbf{r}$. As a result, $J(\mathbf{r},\tau)$ can actually be described by only $\mathbf{r}$ in the Young's interference experiment as shown in Fig. 1.2.

As $\mathbf{r}$ (with origin at the intersection of the optical axis and the detector plane) increases, τ increases accordingly due to the increase of the optical path difference.

Consequently, the overlap between the two wave-packets decreases, which leads to the decrease of the ability to interfere with each other, and hence of the visibility of the interference pattern decreases.

We shall notice that the duration of a wave-packet $\Delta\tau$ is inversely proportional to the spectral bandwidth $\Delta\omega$. Therefore, the size of the area in the detector plane where the two fields passing through the two pinholes are still correlated (giving rise to a visible interference pattern) ultimately depends on the spectral bandwidth $\Delta\omega$. The smaller the spectral bandwidth $\Delta\omega$ is, the larger the coherence area is.

In the extreme case when the interference pattern is generated by a monochromatic point source whose spectral bandwidth $\Delta\omega \to 0$, visibility will be uniformly 1 everywhere in the detector plane because the duration of the wave-packet $\Delta\tau \to \infty$. In practice, a monochromatic point source can be achieved by focusing a laser beam on a spatial filter (a diaphragm).

1.1.2. Spatial coherence

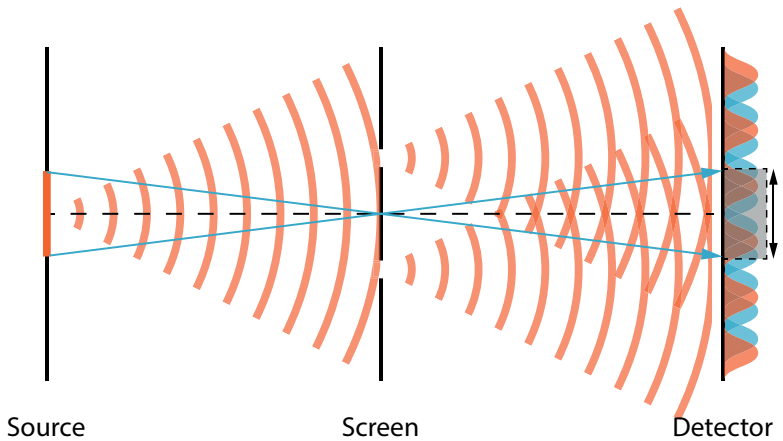

Figure 1.3: Demonstration of the origin of spatial coherence. A planar extended source, which consists of a collection of independent monochromatic point sources, is placed in the source plane. The total intensity distribution is the sum of the shifted interference pattern generated by each point source, and the shift is proportional to the location of the point source. The gray area is covered by the same pitch of different shifted interference pattern. If this area is larger than one pitch of the interference pattern, the fringes will be completely smeared and hence the visibility becomes zero.

As we have demonstrated that when only one monochromatic point source is present, the interference pattern will have uniform visibility in the detector plane. When moving the point source in the source plane, the interference pattern will shift accordingly as shown in Fig. 1.3. This is because the movement alters the phase difference between the two fields at the two pinholes.

In the presence of a collection of independent point sources, we can observe the smearing of the total intensity distribution because the shifted interference patterns superpose with each other. This effect is due to the spatial extension of the source and hence is referred to as the spatial coherence in order to be distinguished from

the temporal coherence effect, which is due to the extension of a wave-packet in the time domain.

In the case shown in Fig. 1.3, if the largest shift of the interference pattern (generated by the point source on the edge of the source) is larger than half pitch of the interference pattern, the fringes will then be completely smeared and hence the visibility becomes zero. Consequently, the fields at these two pinholes are uncorrelated, or in other words, spatially incoherent.

The correlation between the fields at these two pinholes depends on their separation, for a given size of the source. As the separation increases, the pitch of the interference pattern decreases, and so does the correlation. Therefore, using a monochromatic extended source instead of a point source will make the field in the screen plane become spatially partially coherent.

1.2. The theory of image formation

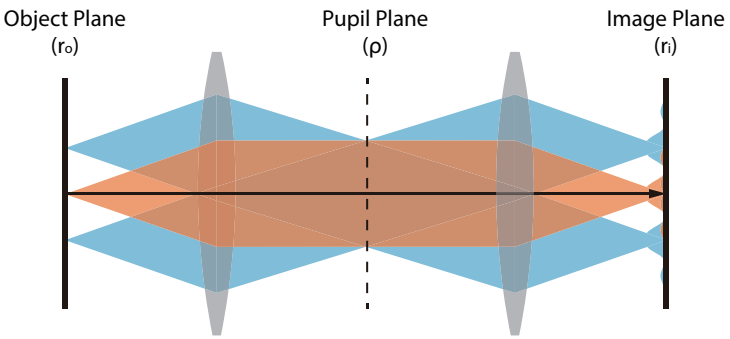

Figure 1.4: Illustration of an imaging process. Each point source that the object field consists of generates an image field. An image of the object is defined as the intensity of the total image field. The spatial coherence property of the object field has a significant impact on the image intensity.

Imaging systems transform an electromagnetic field in the object plane into a target electromagnetic field in the image plane. According to the Huygens' law, each point of the object field is itself a source. In an imaging process, due to the inevitable aberrations of the imaging system, the transformation is never perfect. The total image field is the sum of the image fields emitted by all point sources, which can be written as

$$E_i(\mathbf{r}_i) = \iint H(\mathbf{r}_i;\mathbf{r}_o)E_o(\mathbf{r}_o)\mathrm{d}\mathbf{r}_o. \tag{1.6}$$

where $H(\mathbf{r}_i;\mathbf{r}_i)$ represents the image field (a function of image plane location $\mathbf{r}_i$) generated by the point source at $\mathbf{r}_o$ in the object plane. $H(\mathbf{r}_i;\mathbf{r}_i)$ is also known as the point-spread function (PSF) of the imaging system. In most situations, PSF is assumed to be translation-invariant: $H(\mathbf{r}_i;\mathbf{r}_o) = H(\mathbf{r}_i - \mathbf{r}_o)$, which depends on only the relative distance between $\mathbf{r}_i$ and $\mathbf{r}_o$. Notice that here we assume a unit scaling ratio of the imaging system, namely no magnification or demagnification.

Because only the intensity can be measured by using charge coupled devices (CCD) or complementary metal oxide semiconductor (CMOS) sensors, the "image" of an object is referred to as the intensity of the total image field. The correlation between the image field generated by each point source of the object plays an important role in image formation. We can write the image intensity as

$$\begin{aligned} I(\mathbf{r}_i) &= \langle E_i(\mathbf{r}_i) E_i(\mathbf{r}_i)^* \rangle \\ &= \iint \iint H(\mathbf{r}_i - \mathbf{r}_{o1}) H(\mathbf{r}_i - \mathbf{r}_{o2})^* \langle E_o(\mathbf{r}_{o1}) E_o(\mathbf{r}_{o2})^* \rangle \, \mathrm{d}\mathbf{r}_{o1} \mathrm{d}\mathbf{r}_{o2}, \end{aligned} \tag{1.7}$$

where $\langle \cdot \rangle$ denotes the ensemble averaging. In Eq. (1.7), we define $J_o(\mathbf{r}_{o1}, \mathbf{r}_{o2}) = \langle E_o(\mathbf{r}_{o1}) E_o(\mathbf{r}_{o2})^* \rangle$ as the MCF of the object field, which describes the correlation between fields at $\mathbf{r}_{o1}$ and $\mathbf{r}_{o2}$. We can consider that the MCF $J_o(\mathbf{r}_{o1}, \mathbf{r}_{o2})$ determines the weight of the product $H(\mathbf{r}_i - \mathbf{r}_{o1}) H(\mathbf{r}_i - \mathbf{r}_{o2})^*$. The image intensity Eq. (1.7) thus represents a weighted sum for all possible combinations of $\mathbf{r}_{o1}$ and $\mathbf{r}_{o2}$.

Normally computing the image intensity using Eq. (1.7) is very time-consuming unless using fast methods based on approximations such as modal decomposition approaches [1, 2] or in the cases of translation-invariant MCF [3]. In the cases when the object field is completely coherent or completely incoherent, the computation of the image intensity can be greatly simplified as a convolution, denoted by $*$, which can be calculated using Fourier transform algorithms.

- Spatially coherent: The MCF of the object field is given by

$$J_o(\mathbf{r}_{o1}, \mathbf{r}_{o2}) = E_o(\mathbf{r}_{o1}) E_o(\mathbf{r}_{o2})^*, \tag{1.8}$$

  which indicates that $E(\mathbf{r}_{o1})$ and $E(\mathbf{r}_{o2})$ are correlated for all possible combination of $\mathbf{r}_{o1}$ and $\mathbf{r}_{o2}$. The image intensity is written as

$$I(\mathbf{r}_i) = \left| \iint H(\mathbf{r}_i - \mathbf{r}_o) E_o(\mathbf{r}_o) \mathrm{d}\mathbf{r}_o \right|^2 = \left| H(\mathbf{r}_i) * E_o(\mathbf{r}_i) \right|^2. \tag{1.9}$$

- Spatially incoherent: The MCF of the object field is given by

$$J_o(\mathbf{r}_{o1}, \mathbf{r}_{o2}) = E_o(\mathbf{r}_{o1}) E_o(\mathbf{r}_{o2})^* \delta(\mathbf{r}_{o1} - \mathbf{r}_{o2}), \tag{1.10}$$

  which indicates that $E(\mathbf{r}_{o1})$ and $E(\mathbf{r}_{o2})$ are correlated only when $\mathbf{r}_{o1} = \mathbf{r}_{o2}$ and are uncorrelated elsewhere when $\mathbf{r}_{o1} \neq \mathbf{r}_{o2}$. The image intensity is written as

$$I(\mathbf{r}_i) = \iint |H(\mathbf{r}_i - \mathbf{r}_o)|^2 |E_o(\mathbf{r}_o)|^2 \mathrm{d}\mathbf{r}_o = |H(\mathbf{r}_i)|^2 * |E_o(\mathbf{r}_i)|^2. \tag{1.11}$$

To summarize, for the calculation of the image intensity $I(\mathbf{r}_i)$, we use Eq. (1.9), the coherent imaging formula, when imaging a transmissive/reflective sample illuminated by a laser beam, and use (1.11), the incoherent imaging formula, in the case of e.g. fluorescent imaging. We remark that coherent effect may occur even when the object is illuminated using a incoherent source [4]. In the intermediate state when the imaging is only partially spatially coherent, characterizing the MCF of the object field and computing the image intensity are both very challenging.

1

1.3. Computation of the point-spread function

The point-spread function (PSF) is defined as the image plane field distribution generated by a point source in the object plane. The imaging system is usually simplified by using a black-box model for computing the PSF. This black-box is bounded by the entrance pupil and the exit pupil. The fields at these two pupils are diverging and converging spherical waves centered at the point source and its geometrical image, respectively.

Suppose that the field at the entrance pupil is ideal. The aberrations, caused by both the imaging system and the ambient medium, and the loss of light intensity are described by the amplitude and the phase of the pupil function (modulation to the field at the exit pupil), respectively.

The PSF can be obtained by computing the far-field diffraction of the exit pupil field in the image space. We define that the image space contains the focal plane, in which the geometrical image of the point source locates, and defocused planes, both of which are perpendicular to the optical axis.

In this thesis, we discuss only the computation of the PSF for low NA ($\mathrm{NA} < 0.6$) imaging systems. The situation is also known as the scalar case, in which the effect of polarization is neglected.

We denote coordinates of the exit pupil and the image space by $\mathbf{k}$ and $\mathbf{r}$, respectively. The axial coordinate in the image space is denoted by z with the origin at the intersection of the z axis and the image plane (focal plane). We remark that the location of the image plane is defined with respect to the object plane.

The field distribution of the PSF in the image space can be computed using the Debye diffraction integral [5, 6], which is effectively a Fourier transform from the exit pupil plane ($\mathbf{k}$) to the image space ($\mathbf{r}$):

$$PSF(\mathbf{r}, z) \propto \iint P(\mathbf{k}) \exp\left(i\pi z \frac{\mathrm{NA}^2}{\lambda} |\mathbf{k}|^2\right) \exp\left(-i2\pi \frac{\mathrm{NA}}{\lambda} \mathbf{k} \cdot \mathbf{r}\right) \mathrm{d}^2\mathbf{k} \tag{1.12}$$

where $P(\mathbf{k})$ is the pupil function of the imaging system. The amplitude is often assigned to be uniform (no loss of light intensity), while the phase is often described in terms of the Zernike polynomials. We can thus express the pupil function by

$$P(\mathbf{k}) = \exp\left\{i2\pi \sum_{m,n} \zeta_n^m Z_n^m(\mathbf{k})\right\}, \tag{1.13}$$

where ζ_n^m are Zernike coefficients in the unit of wavelength λ and $Z_n^m(\mathbf{k})$ are the Zernike polynomials defined as

$$Z_n^m(\rho, \theta) = \begin{cases} \sqrt{2(n+1)} R_n^{|m|}(\rho) \cos(|m|\theta), & m > 0 \\ \sqrt{n+1} R_n^{|m|}(\rho), & m = 0 \\ \sqrt{2(n+1)} R_n^{|m|}(\rho) \sin(|m|\theta), & m < 0 \end{cases}, \tag{1.14}$$

where ρ and θ are the radial and the azimuthal coordinates in the pupil plane, respectively, and n and m are the radial and the azimuthal orders, respectively.

Each Zernike polynomial $Z_n^m(\rho,\theta)$ corresponds to anke aberration and the absolute value of the Zernike coefficient $|\zeta_n^m|$ represents the weight of the root-mean-square wavefront error of the corresponding aberration.

The Zernike polynomials contain two parts: the radial part is given by

$$R_n^m(\rho) = \sum_{k=0}^{(n-m)/2} \frac{(-1)^k * (n-k)!}{k!(\frac{n+m}{2}-k)!(\frac{n-m}{2}-k)!}\rho^{n-2k}, \tag{1.15}$$

where ρ is normalized such that $0 \leq \rho \leq 1$, and the azimuthal part is a trigonometric function depending on the sign of order m.

The term $\exp(iz|\mathbf{k}|^2)$ in Eq. (1.12) represents the effect of defocus relative to the image plane ($z = 0$ represents the image plane). Normalizing the image plane coordinates $\mathbf{r}$ by λ/NA and the defocus distance z by $\lambda/(\pi \text{NA}^2)$, respectively. We can further write Eq. (1.12) as the Fourier transform of the pupil function and a defocus term:

$$PSF(\mathbf{r}', z) \propto F\left\{P(\mathbf{k})\exp(iz|\mathbf{k}|^2)\right\}(\mathbf{r}'), \tag{1.16}$$

The sampling of the pupil and the PSF should satisfy the Shannon-Nyquist sampling theorem. Let the length and the interval of the sampling in the pupil plane and in the image plane be denoted by $(L_\mathbf{k}, \Delta_\mathbf{k})$ and $(L_\mathbf{r}, \Delta_\mathbf{r})$, respectively. The sampling theorem states that

$$\begin{aligned} L_\mathbf{k} &\geq \frac{\lambda}{\text{NA}\Delta_\mathbf{r}}, \\ \Delta_\mathbf{k} &\leq \frac{\lambda}{\text{NA}L_\mathbf{r}}. \end{aligned} \tag{1.17}$$

In order to computed the PSF on the given sampling grid (for example to mimic the measurement using a sensor array), the sampling of the pupil should satisfy Eq. (1.17) and vice versa. However, the satisfaction of Eq. (1.17) does not naturally guarantee the optimal accuracy of PSF computation. Usually oversampling is required to avoid the aliasing effect on the boundary of the pupil. Eq (1.16) can be computed using the fast Fourier transform (FFT) or the chirp-z transform (CZT) algorithm [7, 8]. A recent alternative is the semi-analytical approach: the extended Nijboer-Zernike (ENZ) theory [9, 10].

1.4. The inverse problem

Inverse problem is the process of inferring the parameters of a physical model from the measurements. The physical model should take the parameters as the input and produce the measurements as output. Typical inverse problems in optics are diffractive imaging [11–13], image based aberration retrieval [14–16], image restoration [17, 18], 3-dimensional profile metrology in scatterometry for lithography [19, 20], etc.

An inverse problem is said to be well-posed, when the physical model satisfies three criteria:

- a solution exists,

1

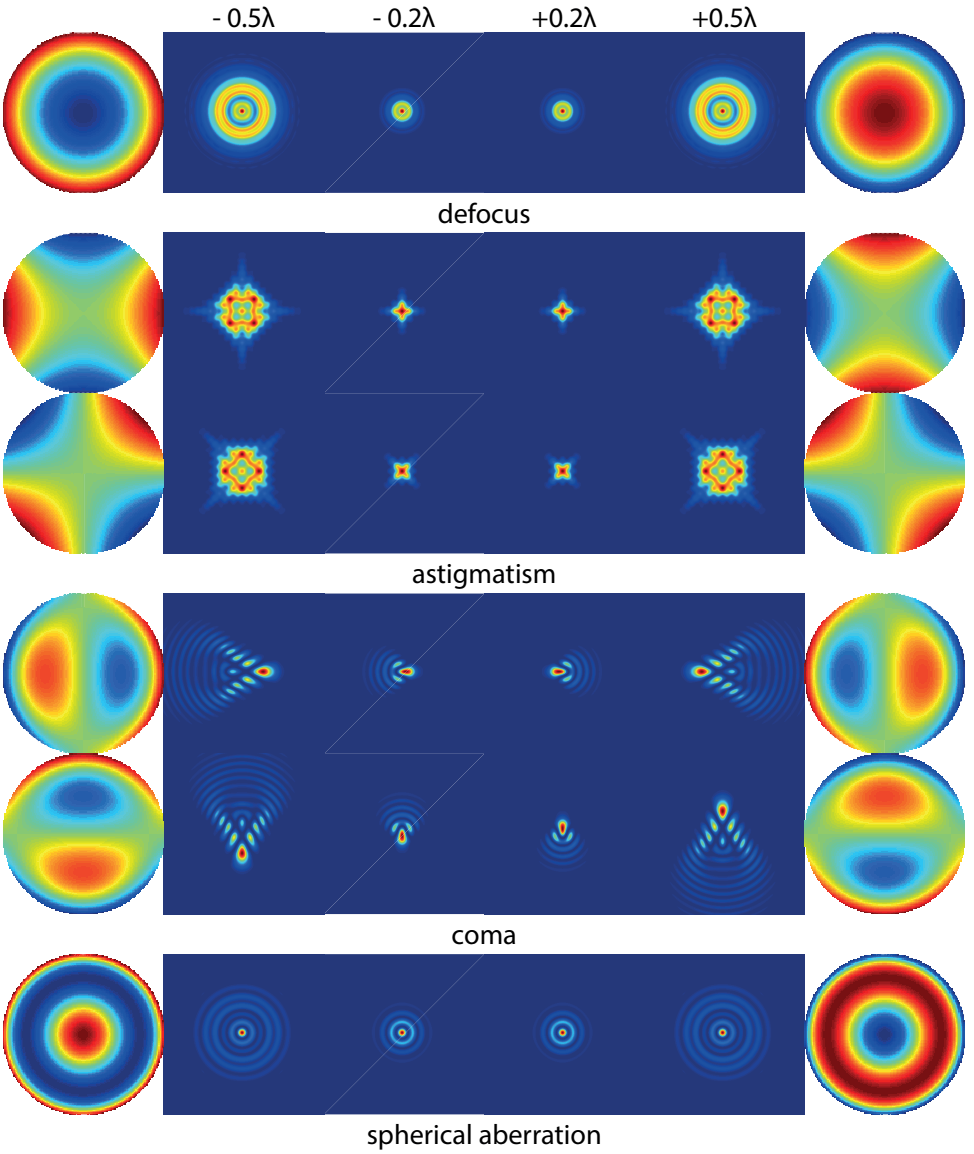

Figure 1.5: The phase of the pupil function and the corresponding PSF intensity for various aberrations.

- the solution is unique,
- the solution depends on the initial condition in a continuous way.

When any of the three criteria is violated, the inverse problem is said to be ill-posed. Particularly, the inverse problem is said to be ill-conditioned when the last criteria is violated. In ill-conditioned inverse problem, a small change in the initial condition leads to an arbitrary large change in the solution. The solution is thus not stable and sensitive to the measurement noise.

Unfortunately, most of the inverse problems are ill-posed. In this section, we will discuss how to formulate an inverse problem using the maximum likelihood estimation based on the noise model and then study how to solve the inverse problem for a nonlinear physical model.

1.4.1. Maximum likelihood estimation

The maximum likelihood estimation (MLE) estimates, for a given physical model, the parameters that most likely reproduce the measurements. Suppose that the measurements (with noise) taken by the pixelated detector and the predictions (without noise) made by the physical model are $g(x,y)$ and $f(x,y)$, respectively.

Due to the noise, the number of photons detected (proportional to the intensity) at each pixel can be modeled as an independent and identically distributed (i.i.d.) random variable. Provided the probability density that the measured intensity obeys, we can formulate a likelihood function that evaluates the likelihood of a given measurement being detected. Thermal noise (Johnson-Nyquist noise) and photon shot noise are typical noises in the detectors. MLE finds the parameters that most likely reproduce the noisy measurements by maximizing the likelihood function or, equivalently, minimizing its logarithm.

Consider the imaging problem as an example. The measured and the predicted image are $g(x,y)$ and $f(x,y;\boldsymbol{\gamma})$, respectively. We assume that the model of imaging depends on the parameters $\boldsymbol{\gamma} = [\gamma_1,\ \gamma_2,\ \cdots]$. Both $g(x,y)$ and $f(x,y;\boldsymbol{\gamma})$ represent distributions of intensities.

Assuming that the dominant noise of the measurement is the thermal noise generated by the thermal agitations of the electronic components of the detector. For the pixel at (x,y), the noisy image intensity $g(x,y)$ is a random variable whose probability density follows the Gaussian distribution with mean given by the perfect (noise free) image intensity $f(x,y;\boldsymbol{\gamma})$ and variance σ_n^2:

$$P_{G,(x,y)}[g(x,y);\boldsymbol{\gamma},\sigma_n^2] = \frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left\{-\frac{[g(x,y)-f(x,y;\boldsymbol{\gamma})]^2}{2\sigma_n^2}\right\}. \tag{1.18}$$

Therefore, the probability density for the entire image is given by

$$P_G[g(x,y);\boldsymbol{\gamma},\sigma_n^2] = \prod_{x,y}\frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left\{-\frac{[g(x,y)-f(x,y;\boldsymbol{\gamma})]^2}{2\sigma_n^2}\right\}. \tag{1.19}$$

Eq. (1.19) is the likelihood function in the case of thermal noise. It is customary to use the natural logarithm of Eq. (1.19) which is given by (assuming that the noise

at each pixel is i.i.d. random variable)

$$\begin{aligned}\ln P_G[g(x,y);\boldsymbol{\gamma},\sigma_n^2] &= \frac{1}{\sqrt{2\pi\sigma_n^2}}\sum_{x,y}-\frac{[g(x,y)-f(x,y;\boldsymbol{\gamma})]^2}{2\sigma_n^2}\\ &\propto -\sum_{x,y}[g(x,y)-f(x,y;\boldsymbol{\gamma})]^2.\end{aligned} \tag{1.20}$$

Notice that the logarithmic likelihood function for thermal noise Eq. (1.20) can be interpreted as the L-2 norm of the difference between the measurement $g(x,y)$ and the prediction $f(x,y;\boldsymbol{\gamma})$. So MLE is equivalent to the method of least squares because maximizing Eq. (1.20) is equivalent to minimizing the least squares.

An alternative assumption is that the dominant noise of the image is due to the fluctuation of the number of photons detected. At pixel (x,y), the number of detected photons $g(x,y)$ follows the Poisson distribution with reference given by the expected photons $f(x,y;\boldsymbol{\gamma})$:

$$P_{P,(x,y)}[g(x,y);\boldsymbol{\gamma}] = \frac{f(x,y;\boldsymbol{\gamma})^{g(x,y)}}{g(x,y)!}\exp[-f(x,y;\boldsymbol{\gamma})], \tag{1.21}$$

The probability density for the entire being measured is given by

$$P_P[g(x,y);\boldsymbol{\gamma},\sigma_n^2] = \prod_{x,y}\frac{f(x,y;\boldsymbol{\gamma})^{g(x,y)}}{g(x,y)!}\exp[-f(x,y;\boldsymbol{\gamma})]. \tag{1.22}$$

Taking the natural logarithm of Eq. (1.22) the photon shot noise likelihood function, we obtain:

$$\begin{aligned}\ln P_P[g(x,y);\boldsymbol{\gamma},\sigma_n^2] &= \sum_{x,y}[g(x,y)\ln f(x,y;\boldsymbol{\gamma}) - \ln g(x,y)! - f(x,y;\boldsymbol{\gamma})]\\ &\propto \sum_{x,y}[g(x,y)\ln f(x,y;\boldsymbol{\gamma}) - f(x,y;\boldsymbol{\gamma})],\end{aligned} \tag{1.23}$$

where the constant term $\ln g(x,y)!$ has been neglected. We shall notice that when the expected number of photons $f(x,y)$ is large, the Poisson distribution that $g(x,y)$ obeys can be approximated by the Gaussian distribution with mean and variance both equal to $f(x,y)$.

1.4.2. Gradient descent method for optimization

Consider a physical model that depends non-linearly on the parameters $\boldsymbol{\gamma}$. For such a non-linear inverse problem, no analytical solution of $\boldsymbol{\gamma}$ can be derived and instead we find the solution of $\boldsymbol{\gamma}$ numerically via iterative optimization. In this thesis we discuss the gradient descent method for optimization, which determines the update of $\boldsymbol{\gamma}$ based on the gradient of an error function with respect to $\boldsymbol{\gamma}$.

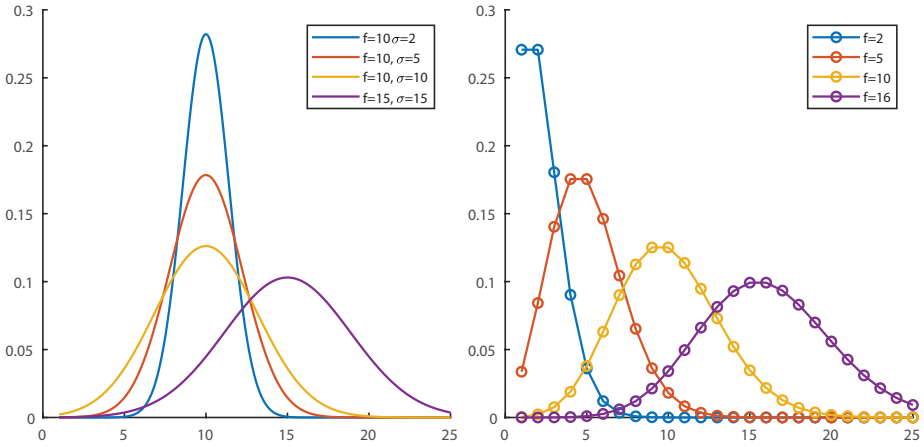

Figure 1.6: The logarithmic likelihood function for thermal noise (left) and photon shot noise (right).

Let us consider the imaging problem as an example again. The parameters $\boldsymbol{\gamma}$ can be regarded as the aberrations. The first step is to define an error function using the least squares as follows:

$$\mathcal{L}(\boldsymbol{\gamma}) = \iint [g(x,y) - f(x,y;\boldsymbol{\gamma})]^2 \, \mathrm{d}x\mathrm{d}y, \tag{1.24}$$

Defining the error function as Eq. (1.24) indicates that the dominant noise of $g(x,y)$ is additive and the probability density follows the Gaussian distribution with mean equal to $f(x,y;\boldsymbol{\gamma})$ and variance given by the noise level.

The gradient descent method starts with a initial guess of $\boldsymbol{\gamma}$. In each iteration, $\boldsymbol{\gamma}$ is updated along a certain direction by a certain step size, and the iterative process repeats until the error function converges, i.e. $\mathcal{L}(\boldsymbol{\gamma})$ or the variation of $\mathcal{L}(\boldsymbol{\gamma})$ is sufficiently small. The iterative process will also stop if any of the constraints, e.g. the maximum iteration number or the maximum time duration, is violated. We illustrated a flow chart of the optimization in Fig. 1.7.

To determine the update direction of $\boldsymbol{\gamma}$, we define the variation of the error function with respect to the aberration as

$$\delta\mathcal{L}(\boldsymbol{\gamma}) = \mathcal{L}(\boldsymbol{\gamma} + \epsilon\boldsymbol{\eta}) - \mathcal{L}(\boldsymbol{\gamma}), \tag{1.25}$$

where ϵ is a number and $\boldsymbol{\eta}(x,y)$ is an arbitrary vector. Taking the Taylor's expansion of $L(\boldsymbol{\gamma} + \epsilon\boldsymbol{\eta})$ at $\epsilon = 0$, we obtain

$$\mathcal{L}(\boldsymbol{\gamma} + \epsilon\boldsymbol{\eta}) = \mathcal{L}(\boldsymbol{\gamma}) + \left[\nabla\mathcal{L}(\boldsymbol{\gamma} + \epsilon\boldsymbol{\eta})\big|_{\epsilon=0}\right]^{\mathrm{T}} \boldsymbol{\eta} + \mathcal{O}(\epsilon), \tag{1.26}$$

where

$$\nabla\mathcal{L}(\boldsymbol{\gamma} + \epsilon\boldsymbol{\eta})\big|_{\epsilon=0} = \left[\frac{\partial\mathcal{L}(\boldsymbol{\gamma})}{\partial\gamma_1}, \frac{\partial\mathcal{L}(\boldsymbol{\gamma})}{\partial\gamma_2}, \cdots\right]$$

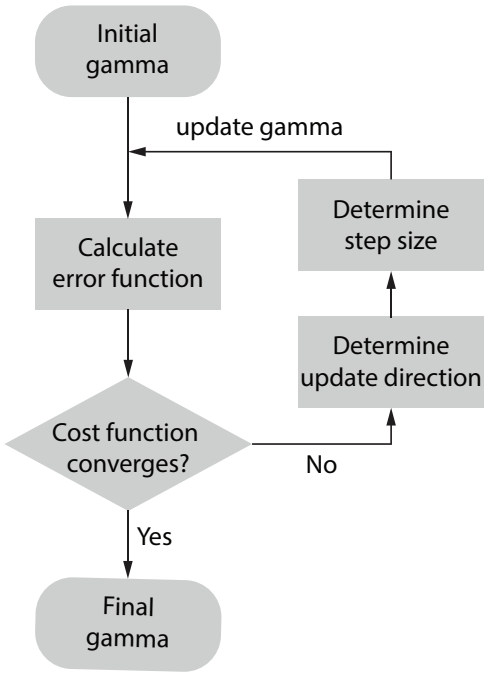

Figure 1.7: The flow chart of the optimization algorithm for solving the inverse problem.

The variation of the error function with respect to $\boldsymbol{\gamma}$ thus is given by

$$\delta\mathcal{L}(\boldsymbol{\gamma}) = \nabla\mathcal{L}(\boldsymbol{\gamma})^{\mathrm{T}}\boldsymbol{\eta}, \tag{1.27}$$

where we have kept only the first order term and neglected the higher order terms. Eq. (1.27) indicates that the variation of the error function depends on the choice of the arbitrary function $\boldsymbol{\eta}$. Because we aim to find $\boldsymbol{\gamma}$ that minimizes the error function, in each iteration, we want to update $\boldsymbol{\gamma}$ in the direction along which the descent of the error function is the steepest. Consequently, the arbitrary function $\boldsymbol{\eta}$ should be

$$\boldsymbol{\eta} \propto -\nabla\mathcal{L}(\boldsymbol{\gamma}) \tag{1.28}$$

and the variation of the error function becomes

$$\delta\mathcal{L}(\boldsymbol{\gamma}) = -\alpha\|\nabla\mathcal{L}(\boldsymbol{\gamma})\|_2^2 \tag{1.29}$$

where α is a scalar that represents the step size and $\|\cdot\|_2^2$ is the L-2 norm.

The determination of the step size along the update direction of $\boldsymbol{\gamma}$ is a question that deserves investigation. One way is to find the optimal α using the line search method which solves the following optimization problem:

$$\alpha = \arg\min_{\alpha} \mathcal{L}(\boldsymbol{\gamma} + \alpha\boldsymbol{\eta}) \tag{1.30}$$

Although the line search method finds α that decreases the error function the most along $\boldsymbol{\eta}$, a complete optimization, which is usually time-consuming, is required in each iteration. An alternative is to find the optimal α that satisfies the following condition:

$$\mathcal{L}(\boldsymbol{\gamma} + \alpha\boldsymbol{\eta})) \leq \mathcal{L}(\boldsymbol{\gamma}) - \alpha\delta, \tag{1.31}$$

where $\delta \in (0, 1)$ is the control parameter of the condition. In each iteration, the initial step size α is reduced by a factor of $\Delta \in (0, 1)$ until the condition is satisfied.

Another approach is to determine α by assuming that the cost function $\mathcal{L}$ shows an approximate parabolic behavior locally as function of α. The value of α yielding a minimum of the local parabolic approximation for $\mathcal{L}$ can then simply be derived from the derivative of this parabolic function.

References

[1] L. Mandel and E. Wolf, *Optical coherence and quantum optics* (Cambridge university press, 1995).

[2] K. Kim and E. Wolf, *A scalar-mode representation of stochastic, planar, electromagnetic sources,* Optics communications **261**, 19 (2006).

[3] M. Singh, H. Lajunen, J. Tervo, and J. Turunen, *Imaging with partially coherent light: elementary-field approach,* Optics express **23**, 28132 (2015).

[4] G. O. Reynolds and J. B. DeVelis, *Review of optical coherence effects in instrument design part ii,* Optical Engineering **20**, 204671 (1981).

[5] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light* (Elsevier, 2013).

[6] E. Wolf, *Electromagnetic diffraction in optical systems-i. an integral representation of the image field,* Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **253**, 349 (1959).

[7] L. Rabiner, R. Schafer, and C. Rader, *The chirp z-transform algorithm,* IEEE transactions on audio and electroacoustics **17**, 86 (1969).

[8] M. Leutenegger, R. Rao, R. A. Leitgeb, and T. Lasser, *Fast focus field calculations,* Optics express **14**, 11277 (2006).

[9] J. Braat, P. Dirksen, and A. J. Janssen, *Assessment of an extended nijboer–zernike approach for the computation of optical point-spread functions,* JOSA A **19**, 858 (2002).

[10] S. Van Haver, *The extended nijboer-zernike diffraction theory and its applications,* (2010).

[11] J. R. Fienup, *Reconstruction of a complex-valued object from the modulus of its fourier transform using a support constraint,* JOSA A **4**, 118 (1987).

[12] H. Faulkner and J. Rodenburg, *Movable aperture lensless transmission microscopy: a novel phase retrieval algorithm,* Physical review letters **93**, 023903 (2004).

[13] M. Guizar-Sicairos and J. R. Fienup, *Phase retrieval with transverse translation diversity: a nonlinear optimization approach,* Optics express **16**, 7264 (2008).

[14] R. G. Paxman, T. J. Schulz, and J. R. Fienup, *Joint estimation of object and aberrations by using phase diversity,* JOSA A **9**, 1072 (1992).

[15] M. G. Löfdahl and G. Scharmer, *Wavefront sensing and image restoration from focused and defocused solar images.* Astronomy and Astrophysics Supplement Series **107**, 243 (1994).

[16] R. G. Paxman, J. H. Seldin, M. G. Lofdahl, G. B. Scharmer, and C. U. Keller, *Evaluation of phase-diversity techniques for solar-image restoration,* (1995).

[17] J. M. Bioucas-Dias and M. A. Figueiredo, *A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration,* IEEE Transactions on Image processing **16**, 2992 (2007).

[18] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems,* SIAM journal on imaging sciences **2**, 183 (2009).

[19] H. Cramer, A. Chen, F. Li, P. Leray, A.-L. Charley, L. Van Look, J. Bekaert, and S. Cheng, *High-speed, full 3d feature metrology for litho monitoring, matching, and model calibration with scatterometry,* in *Metrology, Inspection, and Process Control for Microlithography XXVI*, Vol. 8324 (International Society for Optics and Photonics, 2012) p. 83240R.

[20] A. den Boef, H. Cramer, S. Petra, B. O. F. Auer, J. Schmetz-Schagen, A. Koolen, O. van Loon, G. de Gersem, P. Klandermans, and E. Bakker, *Scatterometry for advanced process control in semiconductor device manufacturing,* in *Fifth International Conference on Optical and Photonics Engineering*, Vol. 10449 (International Society for Optics and Photonics, 2017) p. 1044916.

2

Aberrations Retrieval using Phase Diversity Method

2.1. Background

For advanced imaging systems such as microscope, telescope, and lithographic projection lens, image blurring is mainly caused by the error of the wavefront at the pupil. The ideal wavefront generated by any point source in the object plane has for an perfect imaging system a spherical shape which is centered at the location of the geometrical image of this point source in the image plane. Each type of error, referred to as the aberration, leads to a distinct kind of blurring of the image and hence needs to be dealt with. Typically there are two origins of aberrations: external aberrations, which are caused by the non-uniformity of the ambient medium, for example imaging through a biomedical tissue sample or a turbulent atmosphere, and internal aberrations, which are often caused by vibration, contamination or heating during operation.

One method to deal with the aberrations is to use adaptive optics, which has been studied excessively in astronomical and biomedical imaging. In this method the error of the wavefront at the pupil of the imaging system is measured and then compensated by an active device, e.g. a deformable mirror (DM) or a spatial light modulator (SLM). It is common to use wavefront sensing techniques such as interferometry [1, 2] or a wavefront sensor [3] to measure the wavefront error.

These techniques require a point source for providing the ideal wavefront as reference and an additional imaging system for mapping the actual wavefront onto a detector. The accuracy of these techniques thus relies on the quality of the reference. The additional imaging system also has its own wavefront error that mixes with the wavefront error of the imaging system that is to be measured. This is usually known as the non-common path error, i.e. the imaging and the measurement of the aberrations do not share one common path.

In astronomy [4], either a natural star or an artificial laser guide star is used as a point source. When imaging the eye using scanning laser ophthalmoscopy, it is possible to regard the light scattered by retina cells as if emitted by a point source [5]. In other applications, such a point source maynot be readily available.

Combining an active device and a camera provides a possibility for measuring the wavefront error which differs from the conventional methods mentioned above. This modal approach was proposed by Martin Booth in [6] for confocal microscopy and was later applied to a series of microscopic techniques [7–9]. It works in a probe-and-test manner: the original wavefront is perturbed by a particular mode of aberration, and the information that this perturbation provides is measured in the image. As a consequence, measuring N aberration modes requires at least $2N + 1$ images: each mode requires an original unbiased image and two perturbed images (by opposite bias).

Alternative aberration retrieval methods are based on the propagation of light in free-space. These methods solve the following problem: retrieve the phase of light (the wavefront) in the pupil from the intensity of its Fourier transform, e.g. from the intensity of the point-spread function (PSF), in the image plane. This method was first reported by Gerchberg and Saxton in 1971 [10, 11] who additionally used the distribution of light intensity in the pupil as constraint which had to be assumed to be known. Further development of this method was proposed by Fienup [12, 13]

who used only the support constraint of the light in the pupil.

Measuring the intensity distribution in several planes in the focal volume allows the retrieval of the phase of the light in one of the through-focus planes. There exist two types of methods for achieving this goal. The deterministic methods [14–16] search for the solution to the transport-of-intensity equation (TIE), which requires measuring both the intensity distribution in the plane of interest and its derivative along the direction of the optical axis. The non-deterministic methods [17, 18] aim to find the phase in one plane by propagating back-and-forth among a series of planes: in each plane, the phase is kept, while the amplitude is replaced by the square root of the measured intensity.

In the case of incoherent imaging, the object consists of a collection of mutually incoherent point sources. Most of the above mentioned techniques and methods cannot be used (except for the probe-and-test method [6–9]) in this case because the wavefront in the pupil will be an incoherent superposition of the wavefronts generated by all point sources. In 1992, Paxman *et al.* proposed to retrieve the aberrations from the images of an unknown object using the phase diversity method [19]. In addition to the original image, at least one more image is measured after the phase of the pupil function has been perturbed in a known fashion. Defocus variation, which is introduced by varying the location of the camera sensor on the optical axis, is a commonly used phase diversity. The phase diversity method has been implemented successfully in solar imaging [20, 21].

In contrast to other mentioned aberration measurement and retrieval methods, the phase diversity method can be applied to incoherent imaging of an unknown object. By solving a minimization problem that is independent of the unknown object with the phase diversity method, the aberrations are retrieved and corrected using software instead of hardware. This feature makes it a cost-effective method: it requires only a 1-dimensional translation stage (along the optical axis perpendicular to the image plane) on which the camera sensor is mounted.

In the phase diversity method, only the aberrations that blur the image are retrieved from the image. In other methods, usually the wavefront of a point source is measured and hence all aberrations can be determined in principle. However, the accuracy is limited by the fact that the wavefront determined by a wavefront sensor or by interferometry differs from the actual wavefront that causes the blurring of the image, because the optical path of the wavefront measurement differs from the optical path of imaging. Therefore, to deblur the image of a particular object, the phase retrieval method is superior over other methods.

In this chapter, we investigate the phase diversity by performing a proof-of-principle experiment using a microscope objective as imaging system. We study how the choice of the algorithm, the regularization, and other factors can influence the optimization, and we propose a method to calibrate the measurement settings.

2.2. The phase diversity method

For incoherent imaging, the object is considered to consist of a collection of mutually incoherent point sources. As it is shown in Fig. 2.1, each point source in the object plane generates an image in the image plane, which is known as the PSF. We

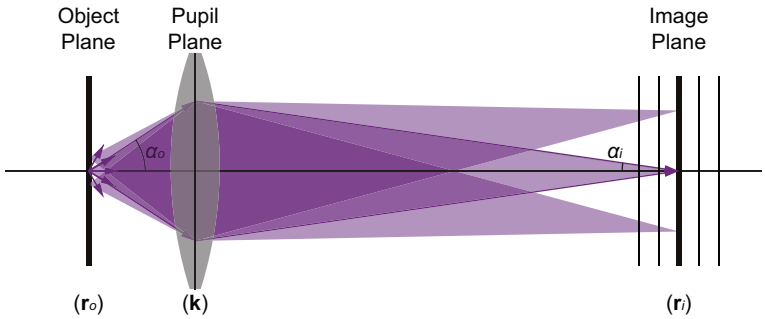

Figure 2.1: Schematic plot of the incoherent imaging system: a microscope objective. Each point source in the object plane generates a PSF in the image plane, whose shape is influenced by the aberrations of this imaging system. The image intensity is the superposition of all individual PSF intensities. Varying the location of the image plane along the optical axis varies the image intensity, which is due to the variation of PSF due to defocus aberration.

assume that the imaging system is shift-invariant (also known as isoplanatic). Then the PSF is independent of the location of the point source in the field of view (FOV). Commercial imaging systems are usually designed to be shift-invariant in the entire FOV (often referred to as the isoplanatic patch), while others are only shift-invariant in a sufficiently small sub-region of the FOV. We will stick to the shift-invariant assumption in the present chapter.

We can express the image formation formula by

$$I(\mathbf{r}_i) = \int H(\mathbf{r}_i - \mathbf{r}_o) O(\mathbf{r}_o) \mathrm{d}\mathbf{r}_o, \tag{2.1}$$

where $\mathbf{r}_o$ and $\mathbf{r}_i$ are 2-dimensional coordinates of the object and image plane, $O(\mathbf{r}_i)$ and $I(\mathbf{r}_o)$ are intensities in the object and image plane, respectively, and $H(\mathbf{r}_i - \mathbf{r}_o)$ is the intensity of the shift-invariant PSF. In this formula, $\mathbf{r}_o$ and $\mathbf{r}_i$ are related by $\mathbf{r}_i = \sigma \mathbf{r}_o$, where $\sigma = \alpha_o / \alpha_i$ is a constant scaling factor given by the ratio between the numerical apertures (NAs) of the object space α_o and the image space α_i.

By Fourier transforming both sides of Eq. (2.1), we obtain:

$$\hat{I}(\mathbf{k}) = \hat{H}(\mathbf{k}) \hat{O}(\mathbf{k}), \tag{2.2}$$

where $\hat{}$ represents the Fourier transform, and $\mathbf{k}$ denotes the pupil plane coordinate, which can also be referred to as the spatial frequency. $\hat{O}(\mathbf{k})$ and $\hat{I}(\mathbf{k})$ are the object and image spectrum, respectively, and $\hat{H}(\mathbf{k})$ is the PSF spectrum, which is also called the optical transfer function (OTF). Suppose that the wavelength of illumination is λ. Then the sampling of $\mathbf{k}$ should satisfy the Shannon-Nyquist sampling theorem with respect to the normalized image plane coordinate $\mathbf{r}_i / (\lambda / \alpha_i)$.

Usually, in the phase diversity method defocus variation is used by varying the image plane along the optical axis. We denote the image plane location by z and set the origin $z = 0$ to be at the location of the nominal best image plane (the plane in which the image quality is the best in absence of aberrations except for

the defocus). In the image plane at z, the OTF is given by

$$\hat{H}_z(\mathbf{k}; P) = \mathcal{F}\{H_z(\mathbf{r}; P)\}, \tag{2.3}$$

where $\mathcal{F}$ denotes the Fourier transform operator, $P(\mathbf{k})$ is the pupil function, and $H_z(\mathbf{r}; P)$ is the PSF intensity, which is given by

$$H_z(\mathbf{r}; P) = \left|\mathcal{F}\left\{P(\mathbf{k}) \exp(-iz|\mathbf{k}|^2)\right\}\right|^2, \tag{2.4}$$

where $\mathbf{r} = \mathbf{r}_i/(\lambda/\alpha_i)$ and $z = z_0/[\lambda/(\pi\alpha_i)^2]$, where z_0 is the actual axial coordinate, are the normalized lateral and axial coordinate, respectively. The image spectrum in the image plane at z is given by

$$\hat{I}_z(\mathbf{k}) = \hat{H}_z(\mathbf{k}; P)\hat{O}(\mathbf{k}). \tag{2.5}$$

Both $\mathbf{k}$ and z are dimensionless because of the normalization. Eq. (2.5) implies that for fixed object spectrum $\hat{O}(\mathbf{k})$ and pupil function $P(\mathbf{k})$, varying the image plane location z varies the image spectrum $\hat{I}_z(\mathbf{k})$ due to the defocus aberration in the OTF $\hat{H}_z(\mathbf{k}; P)$. The pupil function describes the properties of the imaging system. For example, the deviation of its phase from a constant represents the aberrations of the imaging system.

By capturing images in a series of image planes and subsequently computing their 2-dimensional Fourier transforms, we can obtain a system of equations, which depends linearly on the unknown object spectrum $\hat{O}(\mathbf{k})$ and nonlinearly on the unknown pupil function $P(\mathbf{k})$:

$$\begin{aligned} \hat{I}_{z_1}(\mathbf{k}) &= \hat{H}_{z_1}(\mathbf{k}; P)\hat{O}(\mathbf{k}), \\ \hat{I}_{z_2}(\mathbf{k}) &= \hat{H}_{z_2}(\mathbf{k}; P)\hat{O}(\mathbf{k}), \\ \vdots \quad &= \qquad \vdots \\ \hat{I}_{z_\ell}(\mathbf{k}) &= \hat{H}_{z_\ell}(\mathbf{k}; P)\hat{O}(\mathbf{k}), \end{aligned} \tag{2.6}$$

where $z_1, z_2, \cdots z_\ell$ are the locations of the image planes. The image spectrum $\hat{I}_z(\mathbf{k})$ is obtained by taking the Fourier transform of the measured intensity distribution and the OTF $\hat{H}_z(\mathbf{k}; P)$ is computed based on the pupil function $P(\mathbf{k})$ and the coordinates ($\mathbf{r}$ and z) of the planes in which the image is measured. In order to solve for the two unknowns, we need a system consisting of at least two equations: two images measured at two different locations but generated by the same object, i.e. using the same imaging system. Usually, we take one focal image in the nominal best image plane at $z = 0$, which is supposed to be the clearest, and one additional blurred defocused image in an image plane at $z \neq 0$.

In the phase diversity method the minimum is computed of an error function which is the L-2 norm of the difference between $\hat{I}_z(\mathbf{k})$, as obtained by taking the Fourier transform of the measured image, and the predicted $\hat{H}_z(\mathbf{k}, P)\hat{O}(\mathbf{k})$:

$$\begin{aligned} \mathcal{L}\left(\hat{O}, P\right) &= \sum_z \left\|\hat{I}_z(\mathbf{k}) - \hat{H}_z(\mathbf{k}; P)\hat{O}(\mathbf{k})\right\|_2^2 \\ &= \sum_z \int \left|\hat{I}_z(\mathbf{k}) - \hat{H}_z(\mathbf{k}; P)\hat{O}(\mathbf{k})\right|^2 \mathrm{d}\mathbf{k}, \end{aligned} \tag{2.7}$$

In case the exact $\hat{O}(\mathbf{k})$ and $P(\mathbf{k})$ are known and are substituted in Eq. (2.7), this difference actually gives the L-2 norm of the noise over every measurement plane. By defining the error function in Eq. (2.7) as the squared sum of the differences, we implicitly assume that the dominating noise obeys a zero mean Gaussian distribution in [19]. In other situations, the dominating noise may obey a Poisson distribution which depends on the distribution of image intensity, and hence the justification of using Eq. (2.7) is not valid. The key factor here is the level of the light intensity. Thermal noise (Gaussian distributed noise) is dominating in most situations, while shot noise (Poisson distributed noise) is dominating when the light intensity is very low, e.g. equal to the energy of relatively low number of photons.

As is seen in Eq. (2.6), the image spectrum depends linearly on $\hat{O}(\mathbf{k})$ but very nonlinearly on $P(\mathbf{k})$ through the OTF. Therefore, in the phase diversity method first a closed-form expression is derived for the object spectrum in terms of $\hat{I}_z(\mathbf{k})$ and $\hat{H}_z(\mathbf{k};P)$ in which a pupil function is assumed. By substituting the closed-form expression into Eq. (2.7), the unknown object spectrum $\hat{O}(\mathbf{k})$ is eliminated and a problem for only the pupil function $P(\mathbf{k})$ remains. After solving the corresponding problem for $P(\mathbf{k})$ by iterative optimization, the object spectrum is finally reconstructed using the closed-form expression.

2.2.1. The expression for the object spectrum

Our goal is to derive an expression for the object spectrum $\hat{O}(\mathbf{k})$ in terms of the image spectrum $\hat{I}_z(\mathbf{k})$ and the OTF $\hat{H}_z(\mathbf{k};P)$ in all measurement planes. We temporarily treat the pupil function $P(\mathbf{k})$ as if known and hence $\hat{O}(\mathbf{k})$ is the only unknown. Because the expression for $\hat{O}(\mathbf{k})$ should minimizes the error function Eq. (2.7):

$$\hat{O}(\mathbf{k}) = \min_{\hat{O}(\mathbf{k})} \mathcal{L}\left(\hat{O}\right), \tag{2.8}$$

we should set the functional derivative of the error function with respect to $\hat{O}(\mathbf{k})$ equal to zero. Notice that $\hat{O}(\mathbf{k})$ is a complex-valued function, which is given by

$$\hat{O}(\mathbf{k}) = \hat{O}_{\Re}(\mathbf{k}) + i\hat{O}_{\Im}(\mathbf{k}), \tag{2.9}$$

where $\hat{O}_{\Re}(\mathbf{k})$ and $\hat{O}_{\Im}(\mathbf{k})$ are the real and imaginary part, respectively. We can derive the derivative of the error function with respect to the real and imaginary parts:

$$\begin{aligned}
\frac{\partial \mathcal{L}\left(\hat{O}\right)}{\partial \hat{O}_{\Re}} &= \lim_{\epsilon\to 0} \frac{\mathcal{L}\left(\hat{O}+\epsilon\hat{O}_{\Re}\right)-\mathcal{L}\left(\hat{O}\right)}{\epsilon} \\
&= \lim_{\epsilon\to 0} \frac{\sum_z \left\|\left(\hat{I}_z-\hat{H}_z\hat{O}\right)-\epsilon\hat{H}_z\hat{O}_{\Re}\right\|_2^2 - \sum_z \left\|\hat{I}_z-\hat{H}_z\hat{O}\right\|_2^2}{\epsilon} \\
&\approx \sum_z \int -2\Re\left\{\hat{H}_z(\mathbf{k})\hat{O}_{\Re}(\mathbf{k})\left[\hat{I}_z(\mathbf{k})-\hat{H}_z(\mathbf{k})\hat{O}(\mathbf{k})\right]^*\right\}\mathrm{d}\mathbf{k},
\end{aligned} \tag{2.10}$$

and

$$\begin{aligned}\frac{\partial \mathcal{L}\left(\hat{O}\right)}{\partial \hat{O}_{\Im}} &= \lim_{\epsilon\to 0}\frac{\mathcal{L}\left(\hat{O}+i\epsilon\hat{O}_{\Im}\right)-\mathcal{L}\left(\hat{O}\right)}{\epsilon}\\ &= \lim_{\epsilon\to 0}\frac{\sum_z\left\|\left(\hat{I}_z-\hat{H}_z\hat{O}\right)-i\epsilon\hat{H}_z\hat{O}_{\Im}\right\|_2^2-\sum_z\left\|\hat{I}_z-\hat{H}_z\hat{O}\right\|_2^2}{\epsilon}\\ &\approx \sum_z\int -2\Re\left\{i\hat{H}_z(\mathbf{k})\hat{O}_{\Im}(\mathbf{k})\left[\hat{I}_z(\mathbf{k})-\hat{H}_z(\mathbf{k})\hat{O}(\mathbf{k})\right]^*\right\}\mathrm{d}\mathbf{k}.\end{aligned} \tag{2.11}$$

As a result, the derivative of the error function with respect to the object spectrum is

$$\begin{aligned}\frac{\partial \mathcal{L}\left(\hat{O}\right)}{\partial \hat{O}(\mathbf{k})} &= \frac{\partial \mathcal{L}\left(\hat{O}\right)}{\partial \hat{O}_{\Re}(\mathbf{k})}+i\frac{\partial \mathcal{L}\left(\hat{O}\right)}{\partial \hat{O}_{\Im}(\mathbf{k})}\\ &= \sum_z\int -2\Re\left\{\hat{H}_z(\mathbf{k})\left[\hat{O}_{\Re}(\mathbf{k})-\hat{O}_{\Im}(\mathbf{k})\right]\left[\hat{I}_z(\mathbf{k})-\hat{H}_z(\mathbf{k})\hat{O}(\mathbf{k})\right]^*\right\}\mathrm{d}\mathbf{k}\\ &= \sum_z\int -2\Re\left\{\hat{O}(\mathbf{k})^*\left[\hat{H}_z(\mathbf{k})\hat{I}_z(\mathbf{k})^*-|\hat{H}_z(\mathbf{k})|^2\hat{O}(\mathbf{k})^*\right]\right\}\mathrm{d}\mathbf{k},\end{aligned} \tag{2.12}$$

where

$$\hat{O}_{\Re}(\mathbf{k})-\hat{O}_{\Im}(\mathbf{k})=\hat{O}(\mathbf{k})^*. \tag{2.13}$$

In order to make this integral equal to zero, we must have

$$\sum_z\left[\hat{H}_z(\mathbf{k})\hat{I}_z(\mathbf{k})^*-|\hat{H}_z(\mathbf{k})|^2\hat{O}(\mathbf{k})^*\right]=0. \tag{2.14}$$

Finally, we derive the expression for the object spectrum as follows

$$\hat{O}(\mathbf{k})=\frac{\sum_z\hat{I}_z(\mathbf{k})\hat{H}_z(\mathbf{k};P)^*}{\sum_{z'}|\hat{H}_{z'}(\mathbf{k};P)|^2}, \tag{2.15}$$

where z and z' are summation indices in the numerator and denominator, respectively. Eq. (2.15) indicates that provided the pupil function $P(\mathbf{k})$ is known, we can reconstruct the object spectrum $\hat{O}(\mathbf{k})$ using the image spectrum $\hat{I}_z(\mathbf{k})$ and the OTF $\hat{H}_z[\mathbf{k};P(\mathbf{k})]$ in all measurement planes. In this reconstruction, $\hat{I}_z(\mathbf{k})$ is obtained by taking the Fourier transform of the measured intensity distribution and $\hat{H}_z[\mathbf{k};P(\mathbf{k})]$ is computed based on the known pupil function $P(\mathbf{k})$ and the coordinates ($\mathbf{r}$ and z) of the planes in which the images are measured.

2.2.2. The optimization scheme for the pupil function

Now we eliminate the unknown object spectrum $\hat{O}(\mathbf{k})$ from the original error function to derive an error function in which the pupil function $P(\mathbf{k})$ is the only remaining

unknown. For this purpose, we substitute the expression of the object spectrum Eq. (2.15) into the original error function Eq. (2.7) and we obtain

$$\mathcal{L}(P) = \int \sum_z \left| \hat{I}_z(\mathbf{k}) - \hat{H}_z(\mathbf{k};P) \frac{\sum_{z''} \hat{I}_{z''}(\mathbf{k}) \hat{H}_{z''}(\mathbf{k};P)^*}{\sum_{z'} \left|\hat{H}_{z'}(\mathbf{k};P)\right|^2} \right|^2 \mathrm{d}\mathbf{k}, \tag{2.16}$$

where z, z' and z'' summation indices. Note that $\mathcal{L}$ is now a functional of P whereas previously it was a functional of both O and P. This should not cause confusion however. By computing the squared modulus we get (see Appendix I)

$$\begin{aligned} \mathcal{L}(P) &= \int \left\{ \sum_z \left|\hat{I}_z(\mathbf{k})\right|^2 - \frac{\left|\sum_{z''} \hat{I}_{z''}(\mathbf{k}) \hat{H}_{z''}(\mathbf{k})^*\right|^2}{\sum_{z'} \left|\hat{H}_{z'}(\mathbf{k})\right|^2} \right\} \mathrm{d}\mathbf{k} \\ &= \int \left\{ A(\mathbf{k}) - \frac{|C(\mathbf{k})|^2}{B(\mathbf{k})} \right\} \mathrm{d}\mathbf{k}, \end{aligned} \tag{2.17}$$

where

$$A(\mathbf{k}) = \sum_z \left|\hat{I}_z(\mathbf{k})\right|^2, \; B(\mathbf{k}) = \sum_z \left|\hat{H}_z(\mathbf{k})\right|^2, \text{ and } C(\mathbf{k}) = \sum_z \hat{I}_z(\mathbf{k}) \hat{H}_z(\mathbf{k})^*.$$

Eq. (2.17) shows that computation of the error function requires the computation of only 3 quantities, namely A, B, and C. In this computation we do not need to know the object spectrum $\hat{O}(\mathbf{k})$, so Eq. (2.17) only depends on the pupil function $P(\mathbf{k})$. The lateral and axial coordinates ($\mathbf{r}$ and z) of the measurement planes are referred to as hyper-parameters, which will also affect the value of the error function and hence need to be known a priori. Namely, the sampling of $\mathbf{k}$ depends on the sampling of $\mathbf{r}$ and the method for computing the Fourier transform. $\hat{H}_z(\mathbf{k})$ should be computed for the location z of the plane in which the image whose Fourier transform is $\hat{I}_z(\mathbf{k})$ is measured.

It is customary to express the pupil function as

$$P(\mathbf{k}) = |P(\mathbf{k})| \exp[i 2\pi \Phi(\mathbf{k})], \tag{2.18}$$

where $|P(\mathbf{k})|$ represents the amplitude and $\Phi(\mathbf{k})$ is the phase. The amplitude is often assumed to be everywhere 1 inside the pupil, i.e. $P(\mathbf{k}) = 1$ if $|\mathbf{k}| \leq 1$, and 0 elsewhere. The fact that the pupil radius being $\mathbf{k} = 1$ is due to the normalization of the image plane coordinate $\mathbf{r}$ by λ/α_i. The phase represents the wavefront error at the pupil, which is described by a weighted sum of Zernike polynomials:

$$\Phi(\mathbf{k}) = \sum_{m,n} \zeta_n^m Z_n^m(\mathbf{k}), \tag{2.19}$$

where $Z_n^m(\mathbf{k})$ and ζ_n^m are the Zernike polynomials and the associated weights with radial order n and azimuthal order m, respectively. In practice, we use the weights

of a set of 15 ($n = 1, \cdots, 4$) or 37 ($n = 1, \cdots, 6$) Zernike polynomials as unknowns. Therefore, we can significantly reduce the number of unknowns, which previously was the value of the pupil function at every sampling point. Now the unknowns are the weights of Zernike polynomials. We remark that every Zernike polynomial corresponds to a specific type of aberration. Representing the wavefront error by a limited number of Zernike polynomials already made use of a priori information about the aberrations.

Now the error function depends on a vector $\boldsymbol{\zeta}$ of aberration coefficients instead of the entire pupil function $P(\mathbf{k})$ as function of $\mathbf{k}$. We can observe that the dependence of $\mathcal{L}(\boldsymbol{\zeta})$ on $\boldsymbol{\zeta}$ is very non-linear. So, $\boldsymbol{\zeta}$ can only be determined via optimization. Typical optimization algorithms start with an initial guess of $\boldsymbol{\zeta}$ and update $\boldsymbol{\zeta}$ iteratively until a certain stopping criterion is met, e.g. the error function $\mathcal{L}(\boldsymbol{\zeta})$ or its variation is sufficiently small. The scheme for updating $\boldsymbol{\zeta}$ can be written as

$$\boldsymbol{\zeta}_{k+1} = \boldsymbol{\zeta}_k + \Delta\boldsymbol{\zeta}_k, \tag{2.20}$$

where the subscript k denotes the index of iteration, and $\Delta\boldsymbol{\zeta}_k$ is the update of $\boldsymbol{\zeta}_k$. In order to guarantee that the error function at $\boldsymbol{\zeta}_{k+1}$ is not larger than at $\boldsymbol{\zeta}_k$, the convergence constraint must be fulfilled:

$$\mathcal{L}(\boldsymbol{\zeta}_{k+1}) \leq \mathcal{L}(\boldsymbol{\zeta}_k). \tag{2.21}$$

Eq. (2.21) guarantees that $\boldsymbol{\zeta}$ converges to a local minimum, which is in the neighborhood of the initial guess of $\boldsymbol{\zeta}$. The essence of an optimization algorithm is to determine the update $\Delta\boldsymbol{\zeta}_k$ of $\boldsymbol{\zeta}_k$ in each iteration k.

In most optimization algorithms, e.g. quasi-Newton methods, the determination of the update is based on the gradient vector of the error function with respect to the unknown $\boldsymbol{\zeta}$. Each element of the gradient vector is a derivative with respect to an aberration coefficient ζ_ℓ, where ℓ is the Noll's index of the corresponding aberration (Zernike polynomial) [22]. This derivative is given by

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\zeta})}{\partial \zeta_\ell} &= \frac{\partial}{\partial \zeta_\ell} \int \left\{ A(\mathbf{k}) - \frac{|C(\mathbf{k})|^2}{B(\mathbf{k})} \right\} \mathrm{d}\mathbf{k} \\ &= -\int \frac{1}{B(\mathbf{k})^2} \left\{ 2B(\mathbf{k}) \Re\left\{ C(\mathbf{k})^* \frac{\partial C(\mathbf{k})}{\partial \zeta_\ell} \right\} - |C(\mathbf{k})|^2 \frac{\partial B(\mathbf{k})}{\partial \zeta_\ell} \right\} \mathrm{d}\mathbf{k},\end{aligned} \tag{2.22}$$

where

$$\frac{\partial B(\mathbf{k})}{\partial \zeta_\ell} = \sum_z 2\Re\left\{ \hat{H}_z(\mathbf{k})^* \frac{\partial \hat{H}_z(\mathbf{k})}{\partial \zeta_\ell} \right\}, \tag{2.23}$$

and

$$\frac{\partial C(\mathbf{k})}{\partial \zeta_\ell} = \sum_z 2\Re\left\{ \hat{I}_z(\mathbf{k})^* \frac{\partial \hat{H}_z(\mathbf{k})}{\partial \zeta_\ell} \right\}. \tag{2.24}$$

In the plane located at z, the OTF $\hat{H}_z(\mathbf{k})$ is given by the Fourier transform of the intensity $H_z(\mathbf{r})$ of the PSF, which depends on the pupil function $P(\mathbf{k})$, whose phase

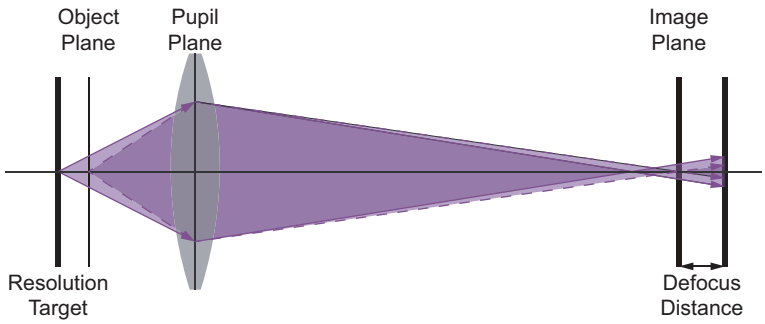

Figure 2.2: Schematic plot of the experimental setup. We introduce defocus aberration to the microscope objective (object side NA $\alpha_o = 0.12$, magnification $4\times$, and operating wavelength at 625 nm) by moving the resolution target away from the nominal object plane. The solid and the dashed line show the imaging with and without defocus aberration. We measure images of the resolution target in the focal plane at $z = 0$ μm (nominal best image plane) and a defocused plane at distance $z = 500$ μm (about 0.72π in normalized axial coordinate).

is a function of the aberration coefficient ζ_ℓ. Using the chain rule, we can derive that

$$\frac{\partial \hat{H}_z(\mathbf{k};\boldsymbol{\zeta})}{\partial \zeta_\ell} = \frac{\partial}{\partial \zeta_\ell}\mathcal{F}\left\{H_z(\mathbf{r};\boldsymbol{\zeta})\right\} = \mathcal{F}\left\{\frac{\partial H_z(\mathbf{r};\boldsymbol{\zeta})}{\partial \zeta_\ell}\right\}, \tag{2.25}$$

and

$$\begin{aligned}\frac{\partial H_z(\mathbf{r};\boldsymbol{\zeta})}{\partial \zeta_\ell} &= \frac{\partial}{\partial \zeta_\ell}\left|\mathcal{F}\left\{P(\mathbf{k})\exp(-iz'|\mathbf{k}|^2)\right\}\right|^2 \\ &= 2\Re\left\{\mathcal{F}\left\{P(\mathbf{k})\exp(-iz'|\mathbf{k}|^2)\right\}^* \mathcal{F}\left\{\frac{\partial P(\mathbf{k})}{\partial \zeta_\ell}\exp(-iz'|\mathbf{k}|^2)\right\}\right\},\end{aligned} \tag{2.26}$$

where

$$\frac{\partial P(\mathbf{k})}{\partial \zeta_\ell} = \frac{\partial}{\partial \zeta_\ell}\exp\left[i2\pi\sum\nolimits_\ell \zeta_\ell Z_\ell(\mathbf{k})\right] = i2\pi Z_\ell(\mathbf{k})P(\mathbf{k}). \tag{2.27}$$

In each iteration of the optimization, computing the gradient vector using Eq. (2.22) - (2.27) is the most time-consuming step. The computational load is proportional to the length of the vector $\boldsymbol{\zeta}$. A possible way for accelerating the optimization is to parallelize the computation of the elements of the gradient vector.

2.2.3. Experimental setup

We validate the phase diversity method by a proof-of-principle experiment. The experimental setup is shown in Fig. 2.2. We use a $4\times$ magnification microscope objective operating at wavelength $\lambda = 625$ nm with object side NA $\alpha_o = 0.12$ and image side NA $\alpha_i = 0.03$. As a result, this microscope objective can achieve resolution $\lambda/(2\alpha_o) = 2.61$ μm and $\lambda/(2\alpha_i) = 10.42$ μm in the object and image side, respectively. Kohler illumination is implemented in the experimental setup. The illumination source is a narrow band collimated LED light source (Thorlabs

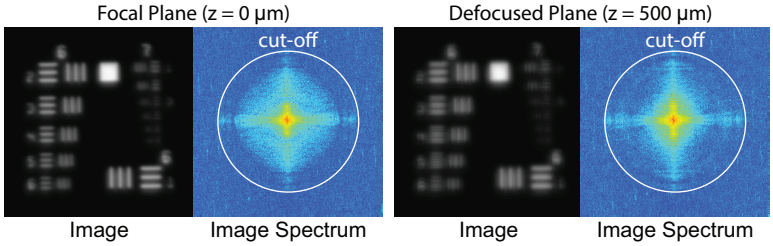

Figure 2.3: Resolution target and measurements of two through-focus images in spatial domain and spatial frequency domain. Group 6 and gourp 7 of the resolution target are imaged by a CCD camera located in the nominal image plane at $z = 0$ and a defocused plane at distance $z = 500$ μm.

M625L4-C2 at 625 nm 490 mW). The size of the illumination source is sufficiently large to guarantee incoherent imaging of the object on the camera detector.

The object is a resolution test target (Thorlabs 1951 USAF Resolution Test Target $\Phi 1''$) with transmissive pattern (bright) on a reflective background (dark), which is made by plating chrome on a soda lime glass substrate. The resolution test target has 6 groups (from -2 to $+7$) of patterns: each consisting of 6 elements with 3 vertical bars and 3 horizontal bars. The smallest bar in group 7 element 6 is about 4.4 μm (equivalent to 288 pairs of lines per millimeter). Because the microscope objective is corrected for spherical aberration induced by the glass coverslip, we introduce spherical aberration by removing the glass coverslip. We furthermore introduce defocus aberration to the microscope objective by moving the object away from the nominal object plane.

A 16-bit CCD camera (SVS-VISTEK eco204MVGE) with pixel size 4.65×4.65 μm and pixel number 1024×776 is placed at a distance $L = 160$ mm, which equals the standard tube length, away from the microscope objective to measure the image. The CCD camera is mounted on a linear precision translation stage (Physik Instrumente M-126.GC1) with step size 100 nm and range 25 mm. The accuracy in the entire range is 2.5 μm. We measure the first image in a defocused plane and the second image in an additional defocused plane at distance $z = 500$ μm (about 0.72π normalized defocus distance) away from the first defocused plane. The direction of defocus is chosen such that the second image is more blurred than the first image.

In the experiment, we introduce 5 settings of defocus aberration by placing the object at 5 defocus positions on the optical axis. For each setting of defocus aberration, we take 2 measurements of images. In total we take 10 images for this proof-of-principle experiment.

2.3. Image filtering using a window function

Image filtering is an inevitable operation of any optimization algorithm that relies on the Fourier transforms of images. In principle, the function of an imaging system is to perform a mapping from the object plane to the image plane, whose sizes are related by the magnification/demagnification of the imaging system. Usually,

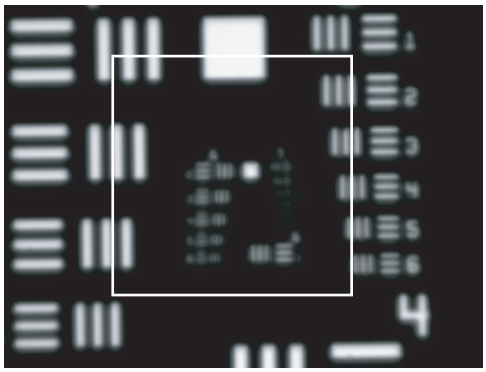

Figure 2.4: An image as measured by the camera sensor. We crop a region (marked by the white box) of the original image to demonstrate the filtering effect,

this image is much larger than the camera sensor, and hence in the measurement process only a small part of this image can be captured. Therefore, measuring an image is equivalent to crop the original image using a window function as shown in Fig. 2.4. We can describe the measurement process by

$$I(\mathbf{r}) = I'(\mathbf{r})W(\mathbf{r}), \tag{2.28}$$

where $I(\mathbf{r})$ and $I'(\mathbf{r})$ are the measured and the original images respectively, and $W(\mathbf{r})$ is the window function. Fourier transforming both sides of Eq. (2.28) yields

$$\hat{I}(\mathbf{k}) = \hat{I}'(\mathbf{k}) * \hat{W}(\mathbf{k}), \tag{2.29}$$

where $*$ denotes the convolution operator. Eq. (2.29) indicates that the spectrum of the measured image $\hat{I}(\mathbf{k})$ is the spectrum of the original image $\hat{I}'(\mathbf{k})$ convoluted by the spectrum of the window function $\hat{W}(\mathbf{k})$. As a consequence, the measurement process can be regarded as filtering the original image by the window function in the spatial domain (ordinary domain).

Note that the size of the Fourier transform of the window function, which has a finite size, is infinite. Therefore, although the spectrum of the original image $\hat{I}'(\mathbf{k})$ is limited by the cut-off spatial frequency of the OTF $\hat{H}(\mathbf{k})$, the spectrum of the measured image $\hat{I}(\mathbf{k})$ is not limited due to the filtering by $\hat{W}(\mathbf{k})$. This effect is referred to as the spectrum leakage. In order to make

$$\hat{I}(\mathbf{k}) = \left[\hat{H}(\mathbf{k})\hat{O}(\mathbf{k})\right] * \hat{W}(\mathbf{k}) \approx \hat{H}(\mathbf{k})\hat{O}(\mathbf{k}) \tag{2.30}$$

so that Eq. (2.5) holds, we need to use a window function whose spectrum is sufficiently narrow. The spectrum of the optimal window function should be a Dirac delta function, which, however, cannot be achieved. The default window function corresponding to using a rectangular camera sensor is rectangular, but we can always apply an additional window function to the measured image in the post-processing step so that the spectrum of the combined window function becomes sufficiently narrow, i.e. having a high ratio between the main lobe and side lobes.

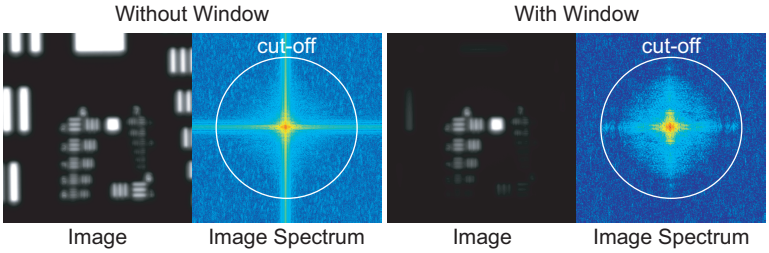

Figure 2.5: Comparison between the focal plane image and its spectrum without (left) and with (right) Chebyshev window. The sizes of the original and the cropped images are 1024×768 and 512×512 respectively.

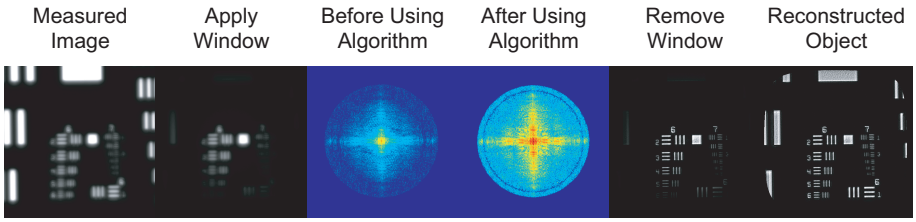

Figure 2.6: Flowchart of applying and removing the window function for the phase diversity method. In order to use the phase diversity algorithm, we need to apply the window function to the measured image and remove it from the reconstructed object (restored image). The removal is regularized by adding a small constant to the window function to avoid noise amplification.

In this work, we use the Chebyshev window as the extra window function. In the spatial frequency domain, it is optimal in that it minimizes the width of the main-lobe for a given attenuation of the side-lobes.

Fig. 2.5 shows the measured image (in the focal plane) and its spectrum with and without Chebyshev window. We can observe that the Chebyshev window can significantly reduce the leakage of the Fourier transform of the image caused by the cropping of the image due to the camera sensor. However, because the Chebyshev window decreases from the center to the edge of the FOV, the information of the object will be lost at places where the value of Chebyshev window is very small (near the edge of the FOV). In the phase diversity method, we first apply the Chebyshev window to the measured image and then remove it from the reconstructed object. The flowchart of this process is shown in Fig. 2.6.

2.4. The Optimization algorithm

The optimization algorithm is the core of the phase diversity method. Here we consider two groups of optimization algorithms: Newton type methods which use the Hessian matrix, and quasi-Newton type methods use a proper update scheme to avoid the use of the Hessian matrix. Typically computing the Hessian matrix is very time consuming, but it provides a quadratic approximation of the landscape of the error function $\mathcal{L}(\boldsymbol{\zeta})$ in the vicinity of the current guess of $\boldsymbol{\zeta}$ and hence it allows

the derivation of the update $\Delta\boldsymbol{\zeta}$ directly (the updated guess $\boldsymbol{\zeta}+\Delta\boldsymbol{\zeta}$ is at the minimum of the quadratic approximation).

Suppose that $\boldsymbol{\zeta}_0$ and $\boldsymbol{\zeta}$ are vectors before and after the update, respectively. Both are $1\times N$ vectors consisting of N aberration coefficients. For $\boldsymbol{\zeta}$ in the vicinity of $\boldsymbol{\zeta}_0$, we approximate the error function at $\boldsymbol{\zeta}_0$ by expanding it using a Taylor series. The error function at $\boldsymbol{\zeta}$ is then approximately given by

$$\mathcal{L}_{app}(\boldsymbol{\zeta}) \approx \mathcal{L}(\boldsymbol{\zeta}_0) + (\boldsymbol{\zeta}-\boldsymbol{\zeta}_0)^{\mathrm{T}}\mathcal{L}'(\boldsymbol{\zeta}_0) + \frac{1}{2}(\boldsymbol{\zeta}-\boldsymbol{\zeta}_0)^{\mathrm{T}}\mathcal{L}''(\boldsymbol{\zeta}_0)(\boldsymbol{\zeta}-\boldsymbol{\zeta}_0), \tag{2.31}$$

where $\mathcal{L}'(\boldsymbol{\zeta}_0) = [\partial\mathcal{L}(\boldsymbol{\zeta}_0)/\partial\zeta_\ell]$ is the $1\times N$ gradient vector (first derivative) and $\mathcal{L}''(\boldsymbol{\zeta}_0) = \left[\partial^2\mathcal{L}(\boldsymbol{\zeta}_0)/\partial\zeta_\ell\partial\zeta_{\ell'}\right]$ is the $N\times N$ Hessian matrix (second derivative).

Because the approximated error function Eq. (2.31) is quadratic, it has a unique extreme (stationary point), which satisfies

$$\mathcal{L}'_{app}(\boldsymbol{\zeta}) \approx \mathcal{L}'(\boldsymbol{\zeta}_0) + \mathcal{L}''(\boldsymbol{\zeta}_0)(\boldsymbol{\zeta}-\boldsymbol{\zeta}_0) = 0. \tag{2.32}$$

Therefore, the optimal update vector can be determined by solving Eq. (2.32):

$$\Delta\boldsymbol{\zeta}_0 = \boldsymbol{\zeta}-\boldsymbol{\zeta}_0 = -\left[\mathcal{L}''(\boldsymbol{\zeta}_0)\right]^{-1}\mathcal{L}'(\boldsymbol{\zeta}_0). \tag{2.33}$$

However, the location of the stationary point of the quadratic error function $\boldsymbol{\zeta}$ may not necessarily be in the vicinity of $\boldsymbol{\zeta}_0$ where the approximation by the first three terms of the Taylor series expansion is accurate. Therefore, the update vector is often defined by

$$\Delta\boldsymbol{\zeta}_0 = l\mathbf{d}, \tag{2.34}$$

where the vector $\mathbf{d}$ and the number l are the direction (a unit vector) and the length of (a scalar) the update vector, respectively.

In the Newton and quasi-Newton type method the update direction $\mathbf{d}$ is chosen to be in the direction pointing to the stationary point of the quadratic error function:

$$\mathbf{d} = -\frac{\left[\mathcal{L}''(\boldsymbol{\zeta}_0)\right]^{-1}\mathcal{L}'(\boldsymbol{\zeta}_0)}{\left\|\left[\mathcal{L}''(\boldsymbol{\zeta}_0)\right]^{-1}\mathcal{L}'(\boldsymbol{\zeta}_0)\right\|}, \tag{2.35}$$

Depending on the initial guess of $\boldsymbol{\zeta}$, both Newton and quasi-Newton type methods can lead to a stationary point, which can be

- a local minimum, where for all directions in $\boldsymbol{\zeta}$ space there is a minimum.
- a local maximum, where for all directions in $\boldsymbol{\zeta}$ space there is a maximum.
- a saddle point, where for some directions in $\boldsymbol{\zeta}$ space there is a minimum whereas for other directions there is a maximum..

In Eq. (2.35), the time to compute the gradient vector $\mathcal{L}'(\boldsymbol{\zeta}_0)$ and the Hessian matrix $\mathcal{L}''(\boldsymbol{\zeta}_0)$ is proportional to N and N^2, respectively. Therefore, computing $\mathcal{L}''(\boldsymbol{\zeta}_0)$ directly either using an analytical formula or via numerical estimation (e.g. finite difference) is time-consuming. Alternatively, direct computation of $\mathcal{L}''(\boldsymbol{\zeta}_0)$ can

be avoided by using a proper update scheme. The most effective update scheme was developed by Broyden [23], Fletcher [24], Goldfarb [25], and Shanno [26] (BFGS), which requires computation of only the gradient vector $\mathcal{L}'(\boldsymbol{\zeta}_0)$. In order to guarantee that $\boldsymbol{\zeta}$ converges to a local minimum, the update scheme of BFGS ensures that $\mathcal{L}''(\boldsymbol{\zeta}_0)$ is positive-definite in every iteration.

The variation of the error function at $\boldsymbol{\zeta}_0$ in the limit of the update length $l \to 0$ is given by

$$\lim_{l\to 0} \frac{\mathcal{L}(\boldsymbol{\zeta}_0 + l\mathbf{d}) - \mathcal{L}(\boldsymbol{\zeta}_0)}{l} \approx \lim_{l\to 0} \frac{2l\mathbf{d}^{\mathrm{T}}\mathcal{L}(\boldsymbol{\zeta}_0)' - l^2\mathbf{d}^{\mathrm{T}}\mathcal{L}(\boldsymbol{\zeta}_0)''\mathbf{d}}{2l} = \mathbf{d}^{\mathrm{T}}\mathcal{L}'(\boldsymbol{\zeta}_0). \tag{2.36}$$

Eq. (2.36) indicates that the steepest descent direction of the error function is

$$\mathbf{d} = -\frac{\mathcal{L}'(\boldsymbol{\zeta}_0)}{\|\mathcal{L}'(\boldsymbol{\zeta}_0)\|}. \tag{2.37}$$

The optimization algorithm with Eq. (2.37) as the update direction is referred to as the steepest descent method or the gradient descent method. We remark that the BFGS direction Eq. (2.35) is not the steepest descent direction Eq. (2.37). As $\boldsymbol{\zeta}$ approaches closer a local minimum, the BFGS direction becomes more accurate than the steepest descent direction, particularly when the local minimum is located in a long and narrow valley.

The update length l can be found by performing a line search along the update direction $\mathbf{d}$ subject to the Wolfe conditions [27]:

$$\mathcal{L}(\boldsymbol{\zeta}_0 + l\mathbf{d}) \le \mathcal{L}(\boldsymbol{\zeta}_0) + c_1 l\mathbf{d}^{\mathrm{T}}\mathcal{L}'(\boldsymbol{\zeta}_0), \tag{2.38}$$

$$-\mathbf{d}^{\mathrm{T}}\mathcal{L}'(\boldsymbol{\zeta}_0 + l\mathbf{d}) \le -c_2\mathbf{d}^{\mathrm{T}}\mathcal{L}'(\boldsymbol{\zeta}_0), \tag{2.39}$$

where c_1 and c_2 are constants satisfying $0 < c_1 < c_2 < 1$. Typical values (e.g. the values used by Matlab) are $c_1 = 10^{-4}$ and $c_2 = 0.9$. The first and the second condition guarantee sufficient descent of the error function $\mathcal{L}(\boldsymbol{\zeta})$ and its variation $\mathbf{d}^{\mathrm{T}}\mathcal{L}'(\boldsymbol{\zeta})$, respectively. Therefore, $\boldsymbol{\zeta}$ always converges to a minimum of $\mathcal{L}(\boldsymbol{\zeta})$, where $\mathbf{d}^{\mathrm{T}}\mathcal{L}'(\boldsymbol{\zeta}) = 0$.

In Fig. (2.7), we compare results obtained using the BFGS and the steepest descent method for 15 and 37 aberration coefficients, respectively. We observe in Fig. (2.7)(a) and (b) that compared to the steepest descent method, the BFGS method not only converges much faster but also finds a local minimum where the values of both the error function $\mathcal{L}$ and its variation $\mathbf{d}^{\mathrm{T}}\mathcal{L}'$ are much lower. Fig. (2.7) also shows that the set of $\boldsymbol{\zeta}$ found by the two methods are the same when $N = 15$ but are different when $N = 37$. By observing Fig. (2.7)(e) and (f), we can see that the performance of the steepest descent method is better, while the BFGS method performs well when N is small but poorly when N is large. This may be because that the Hessian matrix for the current guess of $\boldsymbol{\zeta}$, which is not necessarily positive definite, is now forced to be positive definite by BFGS method. Besides, the accuracy of computing the inverse of the Hessian matrix decreases as the length of $\boldsymbol{\zeta}$ increases.

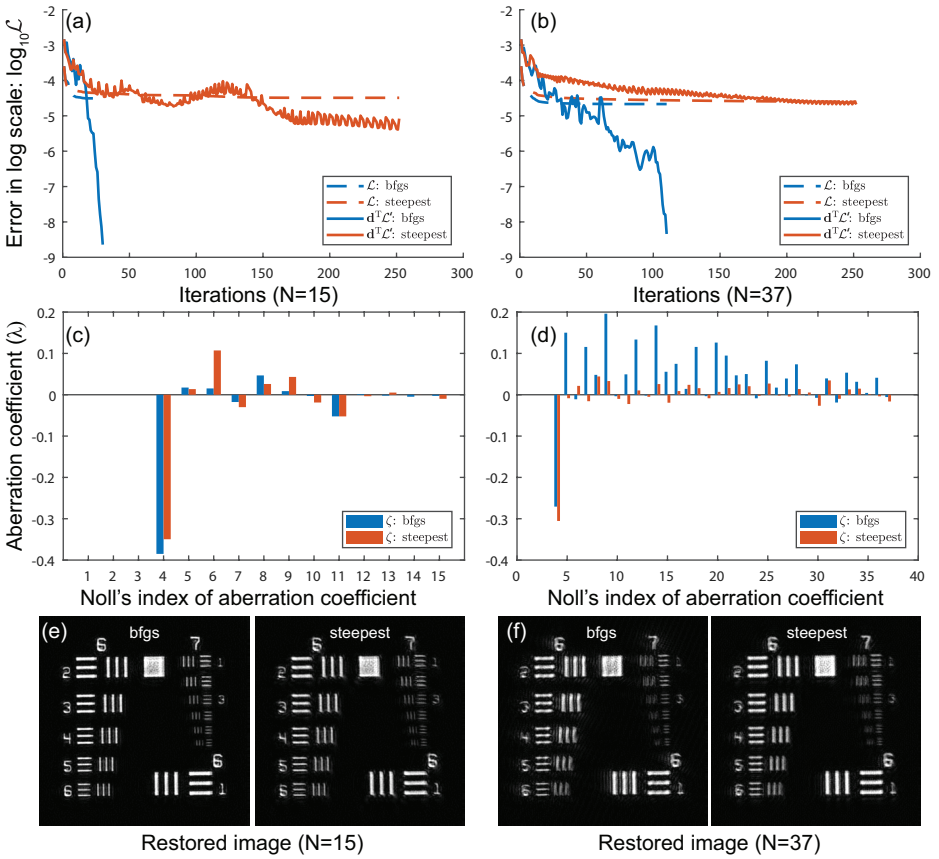

Figure 2.7: Comparison between the optimization for the aberration coefficient vector $\boldsymbol{\zeta}$ consisting of 15 and 37 terms using the BFGS method and the steepest descent method. This comparison is done for the images shown in Fig. 2.3. In (a) and (b) values of the error function $\mathcal{L}(\boldsymbol{\zeta})$ and its variance $\mathbf{d}^T\mathcal{L}(\boldsymbol{\zeta})$ are shown. (c) and (d) show the retrieved aberration coefficient vector $\boldsymbol{\zeta}$ and (e) and (f) show the object reconstructed using the retrieved aberration coefficient vector $\boldsymbol{\zeta}$.

2.5. Regularization of the phase diversity method

In the phase diversity method, the object spectrum $\hat{O}(\mathbf{k})$ is reconstructed by solving an inverse problem given the image spectrum $\hat{I}_z(\mathbf{k})$ and the estimated OTF $\hat{H}_z(\mathbf{k})$ in a every through-focus measurement plane. However, it is well-known that the reconstruction of $\hat{O}(\mathbf{k})$ is ill-posed because dividing $\hat{I}_z(\mathbf{r})$ by $\hat{H}_z(\mathbf{k})$ is extremely unstable. Recall that the formula of image formation in the spatial frequency domain is given by

$$\hat{I}_z(\mathbf{k}) = \hat{O}(\mathbf{k})\hat{H}_z(\mathbf{k}) + \hat{N}_z(\mathbf{k}), \tag{2.40}$$

where $\hat{N}_z(\mathbf{k})$ is the Fourier transform (the spectrum) of the noise $N_z(\mathbf{r})$. A naive solution to this inverse problem is

$$\hat{O}_{\text{naive}}(\mathbf{k}) = \hat{I}_z(\mathbf{k})/\hat{H}_z(\mathbf{k}) = \hat{O}(\mathbf{k}) + \hat{N}_z(\mathbf{k})/\hat{H}_z(\mathbf{k}). \tag{2.41}$$

Eq. (2.41) shows that dividing the noise spectrum $\hat{N}_z(\mathbf{k})$ by the OTF $\hat{H}_z(\mathbf{k})$ will amplify $\hat{N}_z(\mathbf{k})$ at places where $\hat{H}_z(\mathbf{k})$ is small, in particular where it vanishes. This amplification of $\hat{N}_z(\mathbf{k})$ usually occurs on the edge of a disc with radius given by the cut-off spatial frequency. Because an arbitrary small component of $\hat{N}_z(\mathbf{k})$ can be amplified to an arbitrary large value, $\hat{O}(\mathbf{k})$ is very sensitive to the fluctuations of $\hat{N}_z(\mathbf{k})$.

It is customary to deal with ill-posed inverse problem using Tikhonov regularization when Gaussian noise is the dominant noise. In this case, the error function equals the L-2 norm of the noise for the object spectrum $\hat{O}(\mathbf{k})$ that is the exact solution. The Tikhonov regularization modifies this original error function by adding a term proportional to the L-2 norm of the object spectrum:

$$\mathcal{L}_\gamma\left(\hat{O}\right) = \sum_z \|\hat{I}_z(\mathbf{k}) - \hat{H}_z(\mathbf{k};\boldsymbol{\zeta})\hat{O}(\mathbf{k})\|_2^2 + \gamma\|\hat{O}(\mathbf{k})\|_2^2, \tag{2.42}$$

where $\|\cdots\|_2^2$ denotes the L-2 norm, $\|\hat{O}(\mathbf{k})\|_2^2$ is the regularization term and γ is the regularization parameter. We aim to find $\hat{O}(\mathbf{k})$ that minimize not only the original error function but also the regularization term, which can be regarded as "the total energy" of $\hat{O}(\mathbf{k})$. By minimizing "the total energy", Tikhonov regularization guarantees that the value of $\hat{O}(\mathbf{k})$ is bounded and hence will not be significantly affected by the fluctuations of the noise spectrum $\hat{N}(\mathbf{k})$.

We can derive the regularized expression for the object spectrum by setting the derivative of L_γ with respect to $\hat{O}(\mathbf{k})$ to zero:

$$\hat{O}_\gamma(\mathbf{k}) = \frac{\sum_z \hat{I}_z(\mathbf{k})\hat{H}_z(\mathbf{k})^*}{\sum_{z'} |\hat{H}_{z'}(\mathbf{k})|^2 + \gamma} = \frac{C(\mathbf{k})}{B(\mathbf{k}) + \gamma}. \tag{2.43}$$

Comparing Eq. (2.43) with Eq. (2.15), we see that the regularization parameter γ works as an offset to the denominator. When the regularization parameter γ is too small, the resulting object spectrum shows artefacts, while when γ is too large, the object spectrum does not resemble the actual object spectrum anymore and hence the fitting accuracy is poor. Therefore, the optimal value of γ should balance the effect of regularization and fitting accuracy.

We remark that during the optimization, we do not find the optimal γ in each iteration, which would be very time-consuming, but instead use a small value for gamma that is kept constant during the optimization, to prevent dividing by zero. It is shown by both simulations and experiments that the exact value of this small constant does not influence the final result significantly. However, the value of regularization parameter γ does significantly influence the reconstructed object. Therefore, after the optimization we need to determine the optimal γ for the reconstruction of the object spectrum.

We determine this optimal value for γ using the L-curve method [28]. The L-curve is a plot of the values of the error function versus the regularization term as shown at the left of Fig. 2.8. Both values can be parametrized by the regularization parameter γ, and hence the L-curve is given by the set of points:

$$\left(\xi(\gamma) = \sum_z \left\|\hat{I}_z(\mathbf{k}) - \hat{H}_z(\mathbf{k})\hat{O}_\gamma(\mathbf{k})\right\|_2^2,\ \eta(\gamma) = \left\|\hat{O}_\gamma(\mathbf{k})\right\|_2^2\right). \tag{2.44}$$

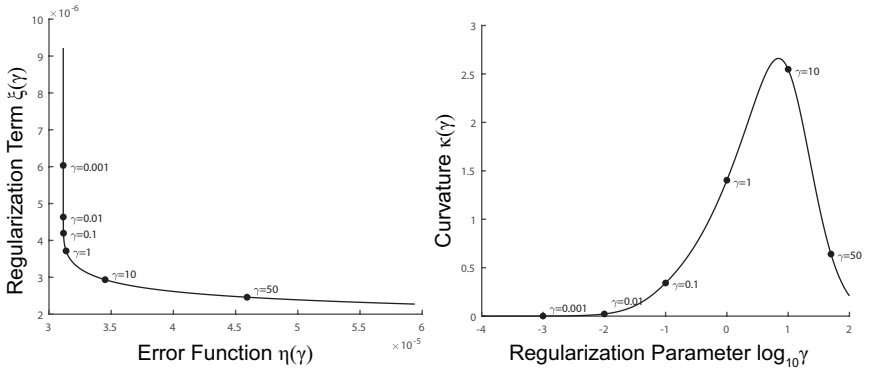

Figure 2.8: Illustration of the L-curve (left) and its curvature (right). The data is based on an optimization using BFGS method for 15 coefficients of aberrations.

Along the L-curve, γ increases monotonically from left to right. The L-curve is separated into two parts by a sharp corner. We can observe that as γ increases, $\eta(\gamma)$ (the reconstruction error) decreases on the vertical part, whereas $\xi(\gamma)$ (the fitting error) increases on the horizontal part. So, the optimum value of γ, which optimally balances the reconstruction error and the fitting error, can be found at the location of the corner of the L-curve.

To determine the location of the corner, we need to calculate the curvature of the L-curve which is given by

$$\kappa(\gamma) = 2\frac{\tilde{\xi}'(\gamma)\tilde{\eta}''(\gamma) - \tilde{\xi}''(\gamma)\tilde{\eta}'(\gamma)}{\sqrt[3]{\tilde{\xi}'(\gamma)^2 + \tilde{\eta}'(\gamma)^2}}, \tag{2.45}$$

where $'$ and $''$ are the first and second derivative, respectively, and

$$\begin{cases} \tilde{\eta}'(\gamma) = \dfrac{\eta'(\gamma)}{\eta(\gamma)} \\ \tilde{\xi}'(\gamma) = \dfrac{\xi'(\gamma)}{\xi(\gamma)} \end{cases}, \tag{2.46}$$

and

$$\begin{cases} \tilde{\eta}''(\gamma) = \dfrac{\eta''(\gamma)\eta(\gamma) - \eta'(\gamma)^2}{\eta(\gamma)^2} \\ \tilde{\xi}''(\gamma) = \dfrac{\xi''(\gamma)\xi(\gamma) - \xi'(\gamma)^2}{\xi(\gamma)^2} \end{cases}, \tag{2.47}$$

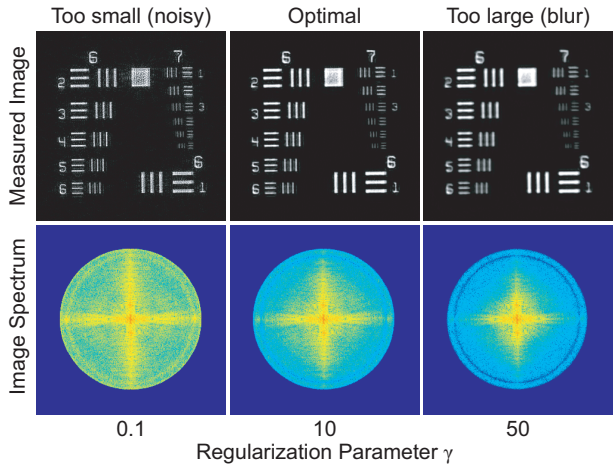

Figure 2.9: Object reconstruction results versus regularization parameter values. Reconstructed object (top) and its Fourier transform (bottom). This plot shows that the reconstructed object will be too noisy when γ is too small and too blurred when γ is too large. The balance is achieved when γ corresponds to the corner of the L-curve where its curvature is maximum (see Fig. 2.8).

where the first and second derivatives of $\eta(\gamma)$ are

$$\begin{aligned}\eta'(\gamma) &= \frac{\mathrm{d}}{\mathrm{d}\gamma}\int \left|\hat{O}_\gamma(\mathbf{k})\right|^2 \mathrm{d}\mathbf{k} = \int \left|\frac{C(\mathbf{k})}{B(\mathbf{k})+\gamma}\right|^2 \mathrm{d}\mathbf{k} \\ &= \int 2\Re\left\{-\left[\frac{C(\mathbf{k})}{B(\mathbf{k})+\gamma}\right]^* \frac{C(\mathbf{k})}{[B(\mathbf{k})+\gamma]^2}\right\}\mathrm{d}\mathbf{k}, \\ \eta''(\gamma) &= \int 2\Re\left\{\left|\frac{C(\mathbf{k})}{[B(\mathbf{k})+\gamma]^2}\right|^2 + \left[\frac{C(\mathbf{k})}{B(\mathbf{k})+\gamma}\right]^* \frac{2C(\mathbf{k})}{[B(\mathbf{k})+\gamma]^3}\right\}\mathrm{d}\mathbf{k},\end{aligned}$$

and the first and second derivatives of $\xi(\gamma)$ are

$$\begin{aligned}\xi'(\gamma) &= \frac{\mathrm{d}}{\mathrm{d}\gamma}\sum_z \int \left|\hat{I}_z(\mathbf{k}) - \hat{H}_z(\mathbf{k})\hat{O}_\gamma(\mathbf{k})\right|^2 \mathrm{d}\mathbf{k} \\ &= \sum_z \int 2\Re\left\{-\left[\hat{I}_z(\mathbf{k}) - \frac{\hat{H}_z(\mathbf{k})C(\mathbf{k})}{B(\mathbf{k})+\gamma}\right]^* \frac{\hat{H}_z(\mathbf{k})C(\mathbf{k})}{[B(\mathbf{k})+\gamma]^2}\right\}\mathrm{d}\mathbf{k}, \\ \xi''(\gamma) &= \sum_z \int 2\Re\left\{\left|\frac{\hat{H}_z(\mathbf{k})C(\mathbf{k})}{[B(\mathbf{k})+\gamma]^2}\right|^2 + \left[\hat{I}_z(\mathbf{k}) - \frac{\hat{H}_z(\mathbf{k})C(\mathbf{k})}{B(\mathbf{k})+\gamma}\right]^* \frac{2\hat{H}_z(\mathbf{k})C(\mathbf{k})}{[B(\mathbf{k})+\gamma]^3}\right\}\mathrm{d}\mathbf{k}.\end{aligned}$$

We plot the curvature $\kappa(\gamma)$ as a function of regularization parameter γ at the right panel of Fig. 2.8. A distinct peak corresponding to the corner of the L-curve can be observed. To find the location of this peak, we formulate an optimization

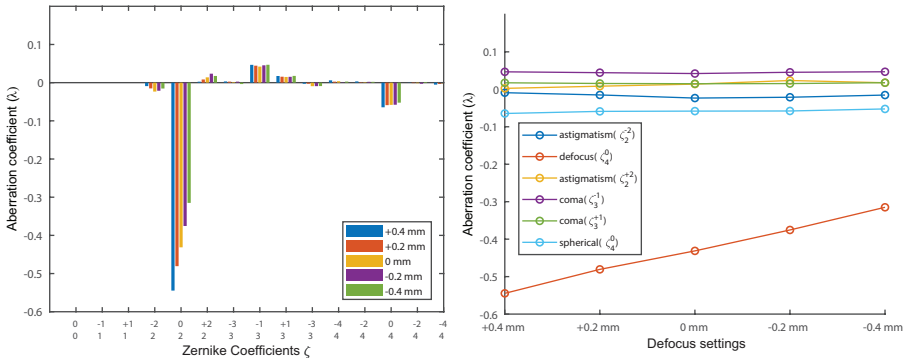

Figure 2.10: Aberration coefficients versus defocus settings. For each defocus setting we retrieve a set of 15 aberration coefficients from the focal image and a defocused image. This image plot shows a linear relation between the defocus aberration and defocus setting. We keep other aberrations constant while varying the defocus aberration.

problem as follows:

$$\gamma = \arg\min_{\gamma}\{-\kappa(\gamma)\}. \tag{2.48}$$

Because γ is the only variable of this optimization problem, we do not need to compute the gradient of $-\kappa(\gamma)$ with respect to γ. This optimization problem can be solved in a few seconds on a desktop, depending on the size and the number of the measured images.

2.6. Experimental Results

In the experiment, we use a $4\times$ magnification microscope objective operating at wavelength $\lambda = 625$ nm with object side NA $\alpha_o = 0.12$ and image side NA $\alpha_i = 0.03$. The object is a resolution test target with the smallest bar in group 7 element 6 equal to about 4.4 µm (equivalent to 288 pairs of lines per millimeter). We introduce a fixed amount of spherical aberration and various amounts of defocus aberration to the measured images to validate the phase diversity method.

Because it is difficult to locate the nominal object plane, we first move the object away from the nominal object plane to introduce a sensible amount of defocus aberration and then move the object towards both directions along the optical axis by 2 steps with 0.2 mm interval. At each object location, we measure two images: one in the nominal image plane at $z = 0$ and another in a defocused plane at distance $z = 500$ µm. We choose the direction of defocusing such that the focal image is clearer than the defocused image. In total 10 images are measured in this experiment for 5 positions of the object.

We plot the retrieval results for the aberration coefficients versus the defocus settings in Fig. 2.10. It is seen that the defocus aberration is much larger than the other aberrations. While we vary the defocus aberration, we keep other aberrations almost constant. Most importantly, we can observe a distinct linear relation between the defocus aberration and defocus setting as expected.

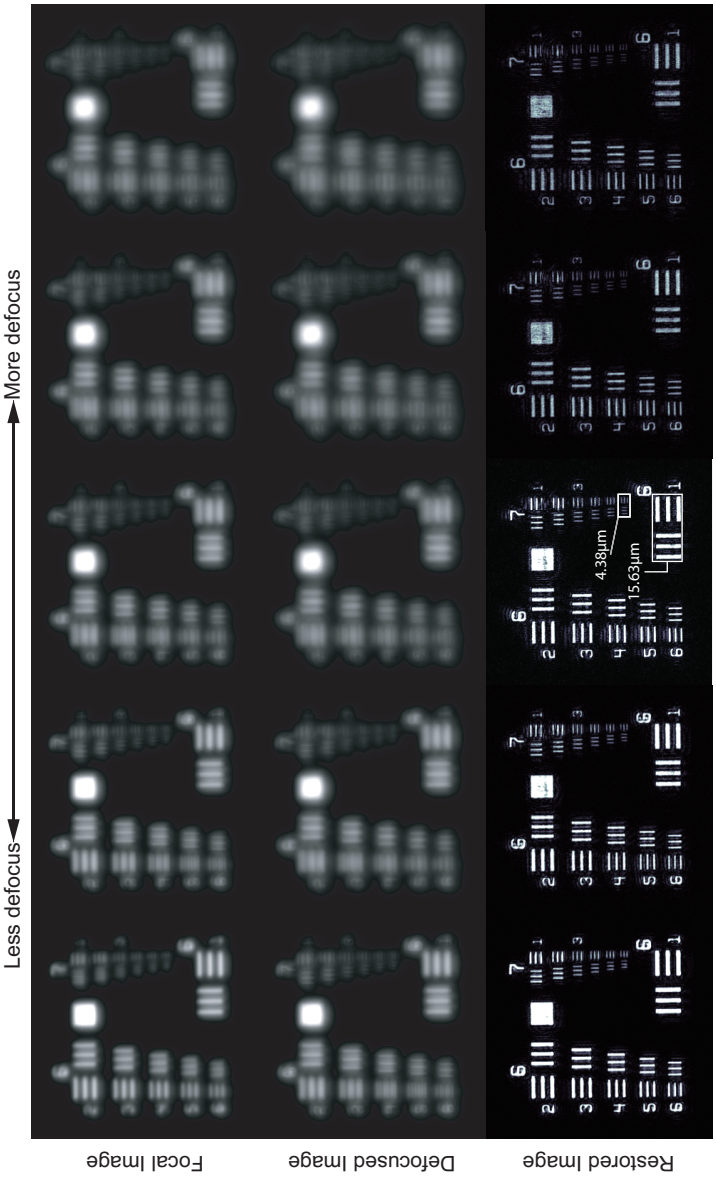

Figure 2.11: Comparison between the object reconstructed based on the retrieved aberration coefficient vector and images measured in the focal and defocused plane for each defocus setting.

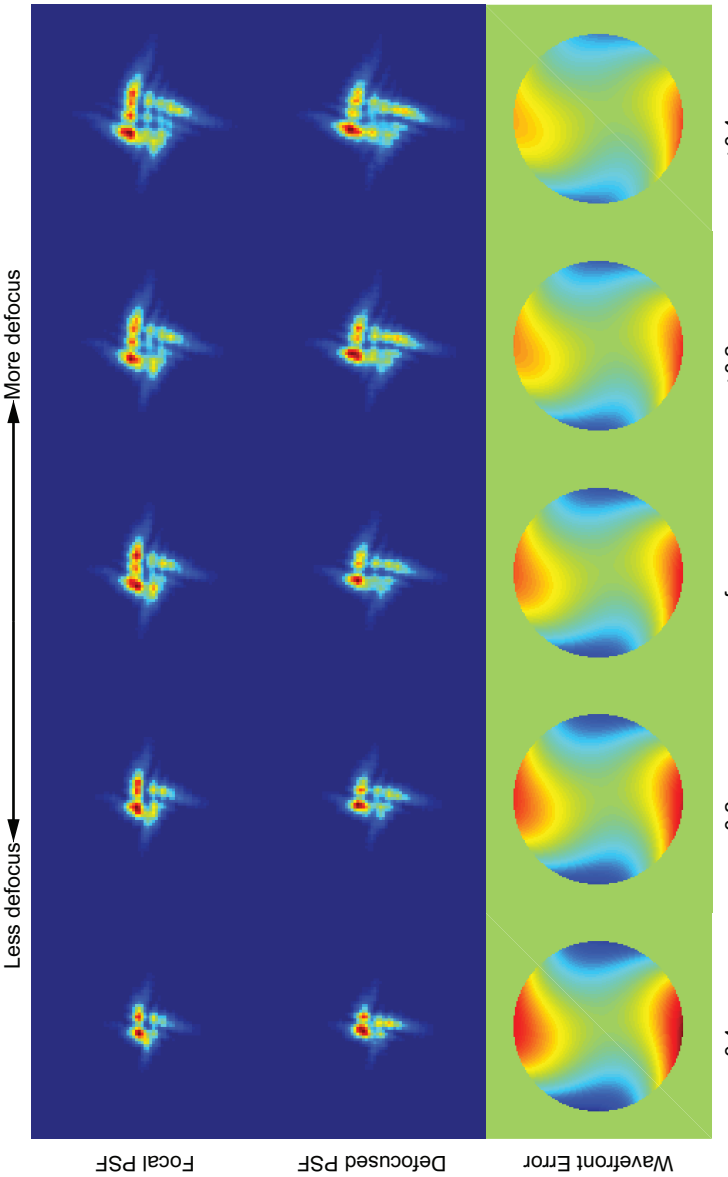

Figure 2.12: Wavefront error (phase of the pupil function) reconstructed based on the retrieved aberration coefficient vector and the corresponding PSF intensities in the focal and defocused plane for each defocus setting.

In Fig. 2.11, images measured in the focal and defocused planes and the restored image (by reconstructing the object) are shown for each of the five object positions. This figure shows that both focal and defocused images become more blurred when the object is further away from the nominal object plane, while the reconstructed object, whose resolution is much higher than that of not only the defocused image but also the focal image, is always equally accurate. The smallest feature of the resolution target, whose size is about 4.38 μm, can clearly be seen in all cases.

It is shown in Fig. 2.12 that the shape of the PSF intensity is consistent with the way how the resolution of the test target changed. The information of the PSF provides essential information for the diagnosis of the imaging system. The work flow of the phase diversity method should be to first retrieve the aberration coefficient vector $\boldsymbol{\zeta}$, which determines the wavefront error, then compute the PSF intensity of each of the through-focus blurred images, and finally obtain a clear image by regularized deconvolution, in which the regularization parameter γ is found by using the L-curve method. In this work flow, information about the wavefront error allows us to compensate the aberrations by hardware, while image restoration allows us to obtain a diffraction-limited image by software.

2.7. Conclusion

In this chapter we have demonstrated that we can retrieve aberrations of imaging systems and perform image restoration using the phase diversity method. We have shown that the retrieved defocus aberration depends linearly on the defocus setting as expected. Furthermore, we showed that the reconstructed object is in all cases less blurred than the measured blurred images. The successful implementation of the phase diversity method relies on many factors. The most important factor is the sampling grid of the measured images. The sampling grid should cover the entire area of the OTF in the spatial frequency domain, so that the complete information about both the object and the aberrations can be acquired. This sampling should be equal to the half size of the diffraction-limited PSF intensity of the imaging system, which is determined by the wavelength and the NA.

In most situations, the information about the images appears in an area much smaller than that of the OTF due to noise and aberrations. Here we need to distinguish the effects due to noise (noisy) and aberrations (blurry): images blurred by aberrations can be restored while those blurred by noise cannot. Usually a more blurred image has lower signal-to-noise ratio. This suggests that we should find a balance between hardware and software aberration correction. Namely, we should perform rough correction by hardware and fine correction by software.

For the phase diversity method and all other image-based methods, the object influences the accuracy of retrieving certain types of aberrations. In the spatial frequency domain, the image spectrum equals the OTF multiplied by the object spectrum. When the object spectrum cannot fill the entire area of the OTF, e.g. when the object is not a point source, the OTF cannot be recovered in areas where the object spectrum is zero. As a result, there will then be higher inaccuracies in the retrieval of certain types of aberrations. It is also challenging to determine the

best object for the retrieval of certain types of aberrations. We remark that the final accuracy of the resolution of the reconstructed object is mainly affected by the noise. In fact, for low noise levels, the aberrations that are important for the accurate reconstruction of the object can always be retrieved accurately.

Finally, the optimization algorithm also plays a significant role here. Depending on the algorithm and the parameters of the algorithm, we may be trapped in one of the local minima, because the dependence on the aberrations of the error function is highly nonlinear and hence there will be more than one minimum. Another issue is the number of aberrations. In this chapter, we have observed crosstalk between aberrations. This implies that the values of the lower order aberration coefficients will depend on the total number of aberrations coefficients.

Appendix I: Derivation of the object free error function

Let us submit the expression for object spectrum Eq. (2.15)

$$\hat{O}(\mathbf{k}) = \frac{\sum_z \hat{I}_z(\mathbf{k})\hat{H}_z[\mathbf{k}; P(\mathbf{k})]^*}{\sum_{z'} |\hat{H}_{z'}[\mathbf{k}; P(\mathbf{k})]|^2} \tag{2.49}$$

into Error function Eq. (2.7)

$$\mathcal{L}\left(\hat{O}, P\right) = \int \sum_z \left|\hat{I}_z(\mathbf{k}) - \hat{H}_z[\mathbf{k}; P(\mathbf{k})]\hat{O}(\mathbf{k})\right|^2 \mathrm{d}\mathbf{k}. \tag{2.50}$$

As a result, we obtain

$$\begin{aligned}\mathcal{L}(P) &= \int \sum_z \left|\hat{I}_z(\mathbf{k}) - \hat{H}_z(\mathbf{k}) \frac{\sum_{z''} \hat{I}_{z''}(\mathbf{k})\hat{H}_{z''}(\mathbf{k})^*}{\sum_{z'} |\hat{H}_{z'}(\mathbf{k})|^2}\right|^2 \mathrm{d}\mathbf{k} \\ &= \int \sum_z \left|\frac{\hat{I}_z(\mathbf{k}) \left(\sum_{z'} \left|\hat{H}_{z'}(\mathbf{k})\right|^2\right) - \hat{H}_z(\mathbf{k}) \left[\sum_{z''} \hat{I}_{z''}(\mathbf{k})\hat{H}_{z''}(\mathbf{k})^*\right]}{\sum_{z'} |\hat{H}_{z'}(\mathbf{k})|^2}\right|^2 \mathrm{d}\mathbf{k}.\end{aligned} \tag{2.51}$$

We now expand the brackets in the numerator by resorting the summation index z, z', and z'':

$$\begin{aligned}&\sum_z \left|\hat{I}_z \left(\sum_{z'} \left|\hat{H}_{z'}\right|^2\right) - \hat{H}_z \left(\sum_{z''} \hat{I}_{z''} \hat{H}_{z''}^*\right)\right|^2 \\ &= \sum_z \left[\left|\hat{I}_z\right|^2 \left(\sum_{z'} \left|\hat{H}_{z'}\right|^2\right)^2 + \left|\hat{H}_z\right|^2 \left|\sum_{z''} \hat{I}_{z''} \hat{H}_{z''}^*\right|^2 - 2 \left(\hat{I}_z \hat{H}_z^*\right) \left(\sum_{z''} \hat{I}_{z''}^* \hat{H}_{z''}\right) \left(\sum_{z'} \left|\hat{H}_{z'}\right|^2\right)\right] \\ &= \left(\sum_z \left|\hat{I}_z\right|^2\right) \left(\sum_{z'} \left|\hat{H}_{z'}\right|^2\right)^2 + \left(\sum_z \left|\hat{H}_z\right|^2\right) \left|\sum_{z''} \hat{I}_{z''} \hat{H}_{z''}^*\right|^2 - 2 \left(\sum_{z'} \left|\hat{H}_{z'}\right|^2\right) \left|\sum_z \hat{I}_z \hat{H}_z^*\right|^2 \\ &= \left(\sum_z \left|\hat{I}_z\right|^2\right) \left(\sum_{z'} \left|\hat{H}_{z'}\right|^2\right)^2 - \left(\sum_{z'} \left|\hat{H}_{z'}\right|^2\right) \left|\sum_z \hat{I}_z \hat{H}_z^*\right|^2.\end{aligned}$$

Dividing the numerator by the denominator leads to

$$\frac{\left(\sum_z \left|\hat{I}_z\right|^2\right)\left(\sum_{z'} \left|\hat{H}_{z'}\right|^2\right)^2 - \left(\sum_{z'} \left|\hat{H}_{z'}\right|^2\right)\left|\sum_z \hat{I}_z \hat{H}_z^*\right|^2}{\left(\sum_{z'} \left|\hat{H}_{z'}\right|^2\right)^2} = \sum_z \left|\hat{I}_z\right|^2 - \frac{\left|\sum_z \hat{I}_z \hat{H}_z^*\right|^2}{\sum_{z'} \left|\hat{H}_{z'}\right|^2}.$$

The object-free error function is thus given by

$$\mathcal{L}(P) = \int \left\{ \sum_z \left|\hat{I}_z(\mathbf{k})\right|^2 - \frac{\left|\sum_z \hat{I}_z(\mathbf{k}) \hat{H}_z(\mathbf{k})^*\right|^2}{\sum_{z'} \left|\hat{H}_{z'}(\mathbf{k})\right|^2} \right\} \mathrm{d}\mathbf{k}. \tag{2.52}$$

References

[1] R. Kingslake, *The interferometer patterns due to the primary aberrations,* Transactions of the Optical Society **27**, 94 (1925).

[2] W. J. Bates, *A wavefront shearing interferometer,* Proceedings of the Physical Society **59**, 940 (1947).

[3] B. C. Platt and R. Shack, *History and principles of shack-hartmann wavefront sensing,* Journal of Refractive Surgery **17**, S573 (2001).

[4] R. K. Tyson, *Principles of adaptive optics* (CRC press, 2015).

[5] A. Roorda, F. Romero-Borja, W. J. D. III, H. Queener, T. J. Hebert, and M. C. Campbell, *Adaptive optics scanning laser ophthalmoscopy,* Opt. Express **10**, 405 (2002).

[6] M. J. Booth, M. A. Neil, R. Juškaitis, and T. Wilson, *Adaptive aberration correction in a confocal microscope,* Proceedings of the National Academy of Sciences **99**, 5788 (2002).

[7] D. Débarre, E. J. Botcherby, M. J. Booth, and T. Wilson, *Adaptive optics for structured illumination microscopy,* Optics express **16**, 9290 (2008).

[8] D. Débarre, E. J. Botcherby, T. Watanabe, S. Srinivas, M. J. Booth, and T. Wilson, *Image-based adaptive optics for two-photon microscopy,* Optics letters **34**, 2495 (2009).

[9] T. J. Gould, D. Burke, J. Bewersdorf, and M. J. Booth, *Adaptive optics enables 3d sted microscopy in aberrating specimens,* Optics express **20**, 20998 (2012).

[10] R. W. Gerchberg, *A practical algorithm for the determination of phase from image and diffraction plane pictures,* Optik **35**, 237 (1972).

[11] R. A. Gonsalves and R. Chidlaw, *Wavefront sensing by phase retrieval,* in *Applications of Digital Image Processing III*, Vol. 207 (International Society for Optics and Photonics, 1979) pp. 32–39.

[12] J. R. Fienup, *Phase retrieval algorithms: a comparison,* Applied optics **21**, 2758 (1982).

[13] J. R. Fienup, *Phase retrieval algorithms: a personal tour,* Applied optics **52**, 45 (2013).

[14] M. R. Teague, *Deterministic phase retrieval: a green's function solution,* JOSA **73**, 1434 (1983).

[15] T. E. Gureyev and K. A. Nugent, *Rapid quantitative phase imaging using the transport of intensity equation,* Optics communications **133**, 339 (1997).

[16] L. Waller, L. Tian, and G. Barbastathis, *Transport of intensity phase-amplitude imaging with higher order intensity derivatives,* Optics express **18**, 12552 (2010).

[17] L. Allen and M. Oxley, *Phase retrieval from series of images obtained by defocus variation,* Optics communications **199**, 65 (2001).

[18] W. Luo, A. Greenbaum, Y. Zhang, and A. Ozcan, *Synthetic aperture-based on-chip microscopy,* Light: Science & Applications **4**, e261 (2015).

[19] R. G. Paxman, T. J. Schulz, and J. R. Fienup, *Joint estimation of object and aberrations by using phase diversity,* JOSA A **9**, 1072 (1992).

[20] M. G. Löfdahl and G. Scharmer, *Wavefront sensing and image restoration from focused and defocused solar images.* Astronomy and Astrophysics Supplement Series **107**, 243 (1994).

[21] R. G. Paxman, J. H. Seldin, M. G. Lofdahl, G. B. Scharmer, and C. U. Keller, *Evaluation of phase-diversity techniques for solar-image restoration,* (1995).

[22] R. J. Noll, *Zernike polynomials and atmospheric turbulence,* JOsA **66**, 207 (1976).

[23] C. G. Broyden, *The convergence of a class of double-rank minimization algorithms,* IMA Journal of Applied Mathematics **6**, 76 (1970).

[24] R. Fletcher, *A new approach to variable metric algorithms,* The computer journal **13**, 317 (1970).

[25] D. Goldfarb, *A family of variable-metric methods derived by variational means,* Mathematics of computation **24**, 23 (1970).

[26] D. F. Shanno, *Conditioning of quasi-newton methods for function minimization,* Mathematics of computation **24**, 647 (1970).

[27] P. Wolfe, *Convergence conditions for ascent methods,* SIAM review **11**, 226 (1969).

[28] P. C. Hansen, *The l-curve and its use in the numerical treatment of inverse problems,* (1999).

3

Spatially-varying Aberrations Retrieval Using a Pair of Periodic Pinhole Array Masks

Yifeng Shao and Mikhail Loktev[*,†]

Parts of this chapter have been published in the proceding of Metrology, Inspection, and Process Control for Microlithography XXXI, **10145**, 101452S [1] and in Optics Express **27**,2 (2019) [2].
[*] Kulicke & Soffa Liteq B.V., Hooge Zijde 32 5626DC Eindhoven, The Netherlands
[†] mloktev@kns.com

3.1. Background

The function of an imaging system is to generate a perfect image (an Airy disc) for every point source in its field-of-view (FOV). However, the wavefront errors in the pupil of the imaging system blur these images by producing various types of aberrations. Imaging systems with large FOV are particularly important for several essential industrial applications, e.g. 3D printing and optical lithography for semiconductor manufacturing.

Optical lithography uses an imaging system to image the mask pattern, which is a magnified version of the image that has to be realized on the wafer, in the photoresist that has been spun on the wafer. In the the regions that are illuminated, a chemical reaction takes place in the photoresist, while in the other regions the photoresist remains unchanged. As a result, a demagnified resist pattern is formed and a semiconductor device, for example a logic or memory device, is manufactured.

The efficiency of optical lithography grows with the size of the FOV. In this chapter, we report experiments performed on a lithography system (a stepper) at Liteq B.V.. The projection lens has to produce a critical dimension ≤ 2 μm for i-line illumination at 355 nm uniformly in a rectangular FOV with size of 52 mm $\times$ 33 mm in one exposure.

To compare with, a normal microscope objective may have resolution 1 μm in a rectangular FOV with size of 1 mm $\times 1$ mm. As a result, the wavefront errors of the imaging system vary more significantly over the FOV in lithography than in microscopy and hence the calibration of spatially-varying wavefront errors of the lithographic projection lens is important for maintaining a uniform imaging performance.

Our goal in this chapter is to develop a method for calibrating the spatial variation of every type of aberration of an imaging system. This will give very valuable information for restoring the functionality of the imaging system. In a commercial lithographic projection lens, there are possibilities for adjusting the lens elements (6 degrees of freedom for positions and tilts). Active devices like deformable mirrors can be used to correct aberrations.

Traditional methods, such as interferometry [3, 4] and wavefront sensing [5] with e.g. a Shack-Hartmann sensor, require a point source in the object plane to provide an ideal reference wavefront (either planar or spherical) and an additional imaging system to map the wavefront in the pupil onto a detector. However, this additional imaging system has its own wavefront errors that will mix with the wavefront errors of the imaging system that is to be calibrated.

To measure the spatial variation of the wavefront errors, interferometry and wavefront sensing methods must be combined with varying the position of the point source. The wavefront errors can only be performed for one position at a time. Thus, the calibration will be time-consuming and may further be influenced by the positioning error introduced during this measurement process.

Alternative methods [6–13] are based on the fact that because wavefront errors lead to blurring of the image, one can in principle infer the wavefront errors from the amount of blurring. In most situations, more than one image is required, and these images should differ from each other in a known fashion [9–13]. A common

choice of diversity is the variation of the defocus setting. These images, of the same object, are blurred by the original wavefront error plus different amounts of extra defocus aberrations. As described in previous chapters, each of these blurred images should be sampled by sufficiently small and sufficiently many pixels. This is automatically guaranteed in imaging systems for image acquisition purpose, e.g. microscopes.

For lithography systems, the smallest feature size of importance for the pattern printed on the wafer (the image) is usually much smaller than the pixel size of any commercially available camera. As a result, one needs to first extract certain features from the image of a periodic object (e.g. a grating), for example the critical dimensions, using a wafer metrology technique called scatterometry [14], and then use the features to retrieve the wavefront errors [15].

An aerial image sensor [16–18] compares the image of a grating with a reference grating image and it measures the lateral and axial image translations caused by the odd and even order aberrations, respectively. By repeating the measurement for several gratings with various pitches and orientations (each combination of pitch and orientation corresponds to a sampling point in the pupil), one can determine the wavefront error in the pupil.

However, because both scatterometry and an aerial image sensor are difficult to be implemented in parallel, the calibration of spatially-varying wavefront errors will be very slow with these two techniques.

It has also been reported in the literature that a scanning electron microscope (SEM) can be used to acquire a high resolution image of the pattern printed in photo-resist [19]. However, the mechanism of SEM imaging may bring artifacts of various kinds to this image. For example, electron scattering is stronger by edges and corners than by flat areas. The nonlinear property of the photo-resist also plays an role because the printed pattern is not exactly equal to the aerial image on the wafer. Although images in multiple regions can be acquired simultaneously using a multi-beam SEM, the image acquisition time in the entire FOV is still too long to be practical.

3.2. Introduction of our method

In this chapter, we propose a fast, accurate and robust method for calibrating the spatially-varying aberrations of an anisoplanatic imaging system. Our method is based on the assumption that the spatially-varying aberrations can always be considered to be spatially-invariant in a sufficiently small sub-region of the FOV (which is often referred to as the isoplanatic patch). As a result, we implemented a measurement system that divides the FOV into a sufficiently large number of such (partially overlapping) sub-regions.

Here the FOV is defined as the object plane area and the corresponding image of this area, imaged through the projection lens, in the image space. A sub-region is defined as an area of the FOV that is imaged onto a pixel of the camera sensor through the camera lens.

We place two periodic pinhole array masks in the object plane (mask 1) and in the image plane (mask 2). We guarantee that the image of mask 1 and mask 2

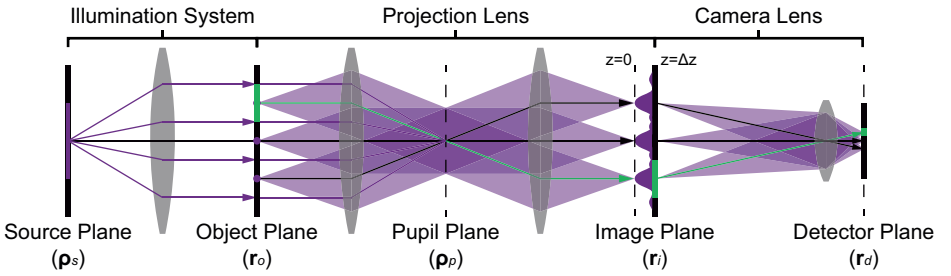

Figure 3.1: Plot of the lithography system and the measurement system. The lithography system consists of a illumination system and a projection lens. The measurement system consists of a camera lens and a camera sensor. Every camera pixel measures the total intensity in the corresponding sub-region in the FOV of the projection lens (one such sub-region is shown in green). So the total number of sub-regions is equal to the number of camera pixels.

have identical pitches. As a result, we can use the pinholes in mask 2 to sample the images of the pinholes in mask 1.

We remark that although the image of mask 1 is not periodic in the entire FOV, because of the spatially-varying aberrations, it is approximately periodic in each sub-region, where the aberrations can be regarded as spatially-invariant (but differ in different sub-regions).

A schematic plot of the lithography system and the measurement system is shown in Fig. 3.1. The lithography system consists of a Köhler illumination system, which uses a planar uniform incoherent monochromatic source, and a telecentric imaging system (e.g. a projection lens). The measurement system consists of an additional imaging system, referred to as the camera lens, and a camera sensor.

The coordinates of the object plane $\mathbf{r}_o$, the image plane $\mathbf{r}_i$, and the detector plane $\mathbf{r}_d$ are conjugated in this configuration as seen in Fig. 3.1 and are related by magnifications that are constant over the FOV. Furthermore, $\mathbf{r}_o$, $\mathbf{r}_i$, and $\mathbf{r}_d$ are scaled Fourier transform (FT) related variables of the pupil coordinates $\boldsymbol{\rho}_p$ and the source coordinates $\boldsymbol{\rho}_s$ of the planar source in Köhler illumination. For a pair of FT related variables, the Shannon-Nyquist sampling theorem must be fulfilled.

3.2.1. Interpretation of the measurement scheme

Our method uses a pair of periodic transmissive pinhole array masks, which are placed in the object and the image plane, respectively. The transmission function of the two masks are $t(\mathbf{r}_o)$ and $\tau(\mathbf{r}_i)$, respectively. We define the image plane as the plane at location Δz on the optical axis, with the nominal best image plane located at $\Delta z = 0$.

Consider a point source at $\boldsymbol{\rho}_s$ in the source plane which provides a plane wave for illumination in the object plane:

$$E_o(\mathbf{r}_o, \boldsymbol{\rho}_s) = S(\boldsymbol{\rho}_s) \exp(i \frac{2\pi}{\lambda f} \mathbf{r}_o \cdot \boldsymbol{\rho}_s), \tag{3.1}$$

where λ is the wavelength, f is the focal length of the illumination lens, and $S(\boldsymbol{\rho}_s)$

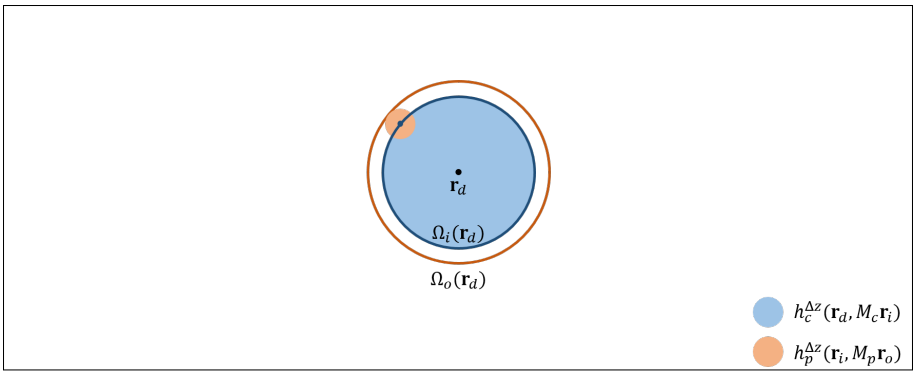

Figure 3.2: Schematic plot of the definition of the sub-regions. For every point $\mathbf{r}_d$ in the detector plane, we can define a sub-region $\Omega_i(\mathbf{r}_d)$ in the image plane (the deep blue circle) and a sub-region $\Omega_o(\mathbf{r}_d)$ in the object plane (the deep orange circle). The size of $\Omega_i(\mathbf{r}_d)$ is exactly equal to the size of the PSF of the camera lens (the light blue disc), while the size of $\Omega_o(\mathbf{r}_d)$ depends on the sizes of both the region $\Omega_i(\mathbf{r}_d)$ and the PSF of the projection lens (the light orange disc).

is the square root of the source intensity. For mask 1 in the object plane, the field in the image plane at Δz is given by

$$E_i^{\Delta z}(\mathbf{r}_i, \boldsymbol{\rho}_s) = \iint h_p^{\Delta z}(\mathbf{r}_i, M_p\mathbf{r}_o) t(M_p\mathbf{r}_o) E_o(M_p\mathbf{r}_o, \boldsymbol{\rho}_s) \mathrm{d}\mathbf{r}_o, \tag{3.2}$$

where $h_p^{\Delta z}(\mathbf{r}_i, M_p\mathbf{r}_o)$ is the PSF, M_p is the magnification, and the subscript p refers to the projection lens. We place mask 2 in the image plane. The resulting field in the detector plane is

$$E_d^{\Delta z}(\mathbf{r}_d, \boldsymbol{\rho}_s) = \iint h_c^{\Delta z}(\mathbf{r}_d, M_c\mathbf{r}_i) \tau(M_c\mathbf{r}_i) E_i^{\Delta z}(M_c\mathbf{r}_i, \boldsymbol{\rho}_s) \mathrm{d}\mathbf{r}_i, \tag{3.3}$$

where $h_c^{\Delta z}(\mathbf{r}_d, M_c\mathbf{r}_i)$ is the PSF, M_c is the magnification, and the subscript c refers to the camera lens. Finally, the intensity distribution in the detector plane due to one source point at $\boldsymbol{\rho}_s$ in the source plane is given by

$$I_d^{\Delta z}(\mathbf{r}_d, \boldsymbol{\rho}_s) = \left| E_d^{\Delta z}(\mathbf{r}_d, \boldsymbol{\rho}_s) \right|^2. \tag{3.4}$$

We remark that both PSFs are fields instead of intensities and Eq.3.4 considers only coherent imaging.

For the definition of the sub-region, we look at the imaging process reversely from the detector plane to the source plane. Suppose we place a point source at $\mathbf{r}_d$ in the detector plane. We denote the region of its image in the image plane, imaged through the camera lens, as $\Omega_i(\mathbf{r}_d)$ and its image in the object plane, imaged through the projection lens, as $\Omega_o(\mathbf{r}_d)$. Finally, we define that $\Omega_i(\mathbf{r}_d)$ and $\Omega_o(\mathbf{r}_d)$ are sub-region of the FOV for point $\mathbf{r}_d$.

Notice that both regions $\Omega_i(\mathbf{r}_d)$ and $\Omega_o(\mathbf{r}_d)$ have finite size because the PSFs of both imaging system eventually attenuate to zero at finite distance. The size of

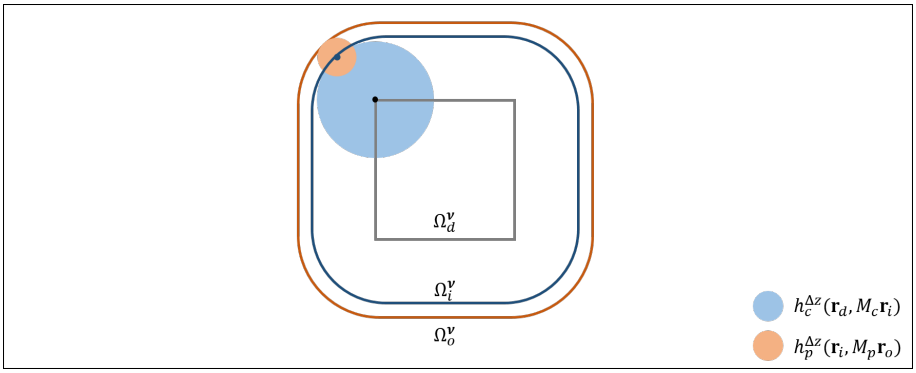

Figure 3.3: Schematic plot of the region of a given pixel in the detector plane, bounded by the gray rectangle, and the region of its image in the image plane, bounded by the deep blue rounded square, and in the object plane, bounded by the deep orange rounded square. The light blue and the light orange discs are the PSF intensity of the camera lens $|h_c^{\Delta z}(\mathbf{r}_d, M_c\mathbf{r}_i)|^2$ and of the projection lens $|h_p^{\Delta z}(\mathbf{r}_i, M_p\mathbf{r}_o)|^2$, respectively.

$\Omega_i(\mathbf{r}_d)$ is exactly equal to the size of the PSF of the camera lens $h_c^{\Delta z}(\mathbf{r}_d, M_c\mathbf{r}_i)$, while the size of $\Omega_o(\mathbf{r}_d)$ depends on the sizes of both the region $\Omega_i(\mathbf{r}_d)$ and the PSF of the projection lens $h_p^{\Delta z}(\mathbf{r}_i, M_p\mathbf{r}_o)$.

We can observe in Fig. 3.2 that the point $\mathbf{r}_d$ has a substantial contribution to all points in $\Omega_i(\mathbf{r}_d)$. Moreover, because at least one point in $\Omega_i(\mathbf{r}_d)$ has substantial contribution to some points in $\Omega_o(\mathbf{r}_d)$, the point $\mathbf{r}_d$ also has substantial contributions to all points in $\Omega_o(\mathbf{r}_d)$. In this way both sub-regions are defined.

We remark that the PSF of the camera lens and of the projection lens are both infinitely large by definition. However, we only need to consider a finitely large area in which the intensity of the PSF is above the threshold defined by the noise. When the lens is free of aberrations, it is customary to take the circular region between the center and the first zeros of the Airy disc as the region of both PSFs (illustrated by the blue and the orange disc). Therefore, both $\Omega_o(\mathbf{r}_d)$ and $\Omega_i(\mathbf{r}_d)$ also have a circular shape (illustrated by the blue and the orange circle).

Now we consider the intensity measured by the pixel, indexed by vector $\boldsymbol{\nu}$, on the camera. Let $\Omega_d^{\boldsymbol{\nu}}$ be the region occupied by a pixel. As a result, $\Omega_o^{\boldsymbol{\nu}}$ and $\Omega_i^{\boldsymbol{\nu}}$ are the union of $\Omega_o(\mathbf{r}_d)$ and $\Omega_i(\mathbf{r}_d)$, respectively, where $\mathbf{r}_d$ runs through Ω_d^{ν}. We illustrate these three regions in green in Fig. 3.1.

In Fig. 3.3, the rectangular region bounded by the gray curve is $\Omega_d^{\boldsymbol{\nu}}$ defined by the pixel $\boldsymbol{\nu}$ in the detector plane, while the regions bounded by the blue curve and by the orange curve are $\Omega_i^{\boldsymbol{\nu}}$ in the image plane and $\Omega_o^{\boldsymbol{\nu}}$ in the object plane, respectively.

For a particular pixel $\boldsymbol{\nu}$ in the detector plane, we can consider only a sub-region of the FOV, i.e. $\Omega_i^{\boldsymbol{\nu}}$ and $\Omega_o^{\boldsymbol{\nu}}$, in which the aberrations are spatially-invariant, although the aberrations are spatially-varying in the entire FOV.

Because the PSF of the projection lens is very small compared to the size of the PSF of the camera lens, we can assume that $\Omega_o^{\boldsymbol{\nu}} = \Omega_i^{\boldsymbol{\nu}}$. In Fig. 3.3 the regions

bounded by the blue and the orange curve thus have identical size and overlap with each other. Notice that if the pixel is very small, we will arrive at the situation depicted in Fig. 3.2. However, if the pixel is very large, both PSFs become very small compared to the size of $\Omega_d^{\boldsymbol{\nu}}$. So, all three regions, after being rescaled by the magnifications, respectively, are identical overlapping rectangles.

In the experiment, the FOV of the projection lens of size 52×33 mm is sampled by a camera sensor of size 500×320 pixel and pixel size 9.9×9.9 μm. Because the camera lens has a magnification $M_c = 1/9.9$, and the projection lens has a unit magnification $M_p = 1$, the sizes of $\Omega_i^{\boldsymbol{\nu}}$ and $\Omega_o^{\boldsymbol{\nu}}$ are both 98.1×98.1 μm. As a result, neighboring sub-regions have considerable overlap with each other.

As mask 1 and mask 2 have pitches of 4.5×4.5 μm, each sub-region contains more than 20×20 pitch. Consider that the camera pixel measures the total intensity in each sub-region. It measures the signal due to more than 400 pitches, which is equivalent to 400 times the signal of 1 pitch, because both the image of mask 1 and mask 2 are periodic in each sub-region.

For wavelength $\lambda = 355$ nm and NA on both side of the projection lens being $\mathrm{NA} = 0.12$, the size of the PSF, i.e. the region between the center and the first zeros of the Airy disc, of the projection lens is about $\lambda/(2\mathrm{NA}) \approx 1.5$ μm. Every sub-region $\Omega_i^{\boldsymbol{\nu}}$ and $\Omega_o^{\boldsymbol{\nu}}$ thus have a size of about 66×66 Airy discs of the projection lens.

The camera lens is operated at F number $F = 22$, which means that the nominal size of the PSF of the camera lens in the image plane is approximately 7.8 μm. Every sub-region thus has a size of about 12×12 Airy discs of the camera lens.

As a consequence, we can fairly assume that the PSFs of both the camera lens and the projection lens are translation-invariant in every sub-region $\Omega_i^{\boldsymbol{\nu}}$ and $\Omega_o^{\boldsymbol{\nu}}$. However, we remark that the projection lens is designed to be isoplanatic for both high resolution and large FOV, while the camera lens is designed to be isoplanatic only for only large FOV.

As a result, we have for the camera lens $h_c^{\Delta z}(\mathbf{r}_d, M_c\mathbf{r}_i) = h_c^{\Delta z}(\mathbf{r}_d - M_c\mathbf{r}_i)$, which is the same translational PSF in different sub-regions. In the contrast, we have $h_p^{\Delta z}(\mathbf{r}_i, M_p\mathbf{r}_o) = h_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - M_p\mathbf{r}_o)$ for the projection lens, which depends on not only the defocus distance Δz of the image plane but also the index $\boldsymbol{\nu}$ of the pixel that the sub-region is imaged onto.

Finally, the intensity distribution measured by the pixel $\boldsymbol{\nu}$ on the camera sensor due to the point source at $\boldsymbol{\rho}_s$ in the source plane is given by

$$I_d^{(\boldsymbol{\nu},\Delta z)}(\boldsymbol{\rho}_s) = \iint_{\Omega_d^{\boldsymbol{\nu}}} \left| \iint_{\Omega_i^{\boldsymbol{\nu}}} \iint_{\Omega_o^{\boldsymbol{\nu}}} h_c^{\Delta z}(\mathbf{r}_d - M_c\mathbf{r}_i)\tau(M_c\mathbf{r}_i) \times h_p^{(\boldsymbol{\nu},\Delta z)}(M_c\mathbf{r}_i - M_p\mathbf{r}_o)t(M_p\mathbf{r}_o)E_o(M_p\mathbf{r}_o,\boldsymbol{\rho}_s)\mathrm{d}\mathbf{r}_o\mathrm{d}\mathbf{r}_i \right|^2 \mathrm{d}\mathbf{r}_d. \tag{3.5}$$

By integrating over the source plane, we obtain the total intensity distribution:

$$I_d^{(\boldsymbol{\nu},\Delta z)} = \iint_{\Omega_s} I_d^{(\boldsymbol{\nu},\Delta z)}(\boldsymbol{\rho}_s)\mathrm{d}\boldsymbol{\rho}_s = \iint_{\Omega_d^{\boldsymbol{\nu}}} \iint_{\Omega_s} J_d^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_d,\boldsymbol{\rho}_s)\mathrm{d}\boldsymbol{\rho}_s\mathrm{d}\mathbf{r}_d. \tag{3.6}$$

where $J_d^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_d,\boldsymbol{\rho}_s)$ is the intensity distribution generated by the point source at $\boldsymbol{\rho}_s$ in the source plane and measured at $\mathbf{r}_d$ in the detector plane.

3.2.2. Discussion about the spatial coherence.

In the following paragraphs, for notation simplicity we will omit the magnification of the projection lens M_p and of the camera lens M_c. Exchanging the order of integration in Eq. (3.6), we obtain

$$I_d^{(\boldsymbol{\nu},\Delta z)} = \iint_{\Omega_i^{\boldsymbol{\nu}}}\iint_{\Omega_i^{\boldsymbol{\nu}}}\iint_{\Omega_o^{\boldsymbol{\nu}}}\iint_{\Omega_o^{\boldsymbol{\nu}}} W_i(\mathbf{r}_{i1},\mathbf{r}_{i2})\tau(\mathbf{r}_{i1})\tau(\mathbf{r}_{i2})^* \times h_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_{i1}-\mathbf{r}_{o1})h_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_{i2}-\mathbf{r}_{o2})^* t(\mathbf{r}_{o1})t(\mathbf{r}_{o2})^* W_o(\mathbf{r}_{o1},\mathbf{r}_{o2})\mathrm{d}\mathbf{r}_{o1}\mathrm{d}\mathbf{r}_{i1}\mathrm{d}\mathbf{r}_{o2}\mathrm{d}\mathbf{r}_{i2}, \tag{3.7}$$

where

$$W_i(\mathbf{r}_{i1},\mathbf{r}_{i2})\iint_{\Omega_d^{\boldsymbol{\nu}}} h_c^{\Delta z}(\mathbf{r}_d-\mathbf{r}_{i1})h_c^{\Delta z}(\mathbf{r}_d-\mathbf{r}_{i2})^*\mathrm{d}\mathbf{r}_d, \tag{3.8}$$

and

$$W_o(\mathbf{r}_{o1},\mathbf{r}_{o2}) = \iint_{\Omega_s} E_o(\mathbf{r}_{o1},\boldsymbol{\rho}_s)E_o(\mathbf{r}_{o2},\boldsymbol{\rho}_s)^*\mathrm{d}\boldsymbol{\rho}_s. \tag{3.9}$$

We define $W_o(\mathbf{r}_{o1},\mathbf{r}_{o2})$ and $W_i(\mathbf{r}_{i1},\mathbf{r}_{i2})$ as the mutual coherence function (MCF) in the object and the image plane, respectively. The MCF describes the correlation between the fields at all combinations of pairs of locations in a plane.

Suppose that the illumination system uses a planar incoherent monochromatic source. Due to the use of the Köhler illumination, each point of the source generates a plane wave in the object plane for illumination. So the total illumination field is the incoherent sum of the plane waves generated by all point sources. The MCF $W_o(\mathbf{r}_{o1},\mathbf{r}_{o2})$ thus describes the spatial coherence of the total illumination field in the object plane.

We focus on the case when the illumination field is fully spatially incoherent for the purpose of measuring the aberrations. Therefore, we have

$$W_o(\mathbf{r}_{o1},\mathbf{r}_{o2}) = I_o(\mathbf{r}_{o1})\delta(\mathbf{r}_{o1}-\mathbf{r}_{o2}), \tag{3.10}$$

where $I_o(\mathbf{r}_o) = \iint_{\Omega_s}|E_o(\mathbf{r}_o,\boldsymbol{\rho}_s)|^2\mathrm{d}\boldsymbol{\rho}_s$ is the sum of the intensities of the plane waves in the object plane. Usually, $I_o(\mathbf{r}_o)$ represents an uniform intensity distribution and hence can be neglected.

For Köhler illumination, spatially incoherent imaging is achieved when the ratio between the NA of the illumination lens and the projection lens is larger than two. When the NA of the projection lens is larger than 0.5, designing an illumination lens, of which the NA is larger than 1.0, is impossible. We remark that our method will also work for spatially coherent and partially coherent imaging.

The MCF $W_i^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_{i1},\mathbf{r}_{i2})$ has the following interpretation: suppose that the detector is our source and we look at a part of our source that is made of one pixel of the detector. The PSF of the camera lens $h_c^{\Delta z}(\mathbf{r}_d-\mathbf{r}_i)$ can thus be interpreted as the field in the image plane generated by a point source in the detector plane.

In our interpretation, we propagate the light reversely from the detector plane to the image plane. While in the Köhler illumination a point source is converted into a plane wave by the illumination lens, here a point source is converted into its own image (the PSF) by the camera lens.

Because the ratio between the NA of the camera lens (about 0.022), which plays the role as the illumination lens in our interpretation, and the projection lens (0.12) is very small, we have

$$W_i(\mathbf{r}_{i1}, \mathbf{r}_{i2}) = H_c^{\Delta z}(\mathbf{r}_{i1})\delta(\mathbf{r}_{i1} - \mathbf{r}_{i2}), \tag{3.11}$$

where $H_c^{\Delta z}(\mathbf{r}_i) = \iint_{\Omega_d^{\boldsymbol{\nu}}} |h_c^{\Delta z}(\mathbf{r}_d - M_c\mathbf{r}_i)|^2 \mathrm{d}\mathbf{r}_d$.

We further guarantee the fulfillment of the condition of spatially incoherent imaging by placing a ground-glass diffuser in the image plane. The ground-glass diffuser introduces random fluctuations to the phase of the field in the image plane, which kill the spatial correlation. The random fluctuations of the amplitude can be regarded as a kind of noise and will be averaged out later.

Finally, substituting Eq. (3.10) and (3.11) into Eq. (3.7), we obtain

$$I_d^{(\boldsymbol{\nu},\Delta z)} = \iint_{\Omega_i^{\boldsymbol{\nu}}} \iint_{\Omega_o^{\boldsymbol{\nu}}} H_c^{\Delta z}(\mathbf{r}_i) I_\tau(\mathbf{r}_i) H_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - \mathbf{r}_o) I_t(\mathbf{r}_o) I_o(\mathbf{r}_o) \mathrm{d}\mathbf{r}_o \mathrm{d}\mathbf{r}_i, \tag{3.12}$$

where $I_\tau(\mathbf{r}_i)$ and $I_t(\mathbf{r}_o)$ are the intensity of mask 2 $|\tau(\mathbf{r}_i)|^2$ and mask 1 $|t(\mathbf{r}_o)|^2$, respectively, and

$$H_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - M_p\mathbf{r}_o) = |h_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - M_p\mathbf{r}_o)|^2$$

is the intensity of the PSF of the projection lens, which depends on the aberrations in the sub-region imaged onto pixel $\boldsymbol{\nu}$. Recall that the spatial variance of the aberrations (the anisoplanatism of the projection lens) is indicated by the dependence on $\boldsymbol{\nu}$. In the sub-region (isoplanatism patch) for each $\boldsymbol{\nu}$, the aberrations as well as the PSF are shift-invariant.

3.2.3. Measurement of the PSF-like image

We rewrite Eq. (3.12) as

$$I_d^{(\boldsymbol{\nu},\Delta z)} = \iint_{\Omega_i^{\boldsymbol{\nu}}} H_c^{\Delta z}(\mathbf{r}_i) I_\tau(\mathbf{r}_i) \left[\iint_{\Omega_o^{\boldsymbol{\nu}}} H_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - \mathbf{r}_o) I_t(\mathbf{r}_o) \mathrm{d}\mathbf{r}_o \right] \mathrm{d}\mathbf{r}_i. \tag{3.13}$$

Notice that $\Omega_o^{\boldsymbol{\nu}}$ and $\Omega_i^{\boldsymbol{\nu}}$ are the object and the image plane sub-region that is imaged onto the pixel $\boldsymbol{\nu}$, respectively. Eq. (3.13) indicates that in the image plane at Δz, the image of mask 1 is transmitted by mask 2 and then modulated by the PSF intensity of the camera lens, and the total intensity is measured by the pixel $\boldsymbol{\nu}$ that the sub-region $\Omega_o^{\boldsymbol{\nu}}$ and $\Omega_i^{\boldsymbol{\nu}}$ are imaged onto.

The measured total intensity thus depends on the PSF, and hence depends on the aberrations, of the projection lens. Fig. 3.4 shows that the total intensity varies spatially due to the spatial variation of the aberrations. We remark that the PSF of the projection lens is related to the index $\boldsymbol{\nu}$ while the PSF of the camera lens is not.

Our method is similar to the Moire methods [20–24], or the aerial image sensor [16–18] in the sense that a blurred image (the image of mask 1) is transmitted by a reference (mask 2). The key is that we use the pinholes in mask 2 to sample the images of the pinholes in mask 1 as shown in Fig. 3.4.

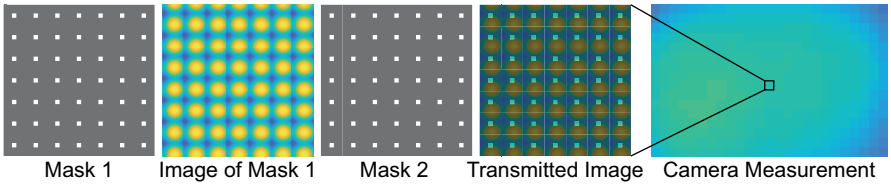

Figure 3.4: The total intensity measured by the camera. The camera pixel $\boldsymbol{\nu}$ measures the total intensity in the sub-region $\Omega_i^{\boldsymbol{\nu}}$ in the image plane at Δz. The total intensity is the sum of the image of mask 1 in the sub-region $\Omega_o^{\boldsymbol{\nu}}$ transmitted by mask 2. In the camera measurement, the total intensity varies due to the variation of the aberrations. Pictures are experimental data.

The image of mask 1 and mask 2 are required to have identical pitches. So each pinhole in mask 2 is aligned with, and hence samples an identical part of, the image of a pinhole in mask 1. As a result, the total intensity in one sub-region is equal to the intensity of one pitch times the number of pitches in one sub-region. Besides, by taking the total intensity we also average out the noise.

Finally, a 2-dimensional scan of mask 2 relative to the image of mask 1 allows the acquisition of a 2-dimensional high-resolution image of mask 1. The total intensity measured by the pixel $\boldsymbol{\nu}$ for scanning position $\Delta\mathbf{r}_i$ is given by

$$I_d^{(\boldsymbol{\nu},\Delta z)} = \iint_{\Omega_i^{\boldsymbol{\nu}}} H_c^{\Delta z}(\mathbf{r}_i) I_\tau(\mathbf{r}_i - \Delta\mathbf{r}_i) \left[\iint_{\Omega_o^{\boldsymbol{\nu}}} H_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - \mathbf{r}_o) I_t(\mathbf{r}_o) \mathrm{d}\mathbf{r}_o \right] \mathrm{d}\mathbf{r}_i. \tag{3.14}$$

Eq. (3.14) describes a 4-dimensional dataset of both the pixel index $\boldsymbol{\nu}$ (2-dimensional) and the scanning position $\Delta\mathbf{r}_i$ (2-dimensional). $I_d^{(\boldsymbol{\nu},\Delta z)}(\Delta\mathbf{r}_i)$ represents the image as a function of $\Delta\mathbf{r}_i$ for the sub-region that is imaged onto the pixel $\boldsymbol{\nu}$.

The sampling of $\boldsymbol{\nu}$ (the sampling of the FOV) depends on the magnification of the camera lens and the sampling given by the camera sensor. The sampling of $\Delta\mathbf{r}_i$ depends on the scanning process and must satisfy the Shannon-Nyquist sampling theorem with respect to the pupil coordinate $\mathbf{k}_p$ of the projection lens.

Suppose that both mask 1 and mask 2 are periodic arrays of rectangular pinholes. We can write the transmission function of both masks as

$$I_t(\mathbf{r}) = I_\tau(\mathbf{r}) = \sum_{\boldsymbol{\mu}} \mathrm{rect}(\mathbf{r} - \boldsymbol{\mu}), \tag{3.15}$$

where $\boldsymbol{\mu}$ denotes the location of the rectangular pinholes in the periodic array. We thus obtain that

$$I_d^{(\boldsymbol{\nu},\Delta z)}(\Delta\mathbf{r}_i) = \sum_{\boldsymbol{\mu}_1} \sum_{\boldsymbol{\mu}_2} \iint_{\Omega_i^{\boldsymbol{\nu}}} H_c^{\Delta z}(\mathbf{r}_i) \mathrm{rect}(\mathbf{r}_i - \boldsymbol{\mu}_2 - \Delta\mathbf{r}_i) I_{\mathrm{rect}}^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - \boldsymbol{\mu}_1) \mathrm{d}\mathbf{r}_i, \tag{3.16}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the locations in mask 1 and mask 2, respectively, and

$$I_{\mathrm{rect}}^{(\boldsymbol{\nu},z)}(\mathbf{r}_i - \boldsymbol{\mu}_1) = \iint_{\Omega_o^{\boldsymbol{\nu}}} H_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - \mathbf{r}_o) \mathrm{rect}(\mathbf{r}_o - \boldsymbol{\mu}_1) \mathrm{d}\mathbf{r}_o. \tag{3.17}$$

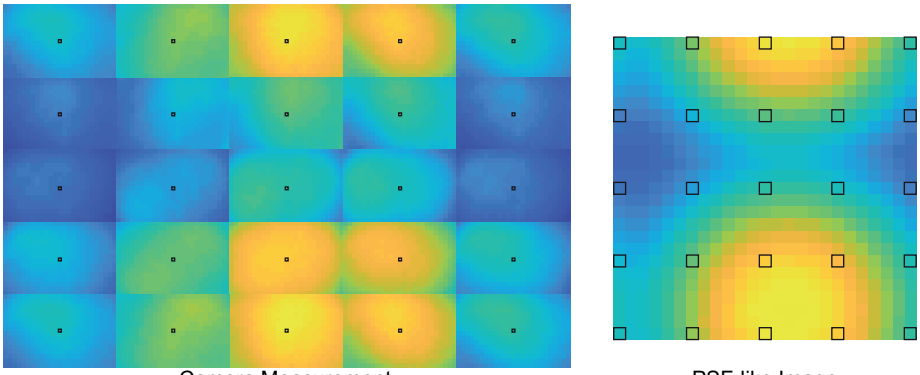

Figure 3.5: A scanning series of camera measurement (left) and the PSF-like image obtained by combining the total intensity measured by one camera pixel for all scanning positions (right). The values at the squares in the PSF-like image (right) are given by the values at the squares in the corresponding camera measurements (left).

We remark that in Eq. (3.16) and (3.17) we only need to consider $\boldsymbol{\mu}_2$ in the neighborhood of $\boldsymbol{\mu}_1$ that makes that the product $\mathrm{rect}(\mathbf{r}_i - \boldsymbol{\mu}_2 - \Delta\mathbf{r}_i) I_{\mathrm{rect}}^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - \boldsymbol{\mu}_1)$ does not vanish for all $\Delta\mathbf{r}_i$. In our method, we scan over one period area and hence $\Delta\mathbf{r}_i$ can only take values in one unit cell.

When the PSF of the projection lens $H_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - \mathbf{r}_o)$ is infinitely small (equivalent to a delta function), and hence $I_{\mathrm{rect}}^{(\boldsymbol{\nu},z)}(\mathbf{r}_i - \boldsymbol{\mu}_1) = \mathrm{rect}(\mathbf{r}_i - \boldsymbol{\mu}_1)$, the condition requires $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ to be identical. The number of $\boldsymbol{\mu}_2$ positions that needs to be considered for every $\boldsymbol{\mu}_1$ increases as the size of $H_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - \mathbf{r}_o)$ increases.

As a result, we can approximate the total intensity measured by the pixel $\boldsymbol{\nu}$ via scanning as

$$I_d^{(\boldsymbol{\nu},\Delta z)}(\Delta\mathbf{r}_i) \approx \mathrm{rect}(\Delta\mathbf{r}_i) \star \left[H_p^{(\boldsymbol{\nu},\Delta z)}(\Delta\mathbf{r}_i) * \mathrm{rect}(\Delta\mathbf{r}_i) \right], \tag{3.18}$$

where $\star$ and $*$ are the operator symbol of correlation and convolution, respectively, and $H_c^{\Delta z}(\Delta\mathbf{r}_i)$ has been neglected. Eq. (3.18) can be interpreted as the PSF intensity of the projection lens, in the sub-region that is imaged onto pixel $\boldsymbol{\nu}$, first convoluted with a pinhole in mask 1 and then correlated with a pinhole in mask 2.

Eq. (3.16) and (3.17) describe an image that is analogous to the PSF intensity, which is obtained by combining the total intensities measured by one camera pixel at all scanning positions. We illustrate this process in Fig. 3.5. The PSF-like image carries sufficient information about the aberrations in each sub-region. We remark that one scanning process allows the measurement of the PSF-like images for all sub-regions in parallel.

The PSF-like images are used to retrieve the defocus, tilt, and telecentricity, and the wavefront errors, in terms of the Zernike polynomials, in each sub-region. Retrieving the former requires estimating the relative shift and peak-to-valley value

of the PSF-like image, while retrieving the latter requires an optimization based on a model which describes the process of illumination, imaging, and measurement.

For both tasks, image acquisition should be performed not only in the nominal best image plane but also in several (at least one) defocused planes to obtain sufficient information. It is worth to mention that the retrieval of the aberrations, like the acquisition of the PSF-like images, can be performed simultaneously for all sub-regions. Therefore, our method is extremely efficient in comparison with existing methods [15–18].

In this chapter, we split our problem into two parts: the retrieval of only the defocus, tilt, and telecentricity (fast but rough), and of the wavefront aberrations in terms of the Zernike polynomials (slow but accurate).

3.3. Description of the experiment*

We validate our method by performing an experiment on a realistic lithography tool (a prototype of the lithography system developed by Liteq B.V.). The projection lens of the this lithography system has unit magnification, NA 0.128 and is designed for an operating wavelength 355 nm. This lithography system allows imaging with ≤ 2 µm critical dimension in a rectangular FOV with size of 52 mm $\times$ 33 mm.

Mask 1 and mask 2 are identical periodic arrays consisting of squared pinholes with pinhole width $w = 2.5$ µm and pitch $p = 4.5$ µm. Incident light is transmitted at the pinholes but reflected at other places. Both masks are fabricated in chrome on a fused silica substrate using e-beam lithography with ≤ 25 nm position accuracy in a rectangular area with size 55 mm $\times$ 35 mm.

The projection lens is illuminated by a Köhler illumination system with as the source a Q-switched pulsed laser operating at a wavelength of 355 nm, a bandwidth of 25 nm, and a repetition rate of 200 kHz. The laser beam is expanded to a diameter of 2 inch and scattered by a 20° engineered diffuser. The scattered laser beam acts like an effective source.

A condenser lens with focal length 200 mm is used to perform an optical Fourier transform between the effective source and mask 1, which are placed in the front and the back focal planes of the condenser lens, respectively. The effective source is large enough to overfill the pupil of the projection lens and hence guarantees a spatially incoherent imaging of mask 1.

The degree of spatial coherence of the imaging process can be tuned by varying the diameter of the expanded laser beam on the diffuser. The scattering angle of the diffuser guarantees that mask 1 is entirely covered by the illuminating light.

During the scanning process, mask 1 is fixed and mask 2 is movable. We mount mask 2 on a piezo stage (Physikinstrumente Plnano XYZ Piezo System) and place it in the image plane of the projection lens. A half inch size CCD camera equipped with a camera lens with 25 mm focal length is used to measure the intensity distribution in the image plane at each scanning position. The magnification of the camera lens is $1/9.9$ (demagnified by by $9.9\times$).

* The experiment is performed by Mikhail Loktev independently at Liteq B.V. on one of its prototype lithography systems.

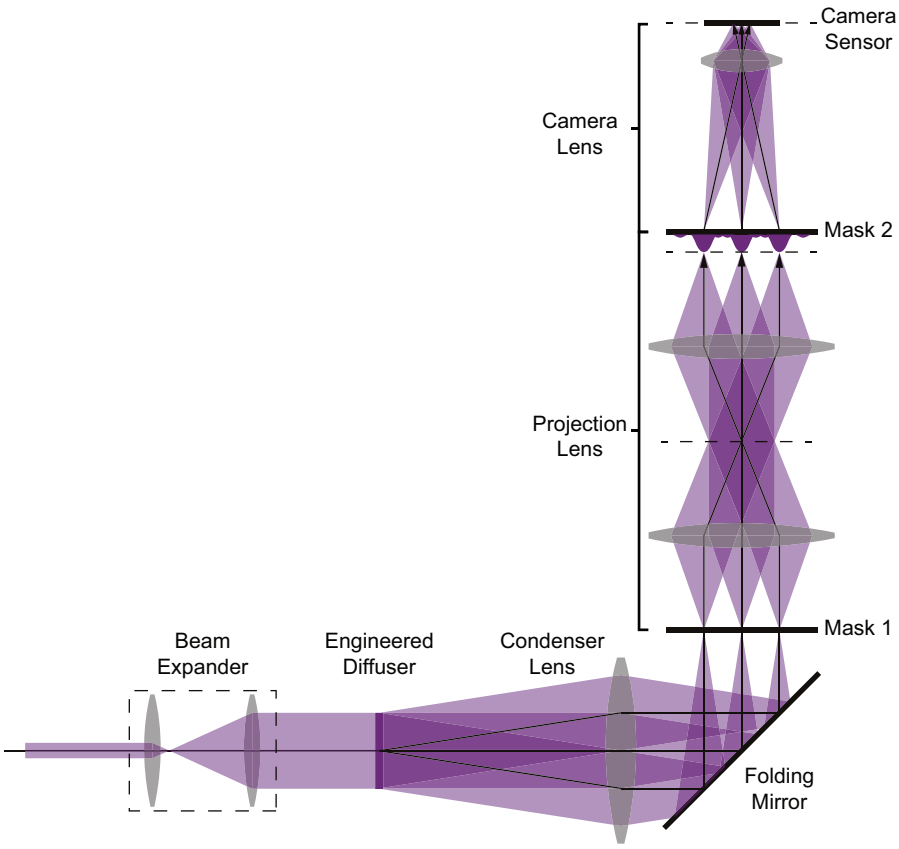

Figure 3.6: Plot of the experimental setup (a test model by Liteq BV). The lithography system is placed reversely in this setup such that the object plane (mask 1) is at the bottom and the image plane is at the top (mask 2). The illumination system is folded by 90 degrees with respect to the projection lens by a planar mirror. The beam expander controls the spatial coherence of the illumination on the object and the scattering angle of the engineered diffuser guarantees that the entire object can be illuminated.

We scan mask 2 with respect to the image of mask 1 in one image plane over one pitch area of 4.5 µm $\times$ 4.5 µm. As a result, every pixel of the camera measures a 2-dimensional high-resolution PSF-like image with size of 25 pixel $\times$ 25 pixel. The pixel size of the PSF-like image is given by the step size of the scanning process, which is 180 nm. The piezo stage offers nanometer accuracy.

We repeat the scanning process at 10 z locations (125 scanning positions in one image plane and 1250 scanning positions in total), located symmetrically on both sides of the nominally best image plane, with maximum defocus distance 25 µm and separation distance 5.56 µm which is equal to $0.26\pi\lambda/(\pi\text{NA})^2$. Therefore, every camera pixel measures 10 PSF-like images by repeating the scanning process at 10 z locations, respectively.

The FOV of the projection lens is divided into a number of sub-regions that is

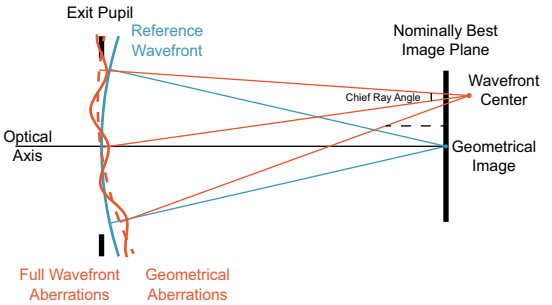

Figure 3.7: Schematic plot of the wavefronts at the exit pupil. The blue curve is the reference wavefront, which is a spherical wavefront with center at the geometrical image of the point source. The dashed and solid orange curves are the wavefronts affected by only tilt and defocus and by all aberrations, respectively.

equal to the number of the camera pixels. Because the optimization algorithm is non-linear and time-consuming, which takes few minutes to complete on a standard personal computer, retrieving the aberrations for all sub-regions is not feasible. We downsample the measurements from 320 pixel $\times$ 500 pixel to 16 pixel $\times$ 32 pixel. The result of the optimizaton algorithm will be 15 matrices, corresponding to the first 15 coefficients of aberrations, with size 16×32. Now we only need to perform retrieval using the optimization algorithm in $16 \times 32 = 512$ sub-regions, instead of in $320 \times 500 = 160,000$ sub-regions.

3.4. The retrieval of distortion, field curvature, and telecentricity*

A projection lens converts a diverging spherical wave with center at a point source in the object plane to a converging spherical wave with center at the geometrical image of this point source in the nominal best image plane. Assuming a perfect wavefront in the entrance pupil, we can describe the "error" of the projection lens by only the wavefront error in the exit pupil.

It is customary to decompose the wavefront errors by the Zernike polynomials, which are orthogonal polynomials defined on the unit disc that form a complete set. Each Zernike polynomial describes a particular type of aberration. Here we adopt the Noll's ordering of the Zernike polynomial [25]. In this section, we discuss the retrieval of the 2^{nd} and 3^{rd} term, the x and y tilt, and the 4^{th} term, the defocus. In later sections, we will discuss the retrieval of all terms of the Zernike polynomial.

In Fig. 3.7 we illustrate a reference wavefront, which is a spherical wavefront with center at the geometrical image of the point source (blue curve), and the wavefronts affected by the aberrations (orange curves). Tilt and defocus only translate the spherical wavefront center with respect to the geometrical image, while other aberrations further deviate the wavefront from the spherical shape. In Fig. 3.7,

* This section is based on the publication in the proceding of Metrology, Inspection, and Process Control for Microlithography XXXI, **10145**, 101452S [1], mainly contributed by Mikhail Loktev at Liteq B.V..

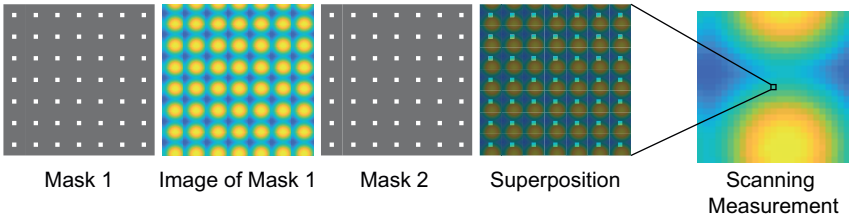

Figure 3.8: Interpretation of the PSF-like image measured by every camera pixel via scanning in one image plane. The value of every pixel is given by the total intensity of the image of mask 1 transmitted by mask 2 in one sub-region for one scanning position. The locations of the maximum pixel value and the image contrast determine the tilt and the defocus, respectively. Pictures are experimental data.

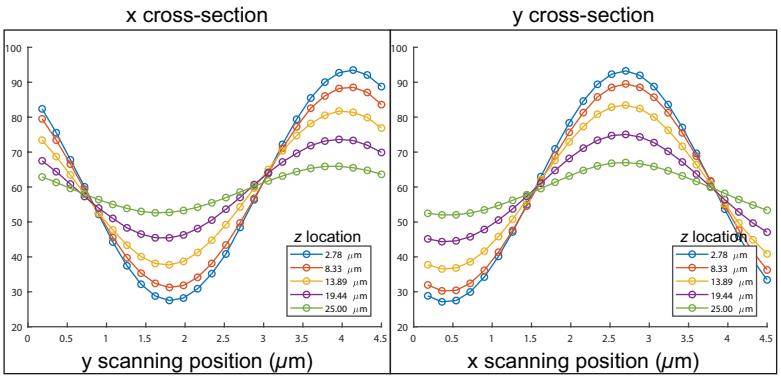

Figure 3.9: The x and y cross-section of the PSF-like images measured by the same camera pixel (in the same sub-region) by performing scanning in different image planes. Both amplitude and phase of these sinusoidal curves depend on the location z of the image plane on the optical axis. Dots and lines are the raw data and the fitted sinusoidal curve, respectively. Pictures are experimental data.

wavefronts affected by only tilt and defocus and by all aberrations are depicted by the dashed and the solid orange curve, respectively.

We define the chief ray as the ray connecting the exit pupil center and the spherical wavefront center. Fig. 3.7 shows that its slope and the intersect with the nominal best image plane gives the telecentricity and the tilt, respectively. For a lithographic projection lens, the chief rays in each sub-region should be all parallel to the optical axis (perpendicular to the image plane) and intersect with the nominal best image plane at the geometrical image. The defocus is given by the distance between the nominal best image plane and the image plane where the spherical wavefront center locates. Spatial variations of the tilt and the defocus in each sub-region are known as distortion and field curvature, respectively.

3.4.1. Description of retrieval method

We illustrate in Fig. 3.8 the PSF-like image measured by the camera pixel indexed by $\boldsymbol{\nu}$ via scanning in one image plane on the optical axis. The value of the pixel at $\mathbf{r}_s$ in the PSF-like image is given by the total intensity of the image of mask 1 and

transmitted by mask 2 for one scanning position $\boldsymbol{r}_s$ in the sub-region that is imaged onto the pixel $\boldsymbol{\nu}$.

The pixel value of the PSF-like image has a maximum when the image of mask 1 and mask 2 are aligned. So we can find the intersection of the chief ray with each image plane by finding the maximal pixel value in each PSF-like image and hence determine the tilt. The image contrast (the ratio between the maximum and minimum pixel value) depends on the relative distance between the image plane in which the scanning is performed and the best nominal image plane. We can find the location of the best nominal image plane by finding the maximal image contrast, which allows us to determine the defocus.

We use the x and y cross-section of the PSF-like image shown in Fig. 3.8 to determine both the location of the maximum pixel value and the image contrast.

We express the x and y cross-section respectively by

$$\begin{aligned} I_x(y) &= \int I(x,y)\mathrm{d}x, \\ I_y(x) &= \int I(x,y)\mathrm{d}y, \end{aligned} \tag{3.19}$$

where $I(x,y)$ is the intensity distribution of the PSF-like image in a sub-region. In order to take the cross-section along one direction, we integrate the PSF-like image along the other direction as seen in Eq. 3.19, instead of taking the cross-section at any particular pixel, e.g. at the pixel whose value is the maximum. This approach helps to stabilize the retrieval method against noise and other aberrations. The resulting cross-sections are sinusoidal curves because both the image of mask 1 (only in the sub-region) and mask 2 have identical periodicities.

We fit the x and y cross-sections respectively by

$$\begin{aligned} I_x(y) &= A_x \sin(B_x y + C_x) + D_x, \\ I_y(x) &= A_y \sin(B_y x + C_y) + D_y, \end{aligned} \tag{3.20}$$

where A, B, C, D are fitting parameters. The fitting is performed using the "fit" function in Matlab with the default 95% confidence and with the initial values of the fitting parameters being chosen empirically. We demonstrate an example of the fitting in Fig. 3.9, in which dots and lines are the raw data and the fitted sinusoidal curve, respectively. The image contrast and the maximal pixel value location are given by the amplitude and the phase, respectively.

Fig. 3.9 shows that the amplitude and the phase of the x and y cross-section vary with respect to the location z of the image plane on the optical axis. Variations of the former and the latter give the location of best nominal image plane (defocus) and the chief ray in terms of the slope (telecentricity) and the intersect with the nominal best image plane (tilt), respectively.

We can observe in Fig. 3.10 a quadratic curve of the image contrast versus the location z. The image contrast is the maximum (thus the image is the sharpest) in the nominal best image plane and decreases as the distance between the nominal best image plane and the image plane of measurement, in which the scanning is performed, increases. In Fig. 3.10, we fit the image contrast curve versus the

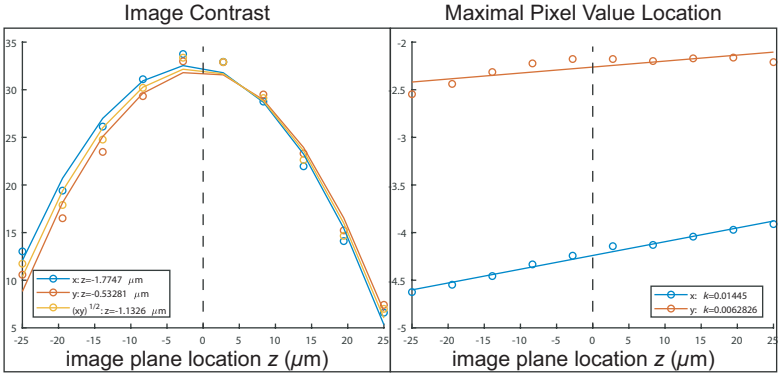

Figure 3.10: The variations of the image contrast and the maximal pixel value location with respect to the image plane location z. The location of the maximum of the image contrast curve (left) gives the location of the nominal best image plane, which gives the defocus, and the maximal pixel value location gives the intersect of chief ray with each image plane (right). The slop of this curve gives the telecentricity and the intersect of this curve with the nominal best image plane gives the tilt aberration. Pictures are experimental data.

location z by a quadratic function defined as

$$\begin{aligned} A_x(z) &= a_x(z - z_{0,x})^2 + b_x, \\ A_y(z) &= a_y(z - z_{0,y})^2 + b_y, \end{aligned} \tag{3.21}$$

where a, b, z_0 are fitting parameters. z_{0x} and z_{0y} give the location of the nominal best image plane according to the x and the y cross-section, respectively. Notice that z_{0x} and z_{0y} may not be identical as shown in Fig. 3.10. So we use the geometric mean $\sqrt{z_{0,x} z_{0,y}}$ to determine the defocus.

The location of the maximal image pixel value gives the intersect of the chief ray with each image plane. Fig. 3.10 shows that the chief ray is a linear function of the location z as expected. Therefore, we fit the chief ray by

$$\begin{aligned} C_x(z) &= c_x z + d_x, \\ C_y(z) &= c_y z + d_y, \end{aligned} \tag{3.22}$$

where c and d are the fitting parameters. d_x and d_y give the x and the y coordinate of the intersect of the chief ray with the nominal best image plane (at $z = 0$), respectably. c_x and c_y give the slope of the chief ray in the x and the y direction, respectively. In each sub-region of a perfect telecentric lithographic projection lens, d_x and d_y should be identical, while c_x and c_y should be identically zero.

We remark that the tilt (the deviation of the chief ray intersection) depends on the initial scanning position: the location of the maximum image pixel value may not necessarily be at the center of each image. Because all images are measured simultaneously by one scanning process, we can select the tilt in one sub-region, usually the sub-region corresponding to the pixel at the center of the camera (at the center of the FOV of the projection lens), to be the reference. We thus further determine the tilt in other sub-regions with respect to the reference.

In summary, in order to determine the tilt and the defocus in each sub-region, we need to do the following analysis:

1. Measuring PSF-like images by performing scanning in a series of image planes at various z locations on the optical axis.
2. Fitting the cross-sections of each PSF-like image by sinusoidal curves to determine the image contrast (amplitude) and the intersect of chief ray with the image plane in which the scanning is performed (phase).
3. Fitting the variations (with respect to the location z) of the image contrast and the chief ray intersection by quadratic curve and linear curve, respectively (Fig. 3.10).
4. Retrieving the location of the nominal best image plane (defocus), the slope of the chief ray (telecentricity) and the intersection of the chief ray with the nominal best image plane (tilt).
5. Finally, combining the retrieved tilt and defocus aberration for each sub-region to obtain the distortion and field curvature in the entire FOV. The telecentricity in the entire FOV can also be obtained.

3.4.2. Experimental Validation

The first step of experimental validation is to align the two masks because misalignment can introduce additional aberrations (mainly distortion) to the aberrations of the projection lens. The effect of the additional distortions is visible in every image measured by the camera and exhibits distinctive patterns associating with particular types of misalignment errors, which allows us to minimize the additional distortions by alignment.

There are 3 typical types of additional distortions caused by misalignment:

- **Magnification Error**: caused by the mismatch between the pitches of the image of mask 1 and mask 2.
- **Rotation Error**: caused by the relative rotation (around the optical axis) between the image of mask 1 and mask 2.
- **Higher Order Distortion**: caused by the misuse of conjugation planes (the pair of object and image planes).

Because the alignment uses visible effects of the additional distortions, the scanning process is not needed during alignment. As a result, our method provides a very useful real-time tool for 3-dimensional alignment of the object plane (mask 1) and the image plane (mask 2).

Fig. 3.11 shows the camera measurement for each step of alignment in the experiment. As seen in Fig. 3.11 (1), we started with severe misalignment that leads to a mixture of magnification error (the rectangular grid pattern), rotation error (rotation with respect to horizontal and vertical directions) and higher order

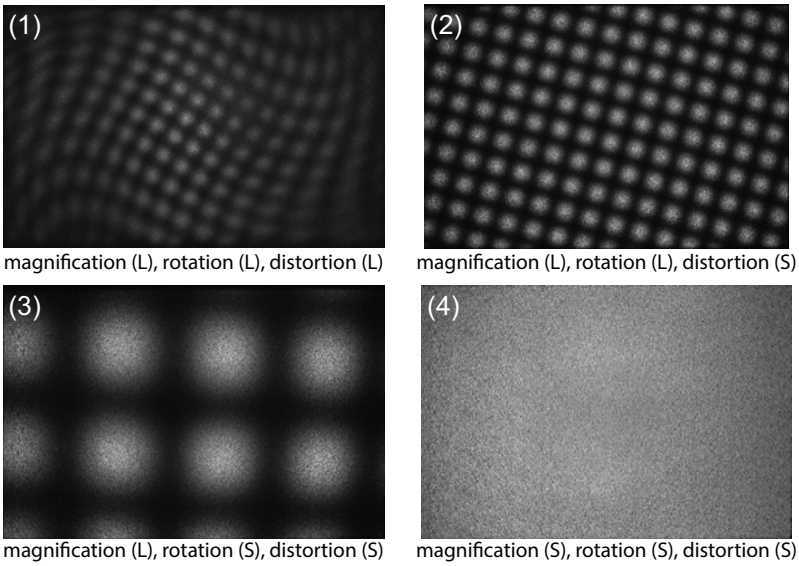

Figure 3.11: Camera measurements for steps (1-4) of masks alignment in the experiment. Effect of each type of distortion (magnification, rotation, and higher order distortion) can be observed visually in each camera measurement. The magnitudes of distortions are described by S (small) and L (large) in the brackets. The zoom settings for step (3) and step (4) are referred to as ”zoom 1” and ”zoom 2”, respectively.

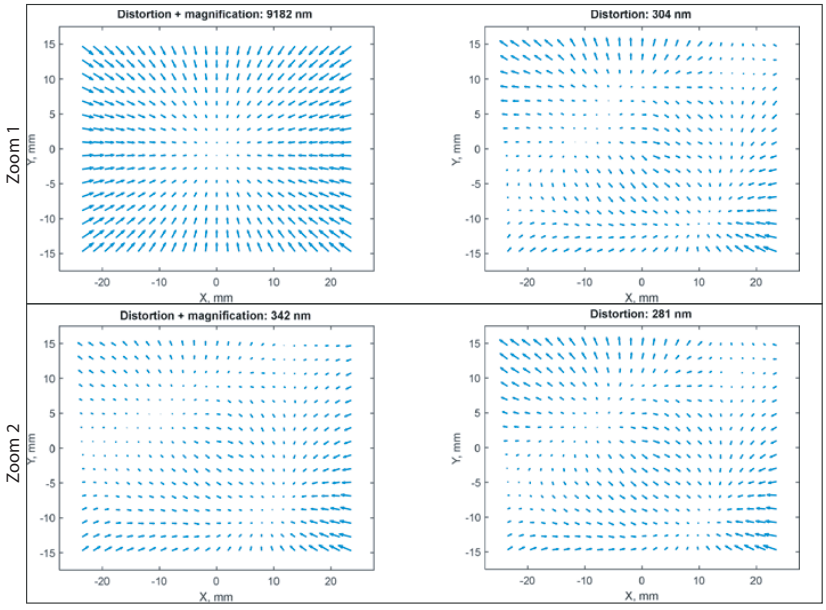

Figure 3.12: Retrieved distortions in zoom 1 and zoom 2 configuration. Left: the total distortion of both magnification error and other distortions. Right: only the contribution by other distortions. The magnitude of distortion in each plot is given by the maximum displacement.

distortion (irregular deformation). By correcting the locations of the conjugated planes on the optical axis, we have reduced the level of higher order distortion from ≥ 9 µm to ≤ 0.5 µm, leading to the camera measurement in Fig. 3.11 (2).

Further reduction of higher order distortion is difficult due to the difficulty of distinguishing the irregular deformation. By correcting the relative rotation between the two masks, we reduced the rotation error, which leads to a reduction of the number of fringes in Fig. 3.11 (3). Finally, by correcting the zoom setting, we further reduced the magnification error, thus reduce the number of fringes to a minimum, and obtained a visually uniform camera measurement as shown by Fig. 3.11 (4).

In order to validate the alignment process, we applied our method to two zoom settings of the projection lens, which are referred to as "zoom 1" and "zoom 2", respectively. The raw camera measurements in zoom 1 and zoom 2 can be seen in Fig. 3.11 (3) and Fig. 3.11 (4), respectively. Actually, zoom 1 and zoom 2 are the configuration before and after the zoom setting correction, which have the same higher order distortion and rotation error, but different magnification error.

In each configuration, we measure a series of PSF-like images via scanning and determine the distortion by combining the retrieved tilt in each sub-region (corresponding to each camera pixel).

Our goal is to verify that whether the zoom setting correction indeed corrects only the magnification error. Notice that the distortion can be decomposed according to the polynomials proposed by Braat, et al in [26]. The decomposition allows us to separate the magnification error (2nd order polynomial) from other distortions (3rd to 10th polynomials). The rotation error, however, does not belong to any of these polynomials, but is distributed in higher order distortions instead.

We present the distortions retrieved in the zoom 1 and zoom 2 configuration in the top and bottom panel of Fig. 3.12. We can observe that the total distortion has been reduced significantly from 9182 nm in zoom 1 to 342 nm in zoom 2 by correcting the zoom setting. In the former, the magnification error dominant, while in the latter, it is not. Besides, the plots of higher order distortions in both configurations show similar patterns and magnitudes. This is in agreement with our expectation that the zoom setting has only a trivial effect on higher order distortion.

We demonstrate the retrieval results of telecentricity and field curvature in Fig. 3.13 and Fig. 3.14, respectively. The 2 masks have been already aligned when performing the retrieval. So the remaining aberrations should be equal to the aberrations of the test projection lens. We can observe in Fig. 3.13 that the comparison between the retrieved and the simulated (using a ray tracing model) results of telecentricity matches each other. The field curvature result also matches the data measured using interferometry as shown in Fig. 3.14. As a consequence, we can finally verify that our method provides an efficient way for accurately retrieving distortion, field curvature, and telecentricity of the projection lens.

3.5. Retrieving full wavefront aberrations*

* This section is based on the publication in Optics Express **27**,2 (2019) [2].

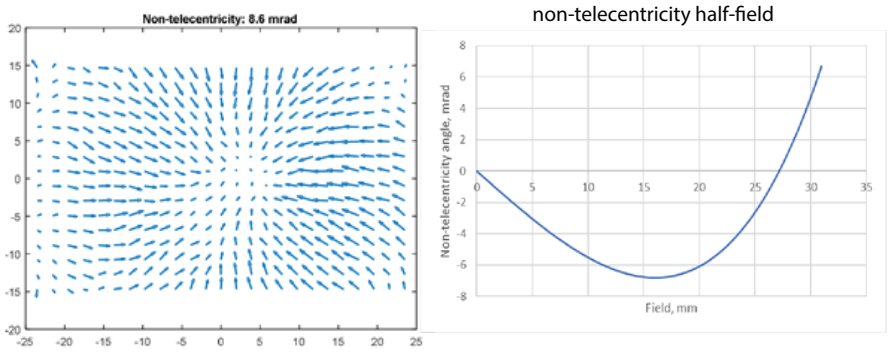

Figure 3.13: Results of telecentricity measurement. Left: experiment retrieval result. Right: ray tracing model simulation result along the x direction. The magnitude of non-telecentricity in the experimental retrieval result is given by the maximum chief ray angle.

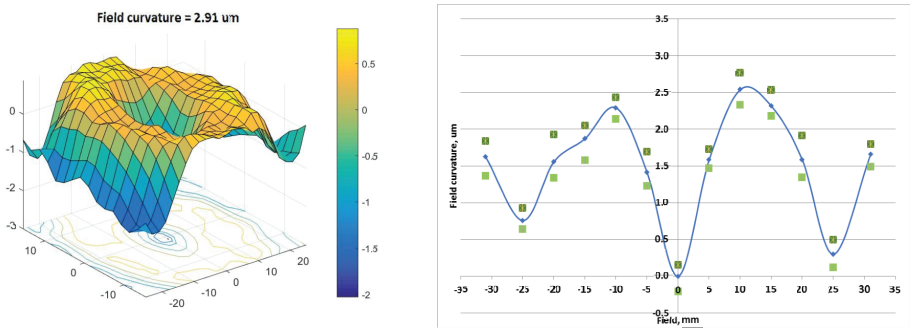

Figure 3.14: Results of field curvature measurement. Left: Experiment retrieval result. Right: Interferometry measurement result along the x direction. The magnitude of field-curvature is given by the maximum defocus distance.

As we described in the previous section, the retrieval of distortion, field curvature and telecentricity depends on the determination of the maximal pixel value location and the contrast of the PSF-like image measured by every camera pixel (Fig. 3.8). However, this determination may also be affected by other aberrations. For example, odd and even order aberration causes lateral shift and blur of the PSF, respectively, as discussed in [16–18]. Therefore, it is more appropriate to develop a method that handles full wavefront aberrations (all terms of the expansion in Zernike polynomials). This can be done using the same data but a different optimization approach.

We first need a model describing the PSF-like image in image plane at Δz measured by camera pixel $\boldsymbol{\nu}$ as a function of the scanning position $\Delta\mathbf{r}_i$. According to

Eq. 3.14, we can express the PSF-like image by

$$\begin{aligned} I_d^{(\boldsymbol{\nu},\Delta z)} &= \iint_{\Omega_i^{\boldsymbol{\nu}}} I_\tau(\mathbf{r}_i - \Delta\mathbf{r}_i) H_c^{\Delta z}(\mathbf{r}_i) \left[\iint_{\Omega_o^{\boldsymbol{\nu}}} H_p^{(\boldsymbol{\nu},\Delta z)}(\mathbf{r}_i - \mathbf{r}_o) I_t(\mathbf{r}_o) \mathrm{d}\mathbf{r}_o \right] \mathrm{d}\mathbf{r}_i \\ &= I_\tau(\Delta\mathbf{r}_i) \star \left[H_p^{(\boldsymbol{\nu},\Delta z)}(\Delta\mathbf{r}_i) * I_t(\Delta\mathbf{r}_i) \right], \end{aligned} \tag{3.23}$$

where $H_c(\Delta\mathbf{r}_i)$ has been neglected because it is much larger than one sub-region and hence it can be considered as a constant as the scanning covers only one sub-region. In Eq. (3.23), $H_p^{(\boldsymbol{\nu},\Delta z)}(\Delta\mathbf{r}_i)$ is the PSF intensity of the projection lens in the image plane at Δz in the sub-region that is imaged onto pixel $\boldsymbol{\nu}$. Notice that the resolution of $H_p^{(\boldsymbol{\nu},\Delta z)}(\Delta\mathbf{r}_i)$ is given by the scanning interval. We remark that Eq. (3.23) can be computed efficiently using the fast Fourier transform (FFT) algorithm.

3.5.1. Computation of the point-spread function

We denote the spatially-varying wavefront error of the projection lens by $\Phi^{\boldsymbol{\nu}}(\boldsymbol{\rho}_p)$, which is a 4-dimensional function of both the coordinate $\boldsymbol{\rho}_p$ of the pupil plane and the index $\boldsymbol{\nu}$ of the camera pixel. $\Phi^{\boldsymbol{\nu}}(\boldsymbol{\rho}_p)$ determines the spatially-varying PSFs in the nominal image plane and the defocused planes, which can be computed using the Debye diffraction integral [27, 28] as follows:

$$\begin{aligned} h_p^{(\boldsymbol{\nu},\Delta z)}(\Delta\mathbf{r}_i) &= \iint \exp\left[i2\pi\Phi^{\boldsymbol{\nu}}(\boldsymbol{\rho}_p)\right] \exp\left[-i\Delta z|\boldsymbol{\rho}_p|^2\right] \exp\left[-i2\pi\left(\Delta\mathbf{r}_i \cdot \boldsymbol{\rho}_p\right)\right] \mathrm{d}\boldsymbol{\rho}_p \\ &= \mathcal{F}\left\{\exp\left[i2\pi\Phi^{\boldsymbol{\nu}}(\boldsymbol{\rho}_p)\right] \exp\left[-i\Delta z|\boldsymbol{\rho}_p|^2\right]\right\}(\Delta\mathbf{r}_i), \end{aligned} \tag{3.24}$$

where $\mathcal{F}$ is the Fourier transform. Eq. (3.24) shows that $\Delta\mathbf{r}_i$, the scanning position, and $\boldsymbol{\rho}_p$ are a pair of Fourier transform variables. The sampling of $\Delta\mathbf{r}_i$ and $\boldsymbol{\rho}_p$ should satisfy the Shannon-Nyquist sampling theorem. Note that Eq. (3.24) describes a purely scalar model that is valid for imaging system with a low NA, i.e. NA ≤ 0.6. For imaging system with high NA, i.e. NA > 0.6, we need to use a vectorial model which incorporates the polarization effect.

In Eq. (3.24), the term $\exp\left[i2\pi\Phi^{\boldsymbol{\nu}}(\boldsymbol{\rho}_p)\right]$ represents the pupil function of the projection lens. The phase $\Phi^{\boldsymbol{\nu}}(\boldsymbol{\rho}_p)$ represents the wavefront error with respect to the ideal wavefront in the sub-region imaged onto pixel $\boldsymbol{\nu}$ and the amplitude is assumed to be uniform, which implies that any vignetting effect as well as any loss of light due to absorption or scattering are neglected in our model. The term $\exp\left[-i\Delta z|\boldsymbol{\rho}_p|^2\right]$ is the defocus term, where Δz is the normalized relative distance between the image plane of measurement and the nominal image plane.

We can decompose the 4-dimensional wavefront error as follows:

$$\Phi^{\boldsymbol{\nu}}(\boldsymbol{\rho}_p) = \sum_{m,n} \zeta_n^m(\boldsymbol{\nu}) Z_n^m(\boldsymbol{\rho}_p), \tag{3.25}$$

where ζ_n^m and Z_n^m are the aberration coefficients and the Zernike polynomials, respectively. For radial order n and azimuthal order m, the coefficient ζ_n^m represents

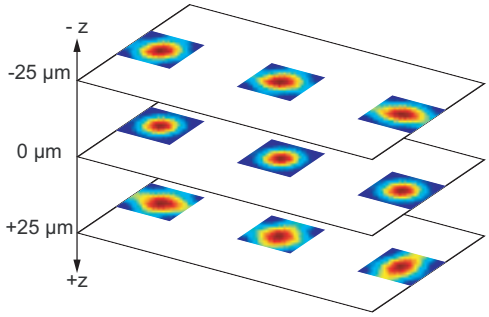

Figure 3.15: PSF-like images measured by three camera pixels (in three sub-regions) via scanning in three image planes. The FOV of the projection lens has a size of 32 mm × 55 mm. Three PSF-like images, each with size of 4.5 µm × 4.5 µm, can be measured simultaneously via one scanning process.

the weight of the Zernike polynomial $Z_n^m(\boldsymbol{\rho}_p)$. The wavefront error $\Phi^{\boldsymbol{\nu}}(\boldsymbol{\rho}_p)$ in the sub-region imaged onto pixel $\boldsymbol{\nu}$ is thus a weighted sum of all the aberrations.

We retrieve a set of coefficients $\boldsymbol{\zeta}(\boldsymbol{\nu})$ for each sub-region. By combining $\boldsymbol{\zeta}(\boldsymbol{\nu})$ retrieved in all sub-regions, we can thus determine the spatial variation of each aberration coefficient and retrieve the spatially-varying aberrations. Here the relation between the Zernike polynomials and the primary Seidel aberrations is beneficial for the diagnosis of projection lens.

3.5.2. Optimization for aberration coefficient retrieval

For each sub-region indexed by $\boldsymbol{\nu}$, we retrieve a set of coefficients $\boldsymbol{\zeta}(\boldsymbol{\nu})$ from a series of PSF-like images $I_d^{(\boldsymbol{\nu},\Delta z)}(\Delta\mathbf{r}_i)$ by solving an non-linear optimization problem. We will now temporarily omit the index $\boldsymbol{\nu}$ and the subscript d for the detector.

We formulate the optimization problem by defining an error function, which is the sum of the squared differences between the measurements and the predictions in all through-focus image planes. We find the solution to this optimization problem by updating the set of coefficients $\boldsymbol{\zeta}$ iteratively until the error function reaches a minimum. The error function is defined as follows:

$$\mathcal{L}(\boldsymbol{\zeta}) = \sum_{\Delta z} \iint \left[I^{\Delta z}(\Delta\mathbf{r}_i) - J^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta})\right]^2 \mathrm{d}\Delta\mathbf{r}_i, \tag{3.26}$$

where $I^{\Delta z}(\Delta\mathbf{r}_i)$ and $J^{\Delta z}(\Delta\mathbf{r}_i,\boldsymbol{\zeta})$ are the measurements and the predictions, respectively.

We remark that retrieving $\boldsymbol{\zeta}$ from only one image measured in the nominal image plane ($\Delta z = 0$) is not sufficient because the solution $\boldsymbol{\zeta}$ will not be unique. Therefore, at least one extra image should be measured in a defocused plane ($\Delta z \neq 0$). Fig. 3.15 shows that the differences between the PSF-like images in different sub-regions indexed by $\boldsymbol{\nu}$ are more significant in the defocused planes than in the nominal best image plane.

According to studies in [29, 30], the defocused planes are preferred to be located symmetrically on both sides of the nominal best image plane. They should be

sufficiently far from the nominal best image plane so the correlation between the PSF-like images are sufficiently small as mentioned in [31, 32]. However, they cannot be too far (further than $5\pi\lambda/(\pi\text{NA})^2$) otherwise Eq. (3.24) the Debye diffraction integral for computing the PSF will be invalid.

To minimize the error function, we need to derive an analytical expression for the gradient of the error function with respect to $\boldsymbol{\zeta}$. Using the chain rule, we obtain:

$$\frac{\partial \mathcal{L}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} = 2\sum_{\Delta z}\iint \left[J^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta}) - I^{\Delta z}(\Delta\mathbf{r}_i)\right]\frac{\partial}{\partial \boldsymbol{\zeta}}J^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta})\mathrm{d}\Delta\mathbf{r}_i. \tag{3.27}$$

The gradient of $J^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta})$ with respect to $\boldsymbol{\zeta}$ is given by

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\zeta}}J^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta}) &= \frac{\partial}{\partial \boldsymbol{\zeta}}\left\{I_\tau(\Delta\mathbf{r}_i)\star\left[H_p^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta}) * I_t(\Delta\mathbf{r}_i)\right]\right\}\\ &= I_\tau(\Delta\mathbf{r}_i)\star\left[\frac{\partial}{\partial \boldsymbol{\zeta}}H_p^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta}) * I_t(\Delta\mathbf{r}_i)\right],\end{aligned} \tag{3.28}$$

where the gradient of $H_p^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta})$ with respect to $\boldsymbol{\zeta}$ is given by

$$\frac{\partial}{\partial \boldsymbol{\zeta}}H_p^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta}) = 2\Re\left\{h_p^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta})^*\frac{\partial}{\partial \boldsymbol{\zeta}}h_p^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta})\right\}. \tag{3.29}$$

For each coefficient of aberration, we have

$$\frac{\partial}{\partial \zeta_n^m}h_p^{\Delta z}(\Delta\mathbf{r}_i;\boldsymbol{\zeta}) = \mathcal{F}\left\{i2\pi Z_n^m(\boldsymbol{\rho}_p)\exp\left[i2\pi\Phi(\boldsymbol{\rho}_p;\boldsymbol{\zeta})\right]\exp\left[-i\Delta z|\boldsymbol{\rho}_p|^2\right]\right\}. \tag{3.30}$$

In each iteration, the computation of the gradient of the error function with respect to each coefficient is the most time-consuming part in our method. We need to compute Eq. (3.27-3.30) as many times as the number of the coefficients. Typically, we need to consider the first 15 aberrations (the primary Seidel aberrations) for a quick diagnosis or the first 37 aberrations for a more thorough diagnosis. We set the initial guess of the coefficients to be all zeros (no aberrations), and we update the coefficients using the "fminunc" routine implemented in the Matlab.

3.5.3. Simulation results

In this section, we will perform qualitative analysis to determine the locations of measurement planes and to optimize the design parameters of the pair of masks. Our analysis is based on simulations.

For the simulations, we assume that both masks consist of identical periodic squared pinhole arrays with pinhole width w and pitch p. We use 37 aberration coefficients at 9 locations in the FOV of the projection lens (based on the Zemax ray tracing simulation data) to simulate the measurements (please see the table of aberration coefficients in Appendix I). Each measurement is normalized to unity and then converted to 16-bit precision data type to mimic the use of a CCD/CMOS camera.

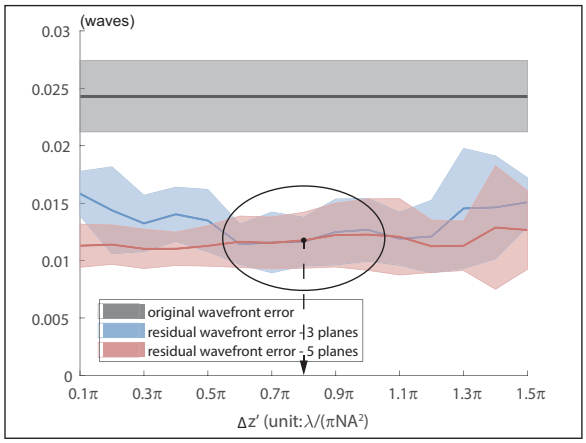

Figure 3.16: Plot of the residual wavefront error versus the locations of measurement planes. Gray curve shows the original wavefront error for simulating the measurements. Blue and red curves show the residual wavefront error corresponding to the aberration coefficients retrieved from 9 sets of simulated measurements in three planes (at $\Delta z = 0$, and $\pm\Delta z'$) and in five planes (at $\Delta z = 0$, $\pm 0.8\pi$, and $\pm\Delta z'$), respectively. Lines and shadings represent the mean and standard deviation of corresponding data. This simulation is noise free.

Determining the locations of measurement planes

In this part, we use masks with $w = 2.5$ μm and $p = 7.5$ μm. We scan mask 2 relative to the image of mask 1 in each measurement plane by 20 steps with step size 375 nm in two orthogonal directions, and we repeat the scanning process in a number of measurement planes with various defocus distance Δz (normalized by $\lambda/(\pi \mathrm{NA})^2$).

We define the residual wavefront error (WE) the squared difference between the original WE, used to simulate the measurements, and retrieved WE, retrieved from the simulated measurements. We evaluate the performance of the optimization algorithm using the root-mean-square (RMS) of the residual WE.

Suppose we take measurements in 3 planes in the focal region: one in the nominal image plane at $\Delta z = 0$ and the other two in two defocused planes located symmetrically on both sides of the nominal image plane at $\Delta z = \pm\Delta z'$. We vary the locations of both defocused planes by varying $\Delta z'$.

We calculate the RMS of the residual WE at each FOV location individually and calculate the mean and the standard derivation (SD) for 9 FOV locations, which are denoted by $\mu(\Delta z')$ and $\sigma(\Delta z')$, respectively.

In Fig. 3.16, we plot $\mu(\Delta z')$ (blue line) and $\sigma(\Delta z')$ (blue shading) for $\Delta z'$ ranging from 0.1π to 1.5π. We observe that the performance of the algorithm is the optimal, namely both $\mu(\Delta z')$ and $\sigma(\Delta z')$ are small, for $\Delta z'$ in the vicinity of 0.8π. Notice that both $\mu(\Delta z')$ and $\sigma(\Delta z')$ are slowly-varying functions of $\Delta z'$. The optimal region is broad and hence the optimization algorithm is not very sensitive to $\Delta z'$.

Next suppose we take measurements in 5 planes in the focal region: one in the nominal image plane at $\Delta z = 0$, and the other four in four defocused planes at

$\Delta z = \pm 0.8\pi$ and $\Delta z = \pm z'$, respectively. We aim to investigate that whether adding two extra defocused planes to the two optimal defocused planes can improve the performance of the optimization algorithm.

In Fig. 3.16 $\mu(\Delta z')$ (red line) and $\sigma(\Delta z')$ (red shading) are the mean and the SD in the case of using 5 measurement planes. Both $\mu(\Delta z')$ and $\sigma(\Delta z')$ are almost independent of $\Delta z'$, which indicates that we cannot improve the performance of the optimization algorithm by using more measurement planes, namely more than 3 measurement planes. However, we do observe that the region of optimal $\Delta z'$, which is already broad, becomes broader. In summary, using 3 optimally chosen measurement planes is sufficient.

Optimizing the mask design and sensitivity analysis

In this part, we investigate the influence of the mask design on the sensitivity of the optimization algorithm with respect to each order of the aberration. Here we adopt the Noll's index to enumerate the coefficients of the aberrations. The mapping from ζ_n^m with radial order n and azimuthal order m to ζ_ℓ with Noll's index ℓ can be found in [25]. We define the retrieval error of the coefficient as

$$\Delta\zeta_\ell(\boldsymbol{v}) = \left[\zeta_\ell(\boldsymbol{v}) - \zeta'_\ell(\boldsymbol{v})\right]^2 \tag{3.31}$$

where $\zeta_\ell(\boldsymbol{v})$ and $\zeta'_\ell(\boldsymbol{v})$ are the original and the retrieved coefficients respectively. In Fig. 3.17 and 3.18, we plot the the mean (line) and the SD (shading) of $\Delta\zeta_\ell(\boldsymbol{v})$ for 9 FOV locations. In both figures, we use 37 pre-calibrated coefficients to simulate the measurements , which are then fitted by the first 15 coefficients (left panel) and by the total number of 37 coefficients (right panel).

We introduce additive random noise, which follows a normal distribution with zero mean and one thousandth variance, to the simulated measurements to investigate the robustness of the optimization algorithm. In Fig. 3.17, we compare the errors of the coefficients retrieved from the noisy (red) and noise-free (blue) measurements. We find that the influence of normally distributed noise on the optimization algorithm is negligible. In practice normally distributed noise is often caused by thermal agitation of electrons in an electronic device, which is proportional to the temperature.

The design parameters of the masks are the pinhole width w and the pitch p. We compare the error of the retrieved coefficients for $w = 1.0$ µm (red) and for $w = 2.5$ µm (blue) in Fig. 3.18. We can observe that the error is independent of w when using the first 15 coefficients (left panel) but depends on w when using the total number of 37 coefficients (right panel).

The left panels in both figures show that we can use only the first 15 coefficients to fit the measurements simulated by 37 coefficients and obtain accurate results: for each aberration, the retrieval error is at least 2 orders of magnitude smaller than the original coefficient in terms of the mean and the SD. The right panels in both figures show that when using the total number of 37 coefficients to fit the measurements, we can obtain more accurate results for lower order (the first 15) coefficients than for higher order (the rest of the total number of 37) coefficients.

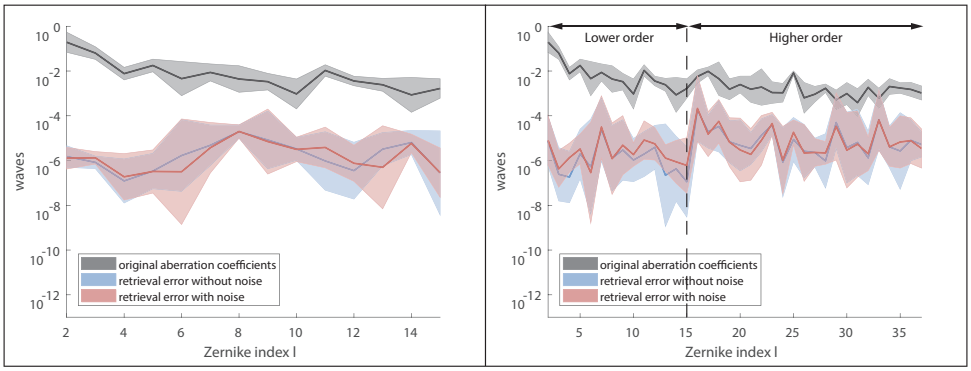

Figure 3.17: Comparison between the retrieval errors of the aberration coefficients with and without noise. We simulate measurements using 37 coefficients, which are fitted by the first 15 coefficients (left panel) and by the total number of 37 coefficients (right panel). The gray graph shows the original coefficients for simulating the measurements. The blue and the red graphs are the errors of the coefficients retrieved from the simulated measurements without and with noise, respectively. Lines and shadings are the mean and the SD of corresponding data.

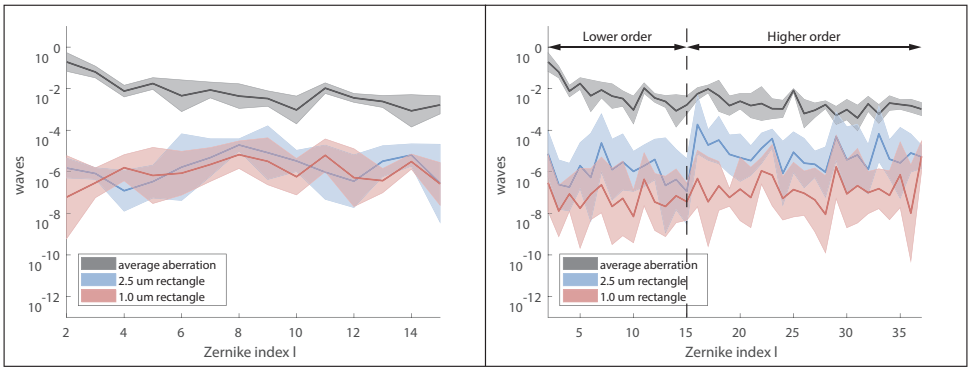

Figure 3.18: Comparison between the retrieval errors of the aberration coefficients as a function of pinhole width. The measurements are simulated using 37 coefficients, which are then fitted by the first 15 coefficients (left panel) and by the total number of 37 coefficients (right panel). The gray graph shows the original coefficients for simulating the measurements. The blue and red graphs are the errors of coefficients retrieved from measurements simulated using periodic mask with pinhole width $w = 1.0$ μm and $w = 2.5$ μm, respectively. Lines and shadings are the mean and the SD of corresponding data.

3.5.4. Experiment results

To validate the optimization algorithm, we compare the measured and the predicted (calculated using the retrieved coefficients of aberrations) PSF-like patterns in 6 through-focus planes at $\Delta z = \pm 2.78$ μm, ± 13.89 μm, and ± 25.00 μm respectively. The comparison result is plotted in Fig. 3.19. It is worth to note that the pitches of the masks used in the simulation (Section 3.5.3) and in the experiment (Section 3.3) are not identical.

Due to the telecentricity error, the PSF-like image measured in different planes is centered at different positions. So the optical algorithm needs extra coefficients

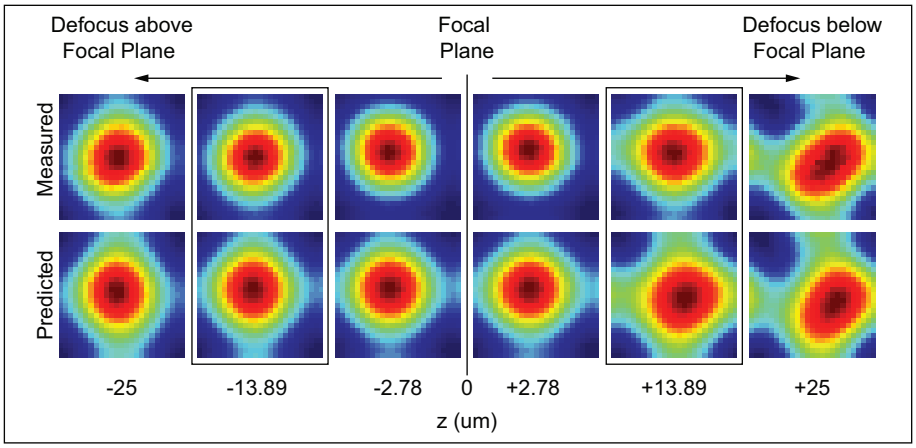

Figure 3.19: Comparison between the measurements and the predictions calculated using the retrieved 15 coefficients of aberrations. Unboxed figures are used for optimization while the boxed figures are not used.

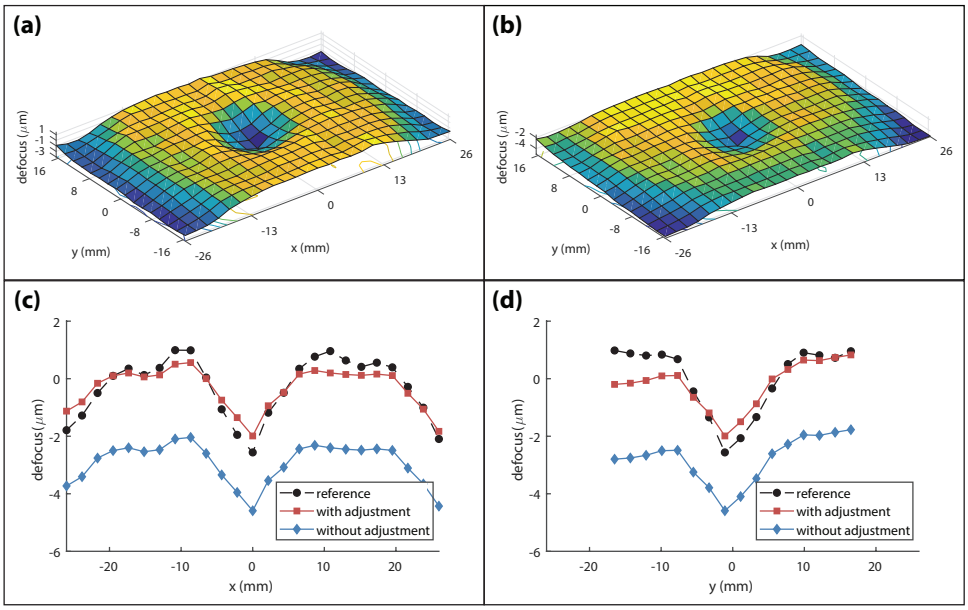

Figure 3.20: Comparison between the defocus aberration (field curvature) retrieved using our algorithm and the reference data in full FOV. (**a**) and (**b**) are the reference data and the retrieved field curvature, respectively. (**c**) and (**d**) are the x and y cross-sections of the plots in (**a**) and (**b**), respectively.

corresponding to the tilt terms of the Zernike polynomials to describe this phenomenon. We need $N_\ell + 2(N_{\Delta z} - 1)$ extra coefficients, where N_ℓ and $N_{\Delta z}$ are the number of the Zernike polynomials and the measurement planes, respectively. Namely, two extra coefficients per measurement plane.

In Fig. 3.19, four of the total six PSF-like patterns are used during the optimization (unboxed figures), while the other two are not used (boxed figures). The comparison in Fig. 3.19 shows that the predictions and the measurements are in good agreement in all six measurement planes. Therefore, we validate the retrieval of the spatially-invariant aberrations in each sub-region. By applying the optimization to all sub-regions, we can thus determine the spatially-varying aberrations in the entire FOV.

In Fig. 3.20, we compare the defocus aberration retrieved using our algorithm with the reference data [1], which is in agreement with interferometry data. The defocus aberration is related to the location of the nominal image plane on the optical axis (z axis). Due to the use of the piezo stage, the error of the location of each measurement plane can be neglected. However, the retrieved defocus may be biased (blue curve) because the nominal image plane may not be located exactly at $\Delta z = 0$.

As a result, we need to adjust the retrieved defocus to zero mean (red curve) and the adjusted defocus is now in agreement with the reference data (black curve). Fig. 3.20 also shows that for a lithography system, determining the defocus at a large number of FOV locations, i.e. obtaining a high resolution measurement of the field curvature, is necessary, because the spatial variation of the defocus can be fast and drastic.

3.6. Conclusion

In this chapter, we developed and validated an efficient, accurate, and robust method for measuring the spatially-varying aberrations of a lithography system. Our method does not require complex equipments, instead, simply uses a pair of periodic pinhole array masks, a 3-dimensional precision translation stage, and a CCD/CMOS camera. As a result, by taking hundreds of intensity measurements we can measure coefficients of aberrations at millions of FOV locations. This drastic difference between the magnitudes is the most attractive feature of our method.

Because our method uses only binary masks with transmission/reflection being either 1 or 0, our method can be applied to any arbitrary wavelength. Although only a low NA imaging system is considered in this chapter, our method also can be applied to measure the spatially-varying aberrations of a high NA imaging system by taking polarization effects into account. We recommend to use either spatially coherent or incoherent illumination for the measurement due to the consideration of computational efficiency. However, in principle, our method can also deal with spatially partially coherent illumination and retrieve the source intensity distribution.

For industrial applications such as lithography, whether choosing a direct method or an image based method depends on practical situations. The direct method usually provides instant and unambiguous measurement of the wavefront error, while the image based method requires parameterization of the wavefront error and an optimization procedure that is time consuming and sometimes yields non-unique result. However, when the spatial variation of the aberrations becomes an issue, the direct method cannot be parallelized for as many FOV locations as our method and the speed turns out to be the bottleneck. In contrast, our method consumes

identical time for optimization at one FOV location or at several FOV locations. The speed can even be further accelerated by using deep learning techniques to avoid optimization at all.

Appendix I: Simulation Data

In the table below we list the 7 sets of aberration coefficients for the simulation in Sub-section 3.5.3. These sets of aberration coefficients are the ray tracing data obtained using Zemax software. The peak-to-valley and the root-mean-square take the centroid of the PSF as the reference point and hence did not consider the piston and the tilt.

Table 3.1: Aberration coefficients for simulation (unit: waves)

Zernike Fringe Coefficient Index	x=+26 mm y=0 mm	x=+26 mm y=+16.5 mm	x=+26 mm y=-16.5 mm	x=0 mm y=0 mm	x=-26 mm y=0 mm	x=-26 mm y=+16.5 mm	x=-26 mm y=-16.5 mm
1	-0.0222	-0.0064	-0.0020	-0.0661	-0.0201	-0.0055	-0.0032
2	0.3733	0.4734	0.4474	-0.0688	0.0368	0.3820	-0.1313
3	0.0422	-0.1233	-0.0460	-0.0401	0.1353	-0.0278	-0.1128
4	-0.0111	-0.0066	-0.0044	-0.0245	-0.0084	-0.0035	-0.0066
5	0.0400	0.0146	0.0133	0.0078	0.0462	0.0153	0.0127
6	-0.0055	0.0114	-0.0165	-0.0019	-0.0001	-0.0146	0.0117
7	0.0052	0.0141	0.0189	0.0039	-0.0022	-0.0173	-0.0176
8	-0.0037	0.0087	-0.0123	-0.0003	-0.0020	0.0099	-0.0134
9	0.0006	-0.0057	-0.0035	0.0049	-0.0023	-0.0079	-0.0043
10	-0.0006	-0.0016	-0.0034	0.0000	-0.0038	0.0009	-0.0015
11	-0.0059	0.0136	-0.0218	-0.0055	-0.0062	0.0133	-0.0192
12	-0.0019	-0.0024	-0.0058	-0.0022	-0.0061	-0.0044	-0.0048
13	-0.0023	-0.0056	0.0007	0.0028	-0.0019	0.0031	-0.0035
14	0.0011	-0.0000	-0.0035	-0.0037	-0.0017	0.0001	0.0027
15	-0.0031	-0.0026	-0.0025	0.0002	0.0023	-0.0021	0.0020
16	-0.0066	0.0042	0.0043	-0.0245	-0.0051	0.0035	0.0037
17	-0.0147	0.0131	0.0093	-0.0053	-0.0144	0.0110	0.0054
18	-0.0043	-0.0187	0.0127	-0.0012	-0.0002	0.0147	-0.0127
19	0.0077	0.0024	-0.0004	-0.0005	-0.0035	-0.0013	-0.0013
20	0.0036	0.0005	0.0017	0.0023	0.0035	0.0058	-0.0046
21	-0.0068	-0.0034	-0.0002	0.0003	-0.0035	-0.0041	-0.0008
22	-0.0002	-0.0102	0.0092	0.0004	0.0003	0.0094	-0.0057
23	-0.0024	0.0004	0.0020	0.0024	0.0002	-0.0007	-0.0034
24	0.0037	-0.0002	-0.0014	-0.0019	-0.0010	0.0005	-0.0015
25	0.0093	0.0086	0.0076	0.0052	0.0091	0.0095	0.0088

Table 3.2: Aberration coefficients for simulation continued (unit: waves)

Zernike Fringe Coefficient Index	x=+26 mm y=0 mm	x=+26 mm y=+16.5 mm	x=+26 mm y=-16.5 mm	x=0 mm y=0 mm	x=-26 mm y=0 mm	x=-26 mm y=+16.5 mm	x=-26 mm y=-16.5 mm
26	-0.0005	-0.0016	0.0002	-0.0015	-0.0035	0.0000	-0.0014
27	-0.0002	0.0023	0.0008	0.0016	-0.0013	-0.0012	0.0007
28	0.0013	-0.0039	-0.0012	0.0023	0.0016	0.0012	0.0014
29	0.0024	0.0006	-0.0002	0.0004	-0.0005	-0.0013	0.0001
30	-0.0007	-0.0021	-0.0001	-0.0011	0.0006	0.0025	0.0030
31	0.0001	-0.0018	0.0003	-0.0001	-0.0003	-0.0023	0.0008
32	-0.0027	-0.0031	-0.0053	-0.0006	-0.0031	-0.0006	-0.0013
33	-0.0004	0.0021	-0.0010	0.0017	0.0003	0.0001	-0.0006
34	0.0004	-0.0015	-0.0028	-0.0021	0.0039	0.0039	0.0027
35	-0.0022	0.0017	0.0033	0.0016	0.0009	0.0009	0.0032
36	-0.0012	-0.0026	-0.0031	-0.0009	-0.0017	-0.0028	-0.0004
37	0.0003	0.0017	0.0015	-0.0011	0.0012	0.0024	-0.0005
peak-to-valley (to centroid)	0.1070	0.0890	0.0950	0.1293	0.1092	0.0938	0.0915
root-mean-square (to centroid)	0.01930	0.0150	0.0162	0.0184	0.0206	0.0152	0.0150

References

[1] M. Loktev and Y. Shao, *Projection lens testing with moiré effect,* in *Metrology, Inspection, and Process Control for Microlithography XXXI*, Vol. 10145 (International Society for Optics and Photonics, 2017) p. 101452S.

[2] Y. Shao, M. Loktev, Y. Tang, F. Bociort, and H. P. Urbach, *Spatially varying aberration calibration using a pair of matched periodic pinhole array masks,* Opt. Express **27**, 729 (2019).

[3] R. Kingslake, *The interferometer patterns due to the primary aberrations,* Transactions of the Optical Society **27**, 94 (1925).

[4] W. J. Bates, *A wavefront shearing interferometer,* Proceedings of the Physical Society **59**, 940 (1947).

[5] B. C. Platt and R. Shack, *History and principles of shack-hartmann wavefront sensing,* Journal of Refractive Surgery **17**, S573 (2001).

[6] R. W. Gerchberg, *A practical algorithm for the determination of phase from image and diffraction plane pictures,* Optik **35**, 237 (1972).

[7] R. A. Gonsalves and R. Chidlaw, *Wavefront sensing by phase retrieval,* in *Applications of Digital Image Processing III*, Vol. 207 (International Society for Optics and Photonics, 1979) pp. 32–39.

[8] J. R. Fienup, *Phase retrieval algorithms: a personal tour,* Applied optics **52**, 45 (2013).

[9] L. Allen and M. Oxley, *Phase retrieval from series of images obtained by defocus variation,* Optics communications **199**, 65 (2001).

[10] G. Zheng, X. Ou, R. Horstmeyer, and C. Yang, *Characterization of spatially varying aberrations for wide field-of-view microscopy,* Optics Express **21**, 15131 (2013).

[11] C. Van der Avoort *, J. J. M. Braat, P. Dirksen, and A. J. E. M. Janssen, *Aberration retrieval from the intensity point-spread function in the focal region using the extended nijboer–zernike approach,* Journal of Modern Optics **52**, 1695 (2005).

[12] L. Waller, L. Tian, and G. Barbastathis, *Transport of intensity phase-amplitude imaging with higher order intensity derivatives,* Optics express **18**, 12552 (2010).

[13] R. Paxman and J. Fienup, *Optical misalignment sensing and image reconstruction using phase diversity,* JOSA A **5**, 914 (1988).

[14] V. F. Paz, S. Peterhänsel, K. Frenner, and W. Osten, *Solving the inverse grating problem by white light interference fourier scatterometry,* Light: Science & Applications **1**, e36 (2012).

[15] S. van Haver, W. M. J. Coene, K. D'havé, N. Geypen, P. van Adrichem, L. de Winter, A. J. E. M. Janssen, and S. Cheng, *Wafer-based aberration metrology for lithographic systems using overlay measurements on targets imaged from phase-shift gratings,* Applied Optics **53**, 2562 (2014).

[16] T. Hagiwara, N. Kondo, I. Hiroshi, K. Suzuki, and N. Magome, *Development of aerial image based aberration measurement technique,* in *Optical Microlithography XVIII*, Vol. 5754 (International Society for Optics and Photonics, 2005) pp. 1659–1669.

[17] J. K. Tyminski, T. Hagiwara, N. Kondo, and H. Irihama, *Aerial image sensor: in-situ scanner aberration monitor,* in *Metrology, Inspection, and Process Control for Microlithography XX*, Vol. 6152 (International Society for Optics and Photonics, 2006) p. 61523D.

[18] A. Y. Bourov, L. Li, Z. Yang, F. Wang, and L. Duan, *Aerial image model and application to aberration measurement,* in *Optical Microlithography XXIII*, Vol. 7640 (International Society for Optics and Photonics, 2010) p. 764032.

[19] N. G. Orji, M. Badaroglu, B. M. Barnes, C. Beitia, B. D. Bunday, U. Celano, R. J. Kline, M. Neisser, Y. Obeng, and A. Vladar, *Metrology for the next generation of semiconductor devices,* Nature electronics **1**, 532 (2018).

[20] R.-S. Chang and C.-C. Lee, *Measurement of aberration by moiré pattern,* in *Optics in Complex Systems*, Vol. 1319 (International Society for Optics and Photonics, 1990) pp. 649–649.

[21] G. Haeusler and W. Jaerisch, *Method of moiré-metrical testing of optical imaging systems,* (1983), uS Patent 4,386,849.

[22] N. Kobayashi, *Lens distortion measurement using moiré fringes,* (1999), uS Patent 5,973,773.

[23] J. Heppner, J. Massig, M. Arnz, M. Kuechel, J. Penzing, and U. Schellhorn, *Moiré method and a system for measuring the distortion of an optical imaging system,* (2004), uS Patent 6,816,247.

[24] U. Wegmann, U. Schellhorn, R. Klaesges, and J. Stuehler, *Moiré method and measuring system for measuring the distortion of an optical imaging system,* (2006), uS Patent 7,019,824.

[25] R. J. Noll, *Zernike polynomials and atmospheric turbulence,* JOsA **66**, 207 (1976).

[26] J. Braat and P. Rennspies, *Effect of lens distortion in optical step-and-scan lithography,* Applied optics **35**, 690 (1996).

[27] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light* (Elsevier, 2013).

[28] E. Wolf and Y. Li, *Conditions for the validity of the debye integral representation of focused fields,* Optics Communications **39**, 205 (1981).

[29] D. J. Lee, M. C. Roggemann, and B. M. Welsh, *Cramér–rao analysis of phase-diverse wave-front sensing,* JOSA A **16**, 1005 (1999).

[30] B. H. Dean and C. W. Bowers, *Diversity selection for phase-diverse phase retrieval,* JOSA A **20**, 1490 (2003).

[31] O. El Gawhary, A. Wiegmann, N. Kumar, S. Pereira, and H. Urbach, *Through-focus phase retrieval and its connection to the spatial correlation for propagating fields,* Optics Express **21**, 5550 (2013).

[32] A. Polo, S. F. Pereira, and P. H. Urbach, *Theoretical analysis for best defocus measurement plane for robust phase retrieval,* Optics letters **38**, 812 (2013).

4

MCF measurement using self-referencing holography

Parts of this chapter have been published in Optics express **26**,4 (2018) [1].

4.1. Introduction

Spatial coherence is among the fundamental properties of light, which describes the statistical correlation between the light field at a pair of locations. The measurement of spatial coherence plays an important role in a broad range of key applications such as beam shaping [2, 3], free-space optical communication through turbulent atmosphere [4], illumination for advanced imaging system (e.g. lithography) [5], and superresolution imaging [6].

The spatial coherence of an arbitrary light beam can be completely characterized by the complex-valued 4-dimensional mutual coherence function (MCF). However, the measurement of the MCF is remarkably challenging. The problem is particularly severe for methods that measure either the intensity-intensity correlation [2, 3] or the interference pattern [7] between light field at all possible combination of pairs of locations.

Other interference based methods methods measure fringe visibility along either a single direction [8–10] or multiple directions [11, 12], which, however, rely on the symmetry of the MCF. An alternative is to describe the MCF by an analytical model, e.g. the Gauss-Schell model [13]. As a result, the MCF can be determined by fitting a set of parameters.

The phase space methods [14–17] measure the diffraction pattern of the light that propagates through a tiny window. It should be considered that the MCF is a function of two two-dimensional spatial coordinates (locations). This approach is equivalent to measuring the Fourier transform with respect to one coordinate, while setting the other coordinate to be at the location of the tiny window. As a result, we only need to scan the tiny window, and take one measurement at each location. However, both phase space methods [14–17] and the methods mentioned in [2, 3, 11, 12] cannot measure both the amplitude and the phase of the MCF. The phase has significant impact on the propagation of light beam and is essential for phase-sensitive applications.

For diffractive imaging, characterizing the spatial coherence of the illumination light beam is essential. Diffractive imaging reconstructs the information of an object from the diffracted far-field intensity pattern. In the x-ray and the electron regime, diffractive imaging is particularly useful due to the lack of high-quality and low-cost optics. Because no spatially coherent illumination source like a laser is available in this regime, the performance of diffractive imaging is severely degraded.

Experimentally, an illumination source that is almost spatially coherent can only be obtained via spatial filtering, but at the cost of severe light loss. Typically, the light generated by a synchrotron source needs to be propagated through a micrometer sized pinhole and then propagated for about a hundred meter distance before illuminating an object. After this process, a significant amount of beam flux is lost.

In order to use spatially partially coherent illumination for diffractive imaging, several modifications to the current algorithms have been developed to take into account the propagation of the MCF instead of the coherent light field. In [18, 19], authors interpret the diffraction pattern as a convolution between a shift-invariant MCF and the coherent diffraction pattern. Alternatives decompose the shift-variant MCF by a set of coherent modes [20, 21] such that the diffraction pattern becomes

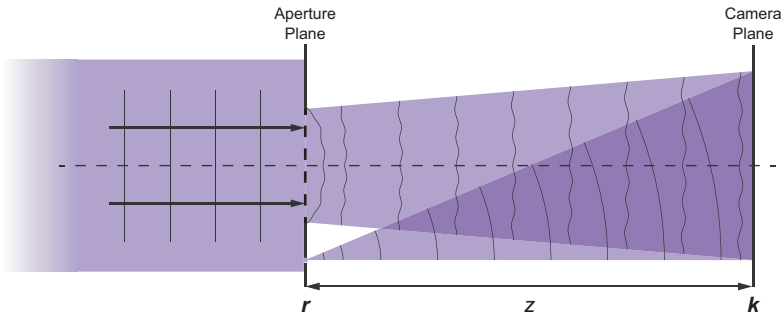

Figure 4.1: Schematic plot of the MCF measurement concept. An arbitrary beam illuminates an object in the aperture plane. The incident beam is split into the object wave and a reference wave by perturbing the light field at a particular "point" in the aperture plane. These two waves interfere with each other in the camera plane and the interference pattern (diffraction pattern) is measured.

a superposition of the diffraction pattern of each coherent mode. However, mode decomposition is effective only when the degree of spatial coherence is rather high and only a few number of modes is required. In the literature, to our knowledge, none of the non-iterative diffractive imaging algorithms, e.g. [22–24], have been adapted to use spatially partially coherent illumination.

4.2. MCF measurement using holography

Here we present a method for measuring the complex-valued 4-dimensional MCF of an arbitrary light beam using holography. Our method neither uses prior knowledge nor imposes any requirement on the structure of the MCF. The concept of our method is sketched in Fig. 4.1. We let the target light beam be transmitted by an aperture and measure the far-field diffraction pattern. Unlike any other methods, our method requires perturbation to the transmission in the aperture plane at a particular "point".

We can realize the point perturbation by using a spatial light modulator (SLM), which can either switch the amplitude between 0 and 1 or varying the phase continuously from $-\pi$ to $+\pi$. The point perturbation allows us to split the incident field into two parts: the original transmitted wave and a reference wave.

Notice that the reference wave is generated by applying the point perturbation to the incident field. The reference wave can be considered as being emitted by a point source located at the perturbation point. As a result, the reference wave is a spherical wave, which becomes a plane wave in the far-field.

We remark that in our method, the correlation between the reference wave and the original transmitted wave is preserved. The diffraction pattern is thus given by the interference pattern between both waves in the far-field.

Using holography, we retrieve the correlation function between the incident field at the perturbation point (reference wave) and all the locations in the aperture plane (original transmitted wave) from the diffraction pattern. We retrieve a 2-dimensional correlation function which can be regarded as a "slice" of the complete

4-dimensional MCF. As a consequence, the 4-dimensional MCF can be measured by a 2-dimensional scanning of the perturbation point.

Our method can be used for diffractive imaging by superposing the aperture with a transmissive object. In this situation, we retrieve the product of the transmission of the object and the correlation function of the illumination beam with respect to the perturbation point, which plays a role as the modulation. By calibrating and compensating the modulation, we can obtain the transmission of the object alone.

4.2.1. Description of the diffraction pattern

The concept of our method is illustrated schematically in Fig. 4.1. Let the coordinates in the aperture plane and the camera plane be denoted by $\mathbf{r}$ and $\mathbf{k}$, respectively. We denote the field of the incident beam by $E_i(\mathbf{r})$. The domain of the aperture is denoted by Ω and the transmission function $T(\mathbf{r})$ in the aperture plane is given by

$$T(\mathbf{r}) = \begin{cases} 1 & \mathbf{r} \in \Omega \\ 0 & \mathbf{r} \notin \Omega \end{cases}. \tag{4.1}$$

The transmission function is 1 inside and 0 outside the aperture. When superposing a transmissive object with the aperture for diffractive imaging, we can simply multiply $T(\mathbf{r})$ with the transmission function of the object $O(\mathbf{r})$.

We perturb the original transmitted wave by varying the transmission function $T(\mathbf{r})$ at a particular "point" $\mathbf{r} = \mathbf{r}_p$ in the aperture plane. The point perturbation is denoted by a Dirac delta function $C\delta(\mathbf{r} - \mathbf{r}_p)$, where C is a complex-valued constant that represents the variation of $T(\mathbf{r}_p)$. The perturbed transmission function is expressed as

$$\begin{aligned} T_p(\mathbf{r}) &= [T(\mathbf{r}) - T(\mathbf{r})\delta(\mathbf{r} - \mathbf{r}_p)] + CT(\mathbf{r})\delta(\mathbf{r} - \mathbf{r}_p) \\ &= T(\mathbf{r}) + C_p\delta(\mathbf{r} - \mathbf{r}_p). \end{aligned} \tag{4.2}$$

where $C_p = (C - 1)T(\mathbf{r}_p)$. Eq. (4.2) shows that at the perturbation point $\mathbf{r} = \mathbf{r}_p$, the original transmission function $T(\mathbf{r}_p)$ is varied by C, while at other locations $T(\mathbf{r})$ remains unchanged. Alternatively, we can also regard Eq. (4.2) as a sum of the original transmission function $T(\mathbf{r})$ at all locations and a Dirac delta function $C_p\delta(\mathbf{r} - \mathbf{r}_p)$ at $\mathbf{r} = \mathbf{r}_p$. As can be seen in Eq. (4.2)□ the incident field will be naturally split into an object wave and a reference wave, which propagates through $T(\mathbf{r})$ and $C_p\delta(\mathbf{r} - \mathbf{r}_p)$, respectively.

The intensity distribution of the diffraction pattern (in the approximation of Fraunhofer diffraction) in the camera plane is written as:

$$\begin{aligned} I(\mathbf{k}) = \iint \iint \langle E_i(\mathbf{r}_1)E_i(\mathbf{r}_2)^*\rangle T_p(\mathbf{r}_1)T_p(\mathbf{r}_2)^* \\ \times \exp\left[-i2\pi\mathbf{k}\cdot(\mathbf{r}_1 - \mathbf{r}_2)\right] \mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2, \end{aligned} \tag{4.3}$$

where $\langle\cdot\rangle$ represents the ensemble averaging and

$$J(\mathbf{r}_1, \mathbf{r}_2) = \langle E_i(\mathbf{r}_1)E_i(\mathbf{r}_2)^*\rangle, \tag{4.4}$$

is the MCF of the incident field, which describes the correlation between the incident field at a pair of locations $\mathbf{r_1}$ and $\mathbf{r}_2$.

$J(\mathbf{r}_1, \mathbf{r}_2)$ is a constant for a spatially coherent beam, meaning that all combinations of locations have identical correlation. However, when the beam is only spatially partially coherent, $J(\mathbf{r}_1, \mathbf{r}_2)$ may have a certain distribution which can be revealed by the far-field diffraction pattern.

Considering the split of the incident wave in the aperture plane, we can rewrite the diffraction pattern as

$$\begin{aligned} I(\mathbf{k}) &= \iint \iint T_p(\mathbf{r}_1) T_p(\mathbf{r}_2)^* J(\mathbf{r}_1, \mathbf{r}_2) \exp\left[-i2\pi \mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)\right] \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \\ &= \iint \iint [T(\mathbf{r}_1) + C_p \delta(\mathbf{r}_1 - \mathbf{r}_p)][T(\mathbf{r}_2) + C_p \delta(\mathbf{r}_2 - \mathbf{r}_p)]^* \\ &\quad \times J(\mathbf{r}_1, \mathbf{r}_2) \exp\left[-i2\pi \mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)\right] \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2. \end{aligned} \tag{4.5}$$

Now we arrange Eq. (4.5) by grouping according to the dependence of each term of $I(\mathbf{k})$ on the object wave $T(\mathbf{r})$ and the reference wave $C_p \delta(\mathbf{r} - \mathbf{r}_p)$. As a result, we obtain 4 terms corresponding to the two quadratic terms and the two cross terms in holography as follows:

$$\begin{aligned} I(\mathbf{k}) &= \iint \iint T(\mathbf{r}_1) T(\mathbf{r}_2)^* J(\mathbf{r}_1, \mathbf{r}_2) \exp\left[-i2\pi \mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)\right] \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \\ &+ \iint \iint [C_p \delta(\mathbf{r}_1 - \mathbf{r}_p)][C_p \delta(\mathbf{r}_2 - \mathbf{r}_p)]^* J(\mathbf{r}_1, \mathbf{r}_2) \exp\left[-i2\pi \mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)\right] \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \\ &+ \iint \iint [C_p \delta(\mathbf{r}_2 - \mathbf{r}_p)]^* T(\mathbf{r}_1) J(\mathbf{r}_1, \mathbf{r}_2) \exp\left[-i2\pi \mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)\right] \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \\ &+ \iint \iint [C_p \delta(\mathbf{r}_1 - \mathbf{r}_p)] T(\mathbf{r}_2)^* J(\mathbf{r}_1, \mathbf{r}_2) \exp\left[-i2\pi \mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)\right] \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2. \end{aligned} \tag{4.6}$$

In analogy to holography, the first and the second term (quadratic terms) contain either the object wave or the reference wave, while the third and the fourth term (two cross terms) contain both the object wave and the reference wave. Our goal is to separate and to extract the two cross terms from the diffraction pattern $I(\mathbf{k})$.

4.2.2. Description of the measurement scheme

As can be seen in Eq. (4.6), the first term is the diffraction pattern for the unperturbed aperture plane, which is generated by only the object wave (the original transmitted wave) and hence shall be denoted by $I_0(\mathbf{k})$, while the second term is the diffraction pattern generated by only the reference wave (the light transmitted only at the perturbation point). In the camera plane, the second term, denoted by $I_p(\mathbf{k})$, gives rise to a constant intensity distribution which equals to the intensity of the incident beam at the perturbation point.

We rewrite the cross terms of Eq. (4.6) as:

$$\begin{aligned} I_{c3}(\mathbf{k}) &= \iint \left[\iint [C_p \delta(\mathbf{r}_2 - \mathbf{r}_p)]^* T(\mathbf{r}_1) J(\mathbf{r}_1, \mathbf{r}_2) \exp\left[-i2\pi \mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)\right] \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \right. \\ &= \left. \iint C_p^* T(\mathbf{r}_1) J(\mathbf{r}_1, \mathbf{r}_p) \exp[-i2\pi \mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_p)] \mathrm{d}\mathbf{r}_1, \right. \end{aligned} \tag{4.7}$$

$$\begin{aligned} I_{c4}(\mathbf{k}) &= \iint \left[\iint [C_p \delta(\mathbf{r}_1 - \mathbf{r}_p)] T(\mathbf{r}_2)^* J(\mathbf{r}_1, \mathbf{r}_2) \exp\left[-i2\pi \mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)\right] \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2 \right. \\ &= \left. \iint C_p T(\mathbf{r}_2)^* J(\mathbf{r}_p, \mathbf{r}_2) \exp[-i2\pi \mathbf{k} \cdot (\mathbf{r}_p - \mathbf{r}_2)] \mathrm{d}\mathbf{r}_2. \right. \end{aligned} \tag{4.8}$$

It should be noted that because $J(\mathbf{r}_p, \mathbf{r}) = J(\mathbf{r}, \mathbf{r}_p)^*$ (Hermitian property of the MCF), the third term $I_{c3}(\mathbf{k})$ and the fourth term $I_{c4}(\mathbf{k})$ are complex conjugate of each other and hence carry exactly identical information.

Inverse Fourier transforming the cross terms, we obtain:

$$\begin{aligned} \hat{I}_{c3}(\mathbf{r}') &= \iint \left[\iint C_p^* T(\mathbf{r}_p + \mathbf{r}) J(\mathbf{r}_p + \mathbf{r}, \mathbf{r}_p) \exp(-i2\pi \mathbf{k} \cdot \mathbf{r}) \mathrm{d}\mathbf{r}\right] \exp(+i2\pi \mathbf{k} \cdot \mathbf{r}') \mathrm{d}\mathbf{k} \\ &= C_p^* T(\mathbf{r}_p + \mathbf{r}') J(\mathbf{r}_p + \mathbf{r}', \mathbf{r}_p), \end{aligned} \tag{4.9}$$

$$\begin{aligned} \hat{I}_{c4}(\mathbf{r}') &= \iint \left[\iint C_p T(\mathbf{r}_p - \mathbf{r})^* J(\mathbf{r}_p, \mathbf{r}_p - \mathbf{r}) \exp(-i2\pi \mathbf{k} \cdot \mathbf{r}) \mathrm{d}\mathbf{r}\right] \exp(+i2\pi \mathbf{k} \cdot \mathbf{r}') \mathrm{d}\mathbf{k} \\ &= C_p T(\mathbf{r}_p - \mathbf{r}')^* J(\mathbf{r}_p - \mathbf{r}' \mathbf{r}_p)^*. \end{aligned} \tag{4.10}$$

In analogy to holography, $\hat{I}_{c4}(\mathbf{r})$ is referred to as the "twin image" of $\hat{I}_{c3}(\mathbf{r})$. Both can be obtained by applying different operations to the same function $C_p^* T(\mathbf{r}) J(\mathbf{r}, \mathbf{r}_p)$:

- $\hat{I}_{c3}(\mathbf{r})$:
  - translation from origin to $-\mathbf{r}_p$: $C_p^* T(\mathbf{r}_p + \mathbf{r}) J(\mathbf{r}_p + \mathbf{r}, \mathbf{r}_p)$.
- $\hat{I}_{c4}(\mathbf{r})$:
  - taking complex conjugate: $C_p T(\mathbf{r})^* J(\mathbf{r}, \mathbf{r}_p)^*$
  - translation from origin to $+\mathbf{r}_p$: $C_p T(\mathbf{r} - \mathbf{r}_p)^* J(\mathbf{r} - \mathbf{r}_p, \mathbf{r}_p)^*$.
  - flipping over the center: $C_p T(\mathbf{r}_p - \mathbf{r})^* J(\mathbf{r}_p - \mathbf{r}, \mathbf{r}_p)^*$.

In our method, the translation that depends on the location of the perturbation point $\mathbf{r}_p$ and the finite size of the transmission function of the aperture $T(\mathbf{r})$ are crucial. Together they determine the spatial separation between the cross terms.

We remark that thanks to the linear dependence of the cross term on both the transmission function $T(\mathbf{r})$ and the correlation function $J(\mathbf{r}, \mathbf{r}_p)$, using this experimental setup, we can either measure the MCF of the incident beam or perform diffractive imaging on a transmissive object.

For MCF measurement, $T(\mathbf{r})$ in the aperture plane needs to be provided, and we can retrieve $J(\mathbf{r}, \mathbf{r}_p)$, the correlation between field at the perturbation point $\mathbf{r}_p$

and all points $\mathbf{r}$. $J(\mathbf{r}, \mathbf{r}_p)$ is a 2-dimensional cross-section of the 4-dimensional MCF $J(\mathbf{r}_1, \mathbf{r}_2)$ by setting one variable to be $\mathbf{r}_2 = \mathbf{r}_p$ and the other variable to be $\mathbf{r}_1 = \mathbf{r}$. Therefore, in order to measure the complete $J(\mathbf{r}_1, \mathbf{r}_2)$, we need to move $\mathbf{r}_p$ to all possible locations and retrieve the corresponding $J(\mathbf{r}, \mathbf{r}_p)$ at each location.

For diffractive imaging, we retrieve the product of $T(\mathbf{r})O(\mathbf{r})$, instead of $T(\mathbf{r})$, and the correlation function $J(\mathbf{r}, \mathbf{r}_p)$. $T(\mathbf{r})O(\mathbf{r})$ is a product of the transmission function in the aperture plane and of the transmissive object. In order to obtain $O(\mathbf{r})$ alone, we need to perform two separate measurements: with and without the object, and we divide the former, which yields $J(\mathbf{r}, \mathbf{r}_p)T(\mathbf{r})O(\mathbf{r})$, by the latter, which yields $J(\mathbf{r}, \mathbf{r}_p)T(\mathbf{r})$.

4.2.3. Explanation of the retrieval process

In our method, the size of the aperture determines the size of each term after inverse Fourier transforming the diffraction pattern. Both quadratic terms are centered at the origin. $\hat{I}_0(\mathbf{r})$ occupies an area twice as large as the aperture, and $\hat{I}_p(\mathbf{r})$ occupies an area twice as large as the perturbation point. The two cross terms have the same sizes but locate at different locations. $\hat{I}_{c3}$ and $\hat{I}_{c4}$ are located at $\mathbf{r} = +\mathbf{r}_p$ and $\mathbf{r} = -\mathbf{r}_p$, respectively.

In Fig. 4.2, we neglect the quadratic term $I_p(\mathbf{r})$ whose size is very small, and we illustrate the relation between the quadratic term $I_0(\mathbf{r})$ and the two cross terms. Three overlap situations may occur depending on the locations of the cross terms, which eventually depends on the location of the perturbation point $\mathbf{r}_p$. We denote the domain of the aperture and the quadratic term $I_0(\mathbf{r})$ by Ω and Ω_0, respectively. These three overlap situations are as follows:

a. $\mathbf{r}_p \in \Omega$ & $\mathbf{r}_p \in \Omega_0$:
The two cross terms overlap with each other and with $I_0(\mathbf{r})$;

b. $\mathbf{r}_p \notin \Omega$ & $\mathbf{r}_p \in \Omega_0$:
The two cross terms overlap with $I_0(\mathbf{r})$ but not with each other;

c. $\mathbf{r}_p \notin \Omega$ & $\mathbf{r}_p \notin \Omega_0$:
The two cross terms overlap neither with each other nor with $I_0(\mathbf{r})$.

The approach for retrieving the two cross terms from the diffraction pattern differs in different situations.

In situation (a), we need to measure three diffraction patterns $I_n(\mathbf{k})$ corresponding to three perturbation constants C_n, where $n = 1, 2, 3$, at the point $\mathbf{r} = \mathbf{r}_p$. As a result, we can create a linear system of equations:

$$\begin{aligned} \hat{I}_1(\mathbf{r}) &= \hat{I}_0(\mathbf{r}) + C_1^* \hat{I}_{c3}(\mathbf{r}) + C_1 \hat{I}_{c4}(\mathbf{r}) \\ \hat{I}_2(\mathbf{r}) &= \hat{I}_0(\mathbf{r}) + C_2^* \hat{I}_{c3}(\mathbf{r}) + C_2 \hat{I}_{c4}(\mathbf{r}), \\ \hat{I}_3(\mathbf{r}) &= \hat{I}_0(\mathbf{r}) + C_3^* \hat{I}_{c3}(\mathbf{r}) + C_3 \hat{I}_{c4}(\mathbf{r}) \end{aligned} \tag{4.11}$$

where $\hat{I}_{c3}(\mathbf{r}) = [T(\mathbf{r}_p + \mathbf{r})J(\mathbf{r}_p + \mathbf{r}, \mathbf{r}_p)]$ and $\hat{I}_{c4}(\mathbf{r}) = [T(\mathbf{r}_p - \mathbf{r})J(\mathbf{r}_p - \mathbf{r}, \mathbf{r}_p)]^*$. By

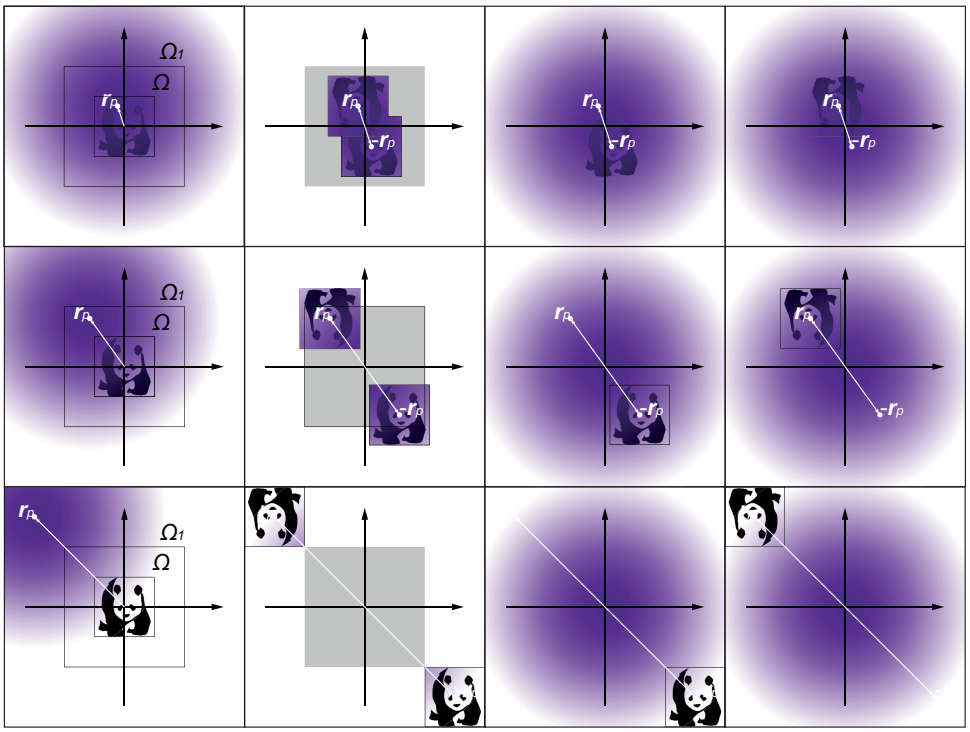

Figure 4.2: Schematic plot of the measurement scheme: The panda represents the object and the surrounding box represents the region of the aperture. The product of the transmission function $T(\mathbf{r})O(\mathbf{r})$ is centered at the origin. The purple distribution represents the correlation function $J(\mathbf{r}, \mathbf{r}_p)$ with respect to the perturbation point. $J(\mathbf{r}, \mathbf{r}_p)$ is centered at the perturbation point where the maximum correlation is $J(\mathbf{r}_P, \mathbf{r}_p)$. For $\hat{I}_{c3}(\mathbf{r})$, $T(\mathbf{r})J(\mathbf{r}, \mathbf{r}_p)$ is translated to become $T(\mathbf{r}_p + \mathbf{r})J(\mathbf{r}_p + \mathbf{r}, \mathbf{r}_p)$, and for $\hat{I}_{c4}(\mathbf{r})$, $T(\mathbf{r})J(\mathbf{r}, \mathbf{r}_p)$ is translated and flipped to become $T(\mathbf{r}_p - \mathbf{r})J(\mathbf{r}_p - \mathbf{r}, \mathbf{r}_p)$.

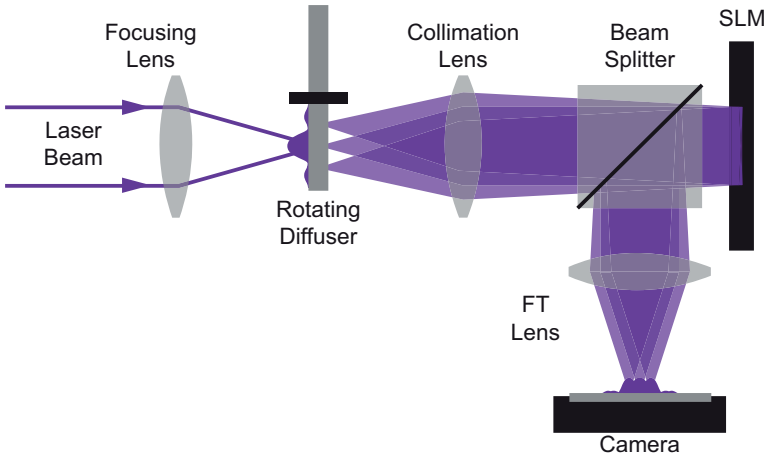

Figure 4.3: Schematic plot of experimental setup. A coherent laser beam is focused onto a rotating diffuser to create an effective incoherent source. The light generated by this source is collimated to illuminate a reflective SLM, on which the effect of superposing a phase object with an aperture is simulated. The far-field intensity distribution of the light reflected by the SLM is measured by the camera using a Fourier transform (FT) lens. By varying the intensity distribution of the focal spot, the MCF of the illumination beam can be varied.

solving the linear system of equations, we obtain

$$\hat{I}_{c3}(\mathbf{r}) = \frac{(C_2 - C_1)^* \left[\hat{I}_3(\mathbf{r}) - \hat{I}_1(\mathbf{r})\right] - (C_3 - C_1)^* \left[\hat{I}_2(\mathbf{r}) - \hat{I}_1(\mathbf{r})\right]}{(C_2 - C_1)^*(C_3 - C_1) - (C_3 - C_1)^*(C_2 - C_1)}, \tag{4.12}$$

and

$$\hat{I}_{c4}(\mathbf{r}) = \frac{(C_2 - C_1) \left[\hat{I}_3(\mathbf{r}) - \hat{I}_1(\mathbf{r})\right] - (C_3 - C_1) \left[\hat{I}_2(\mathbf{r}) - \hat{I}_1(\mathbf{r})\right]}{(C_2 - C_1)(C_3 - C_1)^* - (C_3 - C_1)(C_2 - C_1)^*}. \tag{4.13}$$

In situation (b) we need to measure two diffraction patterns with and without the point perturbation, respectively. By subtracting the later from the former, we can remove the quadratic term $\hat{I}_0(\mathbf{r})$ and obtain the two non-overlapping cross terms. Our approach in situation (b) is similar to the "dOTF" approach in [25].

In situation (c) we can directly obtain the two non-overlapping cross terms because the quadratic term $\hat{I}_0(\mathbf{r})$ does not pose an issue of overlap. Similar approaches, usually known as the Fourier transform holography, can be found in [22–24]. It should be noted that in this situation the total area of the camera sensor was not used efficiently.

4.3. Experimental setup

The schematic plot of experimental setup is shown in Fig. 4.3. We focus a coherent laser beam at a wavelength of 625 nm onto a diffuser. The focused light will then be modulated by a random phase map and generates a speckle pattern. As we rotate the diffuser, a series of speckle patterns (each corresponding to a distinct random

phase map) add together incoherently during the acquisition time. Therefore, we can reduce the degree of coherence of the focused light to zero, and hence create an effective incoherent source, provided that the rotation speed is sufficient fast or the acquisition time is sufficient long.

In the experimental environment, a fast rotation speed causes diffuser vibration while a long acquisition time causes camera saturation. Therefore, finding a balance between these two parameter is the key to a successful experiment. In the setup, we mount the diffuser (a ground glass) on a optical chopper, which rotates at a speed (in frequency) from 20 Hz to 1000 Hz, and we combine images measured at various exposure times to avoid saturation by extending the dynamic range.

Because the light scattered by the diffuser is modulated by a random phase that varies at different times and at different locations on the area illuminated by the focused light, the effective incoherent source can be considered consisting of a collection of independent point sources. The size of and the separation between the point sources are determined by the roughness of the diffuser.

The light generated by the effective incoherent source is collimated to illuminate the SLM. The MCF of the illumination light depends on the intensity distribution of the focused spot (the effective incoherent source). In the experiment, we can obtain two focal spots with two different intensity distributions and hence generate two illumination beams with two different coherence structures: one is the Gaussian correlated beam, also known as Gaussian Schell-model (GSM) beam, and the other is the Gaussian-Airy (GAC) correlated beam.

The difference between these two types of beams is whether the laser beam is truncated by the focusing lens. The intensity distribution of the focal spot, acting as the effective source, is the intensity of the Fourier transform of the field at the focusing lens, which naturally exhibits a Gaussian profile for a normal laser beam. So both the intensity distribution and the coherence structure of the generated beam will exhibit a Gaussian profile without truncation, or a Gaussian-Airy profile with truncation.

The GSM beam can be expressed by

$$J(\mathbf{r}_1, \mathbf{r}_2) = \exp\left(-\frac{\mathbf{r}_1^2 + \mathbf{r}_2^2}{w^2}\right) \exp\left(-\frac{(\mathbf{r}_1 - \mathbf{r}_2)^2}{\sigma^2}\right), \tag{4.14}$$

where w is the width of the Gaussian intensity distribution and σ is the width of the Gaussian coherence structure (degree of coherence). The Gaussian-Airy profile is a result of truncating a Gaussian profile by a circular pupil and then performing an optical Fourier transform. We can interpret the Gaussian-Airy profile by a convolution between the Airy function (the Fourier transform of the circular aperture) and the Gaussian function.

In the experiment, we can also vary the degree of coherence by simply varying the size of the focal spot (while keeping its shape). This can be achieved by translating the focal lens back and forth along the optical axis. In the case when the focal spot size is smaller than the diffraction limit of the collimation lens, we can consider the illumination beam to be spatially coherent. The degree of coherence of the illumination beam decreases as the focal spot size increases.

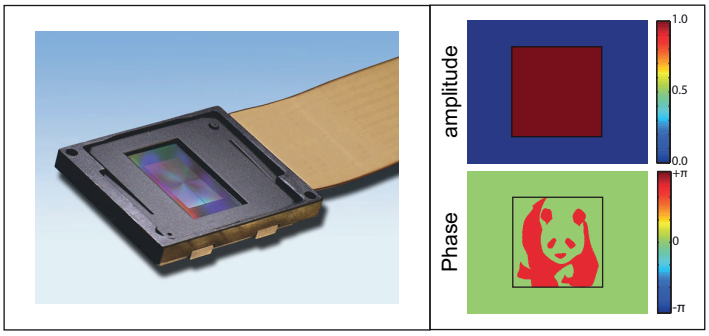

Figure 4.4: Picture of the SLM and the total transmission function of the phase object that is superposed with the aperture. We assign an uniform amplitude and a binary phase in the shape of the panda to the SLM.

In the experiment, we use a phase only reflective SLM (Holoeye PLUTO 2 with size 1980 pixel $\times$ 1080 pixel and resolution 8.0 µm $\times$ 8.0 µm) to simulate the effect of superposing a phase object with an aperture. The aperture is a square with size of 240 pixel $\times$ 240 pixel on the SLM. By using two different phase tilting, we can reflect incident light towards the direction of camera through a beam splitter if it falls inside the aperture or towards another direction if it falls outside the aperture.

The SLM modulates the reflected light according to the phase object, which has a uniform amplitude and a binary phase distribution between 0.1π and 0.9π in the shape of a panda. The resulting phase of the SLM is thus the product of the tilt and the panda. Although here we use a reflection geometry, the result will be also valid for a transmission configuration.

In the experiment, we introduce point perturbation by varying the phase of the incident light in a region of 10 pixel $\times$ 10 pixel on the SLM. The center of this region is given by the location of the perturbation point $\mathbf{r}_p$. When $\mathbf{r}_p$ is located inside the aperture, i.e. in situation (a), we change the phase in the region of perturbation by certain constant values to create a linear system of equations. When $\mathbf{r}_p$ is located outside the aperture in situation (b) and (c), we change the phase tilt in the region of perturbation to simulate the effect of 'opening' (reflected towards the direction of the camera) and 'closing' (reflected towards another direction) a pinhole.

The propagation of the reflected light from the SLM to the camera can be approximated by the Fourier transform. This can be achieved by using the Fraunhofer propagation in free-space, which is particularly useful for short wavelengths, e.g. in the X-ray and the electron regime, where no lens is available. Alternatively this can be achieved by using a Fourier transform (FT) lens with a focal length of 100 mm as we did here. For this purpose, the SLM and the camera should be placed in the front and the back focal plane of the FT lens, respectively.

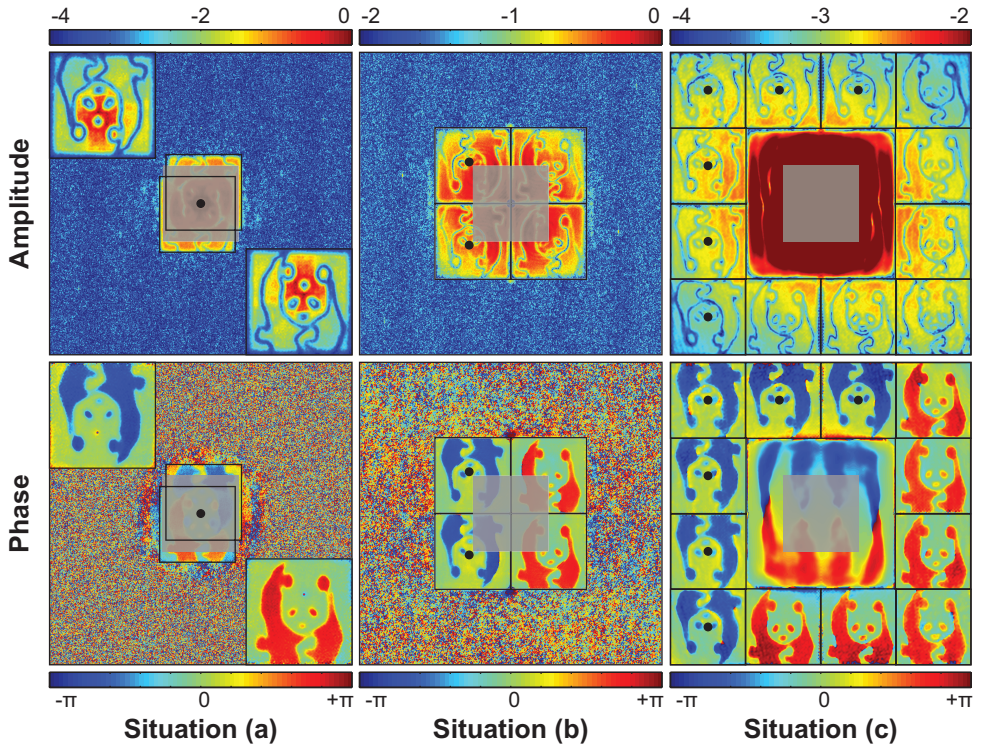

Figure 4.5: Experimental results of varying the perturbation point location for the GSM beam illumination. The gray square shows the aperture and the black dots show the perturbations points. In the phase plot, the red and the blue panda are the phase of the original object and its complex-conjugate, respectively. The panda contour in the amplitude plot is due to abrupt phase transition. The amplitude plot is in logarithmic scale. The quadratic term $\hat{I}_0(\mathbf{r})$ has been removed in (a) and (b). (c) is the direct inverse Fourier transform of the perturbed diffraction pattern.

4.4. Experimental Results

In the experiment, we first focus on using the GSM beam for illumination. In this case, we validate our measurement scheme for three overlap situations and the retrieval of the correlation function for various degrees of coherence. We then use the GAC beam for illumination. Due to the modulation effect of the correlation function, which shows a "oscillating" behavior in both amplitude and phase, we need to perform two measurements with and without the object, respectively. Here we aim to verify that dividing the former by the latter yields the object alone.

4.4.1. Results of varying the perturbation point location for GSM beam illumination

Now we illuminate the phase object that is superposed with the aperture using a GSM beam. The correlation function of the GSM beam has a Gaussian amplitude and an uniform phase, while the object has an uniform amplitude and a phase in

the shape of a panda. Therefore, the retrieved product of the correlation function $J(\mathbf{r}, \mathbf{r}_p)$ and the transmission function $T(\mathbf{r})O(\mathbf{r})$ shows the Gaussian amplitude and the panda phase as illustrated in Fig. 4.5.

The panda shape in the amplitude is due to the fact that at the places of abrupt phase transition, the amplitude is zero. So the amplitude will have a panda contour consisting of zeros. Because the diffraction pattern has limited size and hence its inverse Fourier transform has limited resolution,this invisible contour (with infinitely small width) becomes visible (with enlarged finite width). In the phase plot in Fig. 4.5, the red and the blue panda represent the phase of the original object and its complex-conjugate, respectively.

In Fig. 4.5 we demonstrate experimentally that we can effectively vary the overlap between the quadratic term $\hat{I}_0(\mathbf{r})$ and the two cross terms by varying the location $\mathbf{r}_p$ of the perturbation point. The further the perturbation point is from the origin, the more the two cross terms are separated. Contrarily, $\hat{I}_0(\mathbf{r})$ always stays centered at the origin.

Fig. 4.5 also demonstrates that we vary the correlation function $J(\mathbf{r}, \mathbf{r}_p)$ when varying the location $\mathbf{r}_p$ of the perturbation point. This has two effects: (1) we vary the shape of the correlation function $J(\mathbf{r}, \mathbf{r}_p)$ and (2) we vary the part of $J(\mathbf{r}, \mathbf{r}_p)$ that is measured.

Notice that the MCF of a GSM beam is translation-invariant which depends on only the relative distance $\mathbf{r} - \mathbf{r}_p$: $J(\mathbf{r}, \mathbf{r}_p) = J(\mathbf{r} - \mathbf{r}_p)$. So the shape of $J(\mathbf{r} - \mathbf{r}_p)$ always keeps the same as $J(\mathbf{r})$, except for a translation of its maximum from the origin to $\mathbf{r}_p$. Meanwhile, placing the perturbation point at different locations allows us to measure different parts of $J(\mathbf{r})$. The part of $J(\mathbf{r})$ that can be measured is defined by the domain of the aperture $T(\mathbf{r})$, which always stays at the origin.

By inverse Fourier transforming the diffraction pattern, we observe in Fig. 4.5 that the maximum of the correlation function, $J(\mathbf{r}_p - \mathbf{r}, \mathbf{r}_p)$ and $J(\mathbf{r}_p + \mathbf{r}, \mathbf{r}_p)$, are both at the origin, while the product of aperture and object, $T(\mathbf{r}_p - \mathbf{r})O(\mathbf{r}_p - \mathbf{r})$ and $T(\mathbf{r}_p + \mathbf{r})O(\mathbf{r}_p + \mathbf{r})$, are located at $+\mathbf{r}_p$ and $-\mathbf{r}_p$, respectively.

Finally, we can introduce more than one perturbation point in each situation, as long as there is no overlap between the two cross terms generated by each perturbation point. The larger the number of perturbation points we introduce, the more parts of the correlation function we can measure simultaneously.

4.4.2. Results of varying the degree of coherence for GSM beam illumination

We compare the results of using the GSM beam for two degrees of coherence in Fig. 4.6. In the experiment, we vary the degree of coherence of the GSM beam by varying the size of the focal spot while keeping the Gaussian shape. The perturbation point is placed at the two left corners of the aperture and as a result, the two pairs of cross terms are connected but are not overlapped. Because the correlation function is translation-invariant, we can easily observe its Gaussian profile.

As can be seen in Fig. 4.6, the blurring of the diffraction pattern increases as the degree of coherence of the GSM beam decreases. Because the diffraction pattern is the incoherent sum of the shifted coherent diffraction pattern generated by each

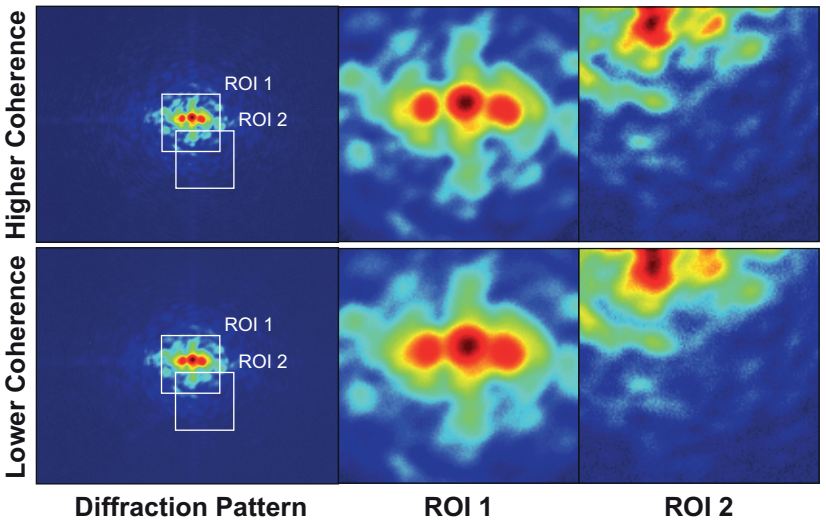

Figure 4.6: The diffraction patterns for GSM beam illumination with two degrees of coherence. Two regions of interest (ROI) marked by two square boxes are illustrated in detail.

point source, the coherent diffraction pattern is smeared due to the use of partially coherent illumination. Because the shift depends on the location of the point source, the blurring of the diffraction pattern is proportional to the size of the source.

Conventionally, the blurring is associated with information loss. It is believed that the more blurred the diffraction pattern is, the more information of the object is lost, especially for the iterative algorithm such as in [20, 21]. However, the results illustrated in Fig. 4.7 show that the blurring influences only the field-of-view (FOV) instead of the resolution.

We remark that we can only reconstruct the correlation function $J(\mathbf{r}, \mathbf{r}_p)$and the total transmission function $T(\mathbf{r})O(\mathbf{r})$ at places where the amplitude of the inverse Fourier transform of the diffraction pattern is not corrupted by the noise as shown in Fig. 4.7. For GSM beam illumination, the size of the FOV is determined by both the noise and the degree of spatial coherence σ defined in Eq. 4.14. For a given noise level, the higher the σ is, the larger the FOV is.

We validate this by fitting the retrieved amplitude, which is contributed by only the correlation function $J(\mathbf{r}, \mathbf{r}_p)$, to a Gaussian distribution to determine the value of σ (the points consisting of the panda contour are neglected). The fitting shows that in the aperture plane, σ is 0.68 and 0.57 mm for the case of higher and lower coherence, respectively.

By comparing the degree of coherence σ to the size of aperture, which is about 2.72 mm along the diagonal, we validate that our method can tolerate partially coherent illumination that conventional iterative algorithms such as [20, 21] cannot. To handle the reduction of the FOV, we can either place the perturbation point in the vicinity of the origin, or place the perturbation point at different locations to reconstruct different parts of the FOV.

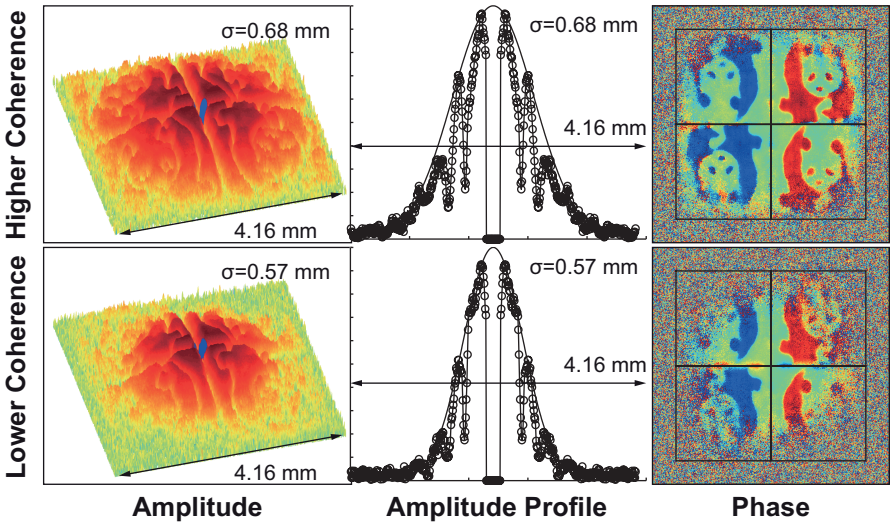

Figure 4.7: The diffraction patterns and the inverse Fourier transform for GSM beam illumination with two degrees of coherence. The quadratic term has been removed. Dots and lines in the amplitude profile plot show the raw data and the fitted curve, respectively.

4.4.3. Results of varying the degree of coherence for GAC beam illumination

Now we consider using a GAC beam for illumination. The correlation function of the GAC beam has a complicated phase, instead of a simple uniform phase that the GSM beam has. Because we only retrieve the sum of the phases of the object and the correlation function, we must calibrate and compensate the correlation function.

We illustrate the diffraction pattern for two degrees of coherence in the case with and without the object in Fig. 4.8 and 4.9, respectively. We observe that just like the GSM beam illumination, the diffraction pattern becomes more blurred as the illumination beam becomes less coherent.

In Fig. 4.8, the amplitude plot is still contributed by only the correlation function except for the panda contour. The amplitude consists of a number of rings and hence shows an "oscillating" behavior. Most importantly, the oscillation is faster (more number of rings) when the coherence is lower (less degree of coherence). Meanwhile, when the coherence is lower we also observe a faster decrease of the correlation versus the increase of distance.

The phase plot in Fig. 4.8 is the sum of the phases of the correlation function and the object. As a result, the panda shape in the phase of the former, although still visible, is obscured by the spiral phase of latter. We also observe a 2π phase jump at places where the corresponding amplitude vanishes. This phenomenon suggests that the correlation function switches signs in between rings.

The obscuration of the object phase by the phase of the correlation function can be calibrated and compensated by performing an extra measurement using only the

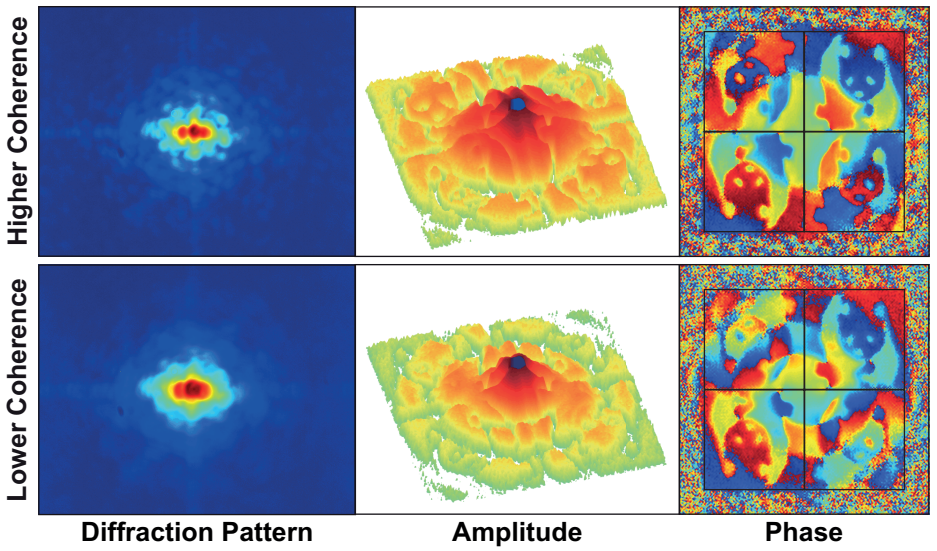

Figure 4.8: Diffraction pattern and the inverse Fourier transform for GAC beam illumination with two degrees of coherence. The transmissive object is superposed with the aperture.

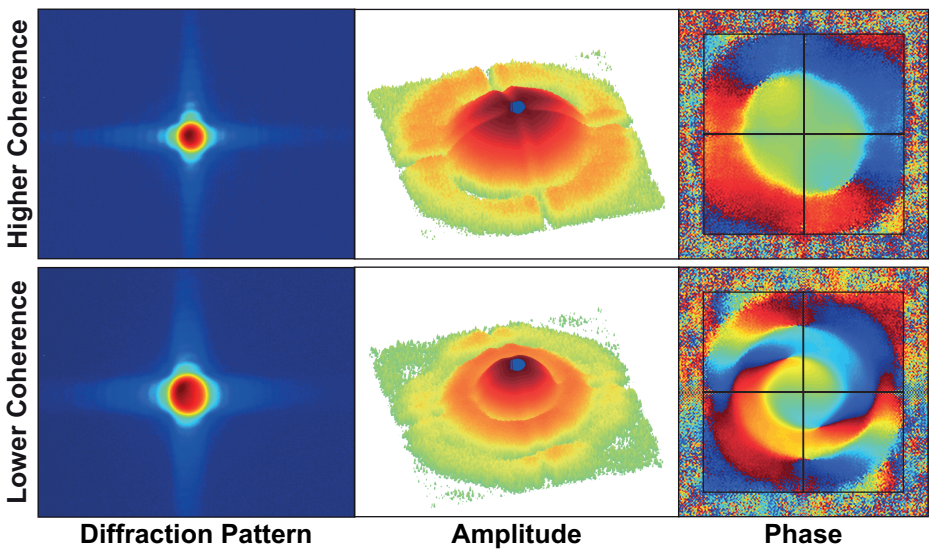

Figure 4.9: Diffraction pattern and the inverse Fourier transform for GAC beam illumination with two degrees of coherence. The transmissive object is not superposed with the aperture.

aperture. In both measurements, we need to place the two perturbation points at the same locations so that we measure the same correlation function $J(\mathbf{r}, \mathbf{r}_p)$. As can be seen in Fig. 4.8 and 4.9, the spirals in the phase match each other.

In the two cases with and without the object, we can retrieve $J(\mathbf{r}, \mathbf{r}_p)T(\mathbf{r})O(\mathbf{r})$ (with object) and $J(\mathbf{r}, \mathbf{r}_p)T(\mathbf{r})$ (without object), respectively. Dividing the former by the later allows us to obtain the object $O(\mathbf{r})$ alone. Fig. 4.10 shows that the resulting $O(\mathbf{r})$ is almost perfect except for the places of 2π phase jump.

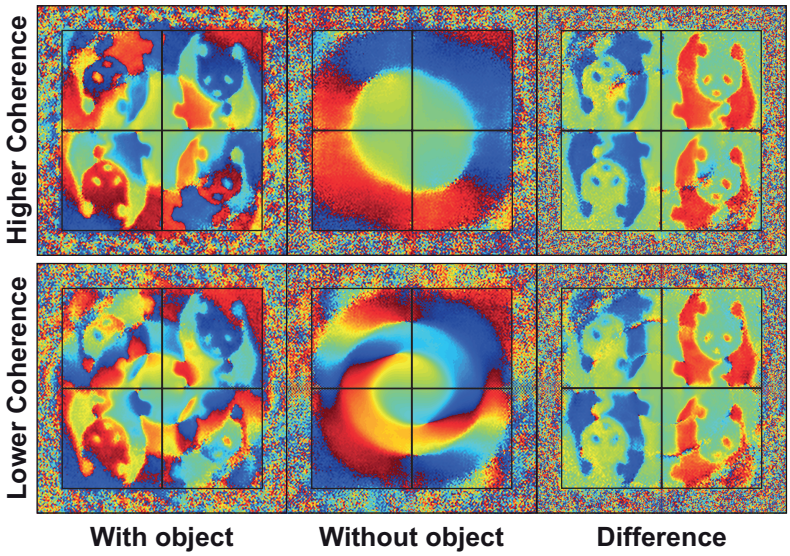

Figure 4.10: The phase of the inverse Fourier transform of the diffraction pattern with object, without object, and their difference. In each plot, the right two pandas are the phase of the original object and the left two pandas are the phase of its complex-conjugate.

4.5. Discussion and Conclusion

In this chapter, we demonstrated that we can retrieve the correlation function $J(\mathbf{r}, \mathbf{r}_p)$ between fields at the perturbation point $\mathbf{r}_p$ and other locations $\mathbf{r}$, and the product of the transmission function of the aperture $T(\mathbf{r})$ and the object $O(\mathbf{r})$.

The key is the extraction of the two cross terms from the diffraction pattern. The process is similar to holography, in which the two cross terms must be separated from the two quadratic terms.

An essential feature that distinguishes our method from holography is that our reference wave is created by perturbing the transmission/reflection function in the aperture plane at a point and hence is correlated with the object wave that is transmitted/reflected by the aperture. This feature cannot be preserved when using a beam splitter to separate the reference wave and the object wave.

Conventionally, it is believed that spatially coherent illumination is required to image a phase object because the phase information will be destroyed by the inco-

herent sum of the randomly fluctuating light fields. We demonstrated that for our method the spatial coherence only affects the FOV instead of the resolution.

In the experimental results, the retrieved product of the transmission function $T(\mathbf{r})O(\mathbf{r})$ is modulated by the correlation function $J(\mathbf{r}, \mathbf{r}_p)$. The size of the FOV is determined by the amplitude of the modulation and the noise level.

The resolution is determined by the size of the perturbed region. In order to use a sufficiently small perturbed region while making the energy of the perturbation being sufficiently high, we need to balance the energies of the light incident on the aperture and the perturbed region.

The signal-to-noise ratio of the diffraction pattern is basically limited by the dynamic range of the camera sensor. Most of the object/aperture generates a diffraction pattern with a strong main peak and weak side-lobes. To reduce the ratio between the mean peak and side-lobes, we can modulate the wavefront of the incident light by using for example a diffuser, so that the diffraction pattern can be measured properly using a limited number of grey levels.

In this chapter, we restrict our research to only monochromatic light. If chromatic light is considered, the scaling of the diffraction pattern must be taken into account. For Fresnel or Fraunhofer diffraction, the size of the diffraction pattern is proportional to the reciprocal of the wavelength. Because the camera sensor has a fixed size, the inverse Fourier transform of the diffraction pattern will have a different resolution at different wavelength. When using optical elements, chromatic aberration also plays a role.

References

[1] Y. Shao, X. Lu, S. Konijnenberg, C. Zhao, Y. Cai, and H. P. Urbach, *Spatial coherence measurement and partially coherent diffractive imaging using self-referencing holography,* Optics express **26**, 4479 (2018).

[2] Y. Cai, Y. Chen, and F. Wang, *Generation and propagation of partially coherent beams with nonconventional correlation functions: a review,* JOSA A **31**, 2083 (2014).

[3] Y. Chen, F. Wang, L. Liu, C. Zhao, Y. Cai, and O. Korotkova, *Generation and propagation of a partially coherent vector beam with special correlation functions,* Physical Review A **89**, 013801 (2014).

[4] G. Gbur, *Partially coherent beam propagation in atmospheric turbulence,* JOSA A **31**, 2038 (2014).

[5] K. Lai, A. E. Rosenbluth, S. Bagheri, J. Hoffnagle, K. Tian, D. Melville, J. Tirapu-Azpiroz, M. Fakhry, Y. Kim, S. Halle, *et al.*, *Experimental result and simulation analysis for the use of pixelated illumination from source mask optimization for 22nm logic lithography process,* in *Optical Microlithography XXII*, Vol. 7274 (International Society for Optics and Photonics, 2009) p. 72740A.

[6] K. M. Douglass, C. Sieben, A. Archetti, A. Lambert, and S. Manley, *Super-resolution imaging of multiple cells by optimized flat-field epi-illumination,* Nature photonics **10**, 705 (2016).

[7] H. Partanen, J. Turunen, and J. Tervo, *Coherence measurement with digital micromirror device,* Optics letters **39**, 1034 (2014).

[8] F. Pfeiffer, O. Bunk, C. Schulze-Briese, A. Diaz, T. Weitkamp, C. David, J. Van Der Veen, I. Vartanyants, and I. Robinson, *Shearing interferometer for quantifying the coherence of hard x-ray beams,* Physical review letters **94**, 164801 (2005).

[9] S. Divitt and L. Novotny, *Spatial coherence of sunlight and its implications for light management in photovoltaics,* Optica **2**, 95 (2015).

[10] D. Morrill, D. Li, and D. Pacifici, *Measuring subwavelength spatial coherence with plasmonic interferometry,* Nature photonics **10**, 681 (2016).

[11] S. Marathe, X. Shi, M. J. Wojcik, N. G. Kujala, R. Divan, D. C. Mancini, A. T. Macrander, and L. Assoufid, *Probing transverse coherence of x-ray beam with 2-d phase grating interferometer,* Optics express **22**, 14041 (2014).

[12] X. Shi, S. Marathe, M. J. Wojcik, N. G. Kujala, A. T. Macrander, and L. Assoufid, *Circular grating interferometer for mapping transverse coherence area of x-ray beams,* Applied Physics Letters **105**, 041116 (2014).

[13] X. Liu, F. Wang, L. Liu, Y. Chen, Y. Cai, and S. A. Ponomarenko, *Complex degree of coherence measurement for classical statistical fields,* Optics letters **42**, 77 (2017).

[14] C. Tran, G. Williams, A. Roberts, S. Flewett, A. Peele, D. Paterson, M. de Jonge, and K. Nugent, *Experimental measurement of the four-dimensional coherence function for an undulator x-ray source,* Physical review letters **98**, 224801 (2007).

[15] L. Waller, G. Situ, and J. W. Fleischer, *Phase-space measurement and coherence synthesis of optical beams,* Nature Photonics **6**, 474 (2012).

[16] J. K. Wood, K. A. Sharma, S. Cho, T. G. Brown, and M. A. Alonso, *Using shadows to measure spatial coherence,* Optics letters **39**, 4927 (2014).

[17] K. A. Sharma, T. G. Brown, and M. A. Alonso, *Phase-space approach to lensless measurements of optical field correlations,* Optics express **24**, 16099 (2016).

[18] J. Clark, X. Huang, R. Harder, and I. Robinson, *High-resolution three-dimensional partially coherent diffraction imaging,* Nature communications **3**, 993 (2012).

[19] N. Burdet, X. Shi, D. Parks, J. N. Clark, X. Huang, S. D. Kevan, and I. K. Robinson, *Evaluation of partial coherence correction in x-ray ptychography,* Optics express **23**, 5452 (2015).

[20] L. Whitehead, G. Williams, H. Quiney, D. Vine, R. Dilanian, S. Flewett, K. Nugent, A. G. Peele, E. Balaur, and I. McNulty, *Diffractive imaging using partially coherent x rays,* Physical review letters **103**, 243902 (2009).

[21] P. Thibault and A. Menzel, *Reconstructing state mixtures from diffraction measurements,* Nature **494**, 68 (2013).

[22] I. McNulty, J. Kirz, C. Jacobsen, E. H. Anderson, M. R. Howells, and D. P. Kern, *High-resolution imaging by fourier transform x-ray holography,* Science **256**, 1009 (1992).

[23] S. Eisebitt, J. Lüning, W. Schlotter, M. Lörgen, O. Hellwig, W. Eberhardt, and J. Stöhr, *Lensless imaging of magnetic nanostructures by x-ray spectroholography,* Nature **432**, 885 (2004).

[24] L.-M. Stadler, C. Gutt, T. Autenrieth, O. Leupold, S. Rehbein, Y. Chushkin, and G. Grübel, *Hard x ray holographic diffraction imaging,* Physical review letters **100**, 245503 (2008).

[25] J. L. Codona, *Differential optical transfer function wavefront sensing,* Optical Engineering **52**, 097105 (2013).

5

Spatially partially coherent diffractive imaging using pinhole array mask

Parts of this chapter have been published in Advanced Photonics **1.1** (2019): 016005 [1].

5.1. Background

Coherent diffractive imaging (CDI) is an important tool for the reconstruction of the complex-valued transmission/reflection function of an object from the far-field diffraction patterns. CDI has been widely applied in material and biological sciences [2, 3]. In 1991, Miao et al. first experimentally realized imaging of a sub-micrometer sized non-crystalline specimen using CDI [4].

Many CDI approaches have been developed in the past decades, which can be divided into two types: the iterative methods [5–10] and the non-iterative methods [11–15]. These CDI approaches all require completely coherent illumination, and hence have limited applications at short wavelengths, e.g. in the X-ray and electron regime, or in unstable experimental environments. For example, the degradation of spatial coherence may be caused by the disturbance due to the mechanical vibration or by the fluctuation of the ambient medium [16, 17].

Iterative algorithms retrieve the phase of the object by propagating the field back and forth between the object plane and the far-field diffraction plane, and imposing constraints on the field in both planes. Gerchberg and Saxton pioneered the iterative algorithms in 1972 by proposing a method using two intensities measured in the object plane and in the far-field respectively [5]. Iterative algorithms using only one intensity measurement of the far-field diffraction pattern were proposed by Fienup [6, 7], which require prior knowledge e.g. the object support.

Recently, ptychographic algorithms have become an essential technique for imaging nano-scale objects using short wavelength sources [9]. Ptychographic algorithms scan the illumination probe over the sample and take a measurement at each scanning position. The key feature is the overlap between the illuminated areas at the neighboring scanning positions. The overlap improves the convergence of the ptychographic algorithms [8].

For spatially partially coherent (SPC) illumination, the propagation of light is described using the mutual coherence function (MCF) instead of the field. The first modification of the iterative algorithm was reported by Whitehead, et al.[16]. Later Thibault, et al. demonstrated that the ptychographic algorithms can also be modified to work for SPC illumination [17]. In both works, the MCF is decomposed by a weighted sum of coherent modes. The accuracy of mode decomposition relies on the number of modes for accurately representing the MCF. The number increases as the spatial coherence of the illumination decreases.

Compared to iterative methods, non-iterative methods [11, 12] do not suffer from issues such as stagnation or non-uniqueness of the solution to the problem of phase retrieval [18, 19]. For example, in holography, the field transmitted by the object is perturbed such that the object's transmission/reflection function can be directly extracted from the inverse Fourier transform of the diffraction pattern. This perturbation can be achieved by introducing a pinhole e.g. in the Fourier transform holography (FTH) [20–22] or by changing the transmission/reflection function at a particular point of the object, e.g. with so-called the Zernike quantitative phase imaging [23]. Alternative methods extract the object information from the autocorrelation, which is obtained by inverse Fourier transforming a three-dimensional data set (e.g. the data set measured by varying focus [13] or any other optical

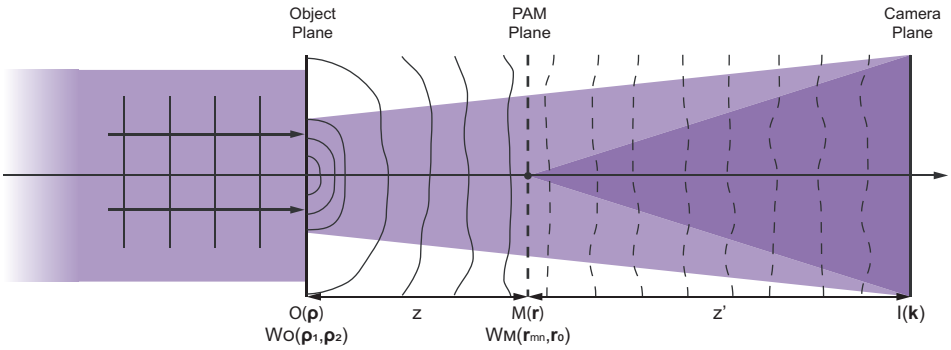

Figure 5.1: Schematic plot of the conceptual experimental setup. A pinhole array mask (PAM) is placed in between the object and the camera. The PAM is specially designed, consisting of a periodic array of measurement pinholes and a reference pinhole. The camera measures the interference pattern between the light transmitted by the measurement pinholes and by the reference pinhole. The design allows us to retrieve the correlation function of the incident light between fields at the measurement pinholes and at the reference pinhole. The object can be reconstructed by reversely propagating the reconstructed correlation function from the PAM to the object.

parameter [14]).

The above mentioned non-iterative methods [13–15] can be used when the illumination is spatially partially coherent. Compared to iterative methods, using non-iterative methods can avoid errors due to the truncation of the number of modes for representing the MCF. However, for non-iterative methods, the field-of-view (FOV) of the reconstructed object is limited by the degree of the spatial coherence of the illumination.

To be precise, what is reconstructed is the product of the transmission function of the object and the correlation function of the illumination with respect to the perturbation point, which attenuates more rapidly as a function of the object coordinate if the degree of spatial coherence is lower.

In FTH [15], the inverse Fourier transform of the diffraction pattern consists of four terms. To reconstruct the object using only one measurement, the location of the perturbation point should be sufficiently far from the object to ensure a spatial separation between the two cross terms and the two quadratic terms are separated. This results in a rather small FOV because the correlation of the fields at the object and at the perturbation point is low.

5.2. Introduction to the method

A schematic plot of the conceptual experimental setup for our method is shown in Fig. 5.1. In this chapter, we consider diffraction imaging in transmission mode. We illuminate a transmissive object by SPC light and measure the far-field diffraction pattern using a camera sensor. Unlike traditional configuration for diffractive imaging, we insert a pinhole array mask (PAM) in between the object plane and the camera plane. Its location is chosen such that the propagation from the object to the PAM and from the PAM to camera obey the approximation of Fresnel and

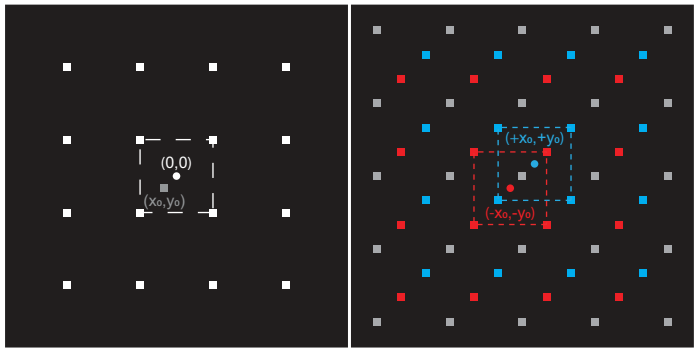

Figure 5.2: Layout of the PAM mask (left) and the inverse Fourier transform of the diffraction pattern (right). Left: The periodic array of measurement pinholes (white) and the reference pinhole (gray). Right: The quadratic term (gray), and the two cross terms (red and blue).

Fraunhofer propagation, respectively.

As can be seen in Fig. 5.1, the coordinates of the object, PAM, and camera plane are $\boldsymbol{\rho}$, $\mathbf{r}$, and $\mathbf{k}$, respectively. Our method consists of two steps:

1. We adopt a special design of the PAM which consists of a periodic array of measurement pinholes and a reference pinhole. In the PAM plane we retrieve the correlation function of the incident light $W_M(\mathbf{r}_{mn}, \mathbf{r}_0)$ between the fields transmitted by the measurement pinholes at $\mathbf{r}_{mn}$ and by the reference pinhole at $\mathbf{r}_0$ from the diffraction pattern $I_k(\mathbf{k})$ measured in the camera plane.

2. We use a differential method, which requires two diffraction patterns with and without perturbation to the transmission function $O(\mathbf{r})$ of the object at a particular point $\boldsymbol{\rho}_0$. In the object plane we reconstruct the product of $O(\boldsymbol{\rho})$ and the correlation function $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_0)$ between the fields at any location $\boldsymbol{\rho}$ and at the location of the point perturbation $\boldsymbol{\rho}_0$ from $W_M(\mathbf{r}_m, \mathbf{r}_0)$ retrieved in the PAM plane.

We remark that the sampling in the PAM plane is determined by $\mathbf{r}_{mn}$, for which the interval and the range of the sampling are given by the pitch and the size of the PAM, respectively. Due to the Fresnel propagation, the sampling interval in the object plane is decreased compared to the sampling interval in the PAM plane. The propagation distance z determines the decrease of the sampling interval.

5.2.1. Step 1: Retrieval of the MCF in the PAM Plane

The specially designed PAM is illustrated in Fig. 5.2, in which the gray square and the white squares represent the reference pinhole and the measurement pinholes, respectively. We can write the transmission function of the PAM as

$$M(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_0) + \sum_{mn} \delta(\mathbf{r} - \mathbf{r}_{mn}), \tag{5.1}$$

where $\mathbf{r}_0 = (x_0, y_0)$ is the location of the reference pinhole and $\mathbf{r}_{mn} = (mp_x, np_y)$ is location of the measurement pinhole in the periodic array, where m, n are the indices of the measurement pinhole and p_x, p_y are the pitches of the 2-dimensional periodic array. The width of the reference pinhole and the measurement pinholes are identical and are given by w_x, w_y.

The incident light is transmitted by the PAM and then generates a diffraction pattern in the camera plane. We denote the MCF of the light in the PMA plane by $W_M(\mathbf{r}_1, \mathbf{r}_2)$, which describes the correlation between the fields at $\mathbf{r}_1$ and $\mathbf{r}_2$. Because the propagation from the PMA to the camera satisfies the condition of Fraunhofer propagation, we can express the diffraction pattern as:

$$I(\mathbf{k}) = \iint \iint W_M(\mathbf{r}_1, \mathbf{r}_2) M(\mathbf{r}_1) M(\mathbf{r}_2)^* \exp[-i2\pi\mathbf{k} \cdot (\mathbf{r}_1 - \mathbf{r}_2)] \mathrm{d}\mathbf{r}_1 \mathrm{d}\mathbf{r}_2. \tag{5.2}$$

Eq. (5.2) consists of four terms:

$$\begin{aligned} I(\mathbf{k}) = {} & W_M(\mathbf{r}_0, \mathbf{r}_0) \\ & + \sum_{m_1 n_1} \sum_{m_2 n_2} W_M(\mathbf{r}_{m_1 n_1}, \mathbf{r}_{m_2 n_2}) \exp[-i2\pi\mathbf{k} \cdot (\mathbf{r}_{m_1 n_1} - \mathbf{r}_{m_2 n_2})] \\ & + \sum_{mn} W_M(\mathbf{r}_{mn}, \mathbf{r}_0) \exp[-i2\pi\mathbf{k} \cdot (\mathbf{r}_{mn} - \mathbf{r}_0)] \\ & + \sum_{mn} W_M(\mathbf{r}_0, \mathbf{r}_{mn}) \exp[+i2\pi\mathbf{k} \cdot (\mathbf{r}_{mn} - \mathbf{r}_0)]. \end{aligned} \tag{5.3}$$

Please note that when deriving Eq. (5.3) we have used the property of the Dirac delta function. As a result, the continuous integration becomes a discrete sum. In Eq. (5.3), $W_M(\mathbf{r}_0, \mathbf{r}_0)$ represents a constant intensity distribution, which will be neglected in the following derivations.

By inverse Fourier transforming the diffraction pattern Eq. (5.3), we obtain

$$\begin{aligned} \mathcal{F}^{-1}[I(\mathbf{k})](\mathbf{r}) = {} & \sum_{m_1 n_1} \sum_{m_2 n_2} W_M(\mathbf{r}_{m_1 n_1}, \mathbf{r}_{m_2 n_2}) \delta[\mathbf{r} - (\mathbf{r}_{m_1 n_1} - \mathbf{r}_{m_2 n_2})] \\ & + \sum_{mn} W_M(\mathbf{r}_{mn}, \mathbf{r}_0) \delta[\mathbf{r} - (\mathbf{r}_{mn} - \mathbf{r}_0)] \\ & + \sum_{mn} W_M(\mathbf{r}_0, \mathbf{r}_{mn}) \delta[\mathbf{r} + (\mathbf{r}_{mn} - \mathbf{r}_0)], \end{aligned} \tag{5.4}$$

where F^{-1} denotes the operation of inverse Fourier transform. The three terms in Eq. (5.4) are:

1. $W_M(\mathbf{r}_{m_1 n_1}, \mathbf{r}_{m_2 n_2})$ locates on the periodic array defined by $\mathbf{r} = \mathbf{r}_{m_1 n_1} - \mathbf{r}_{m_2 n_2} = [(m_1 - m_2)p_x, (n_1 - n_2)p_y]$ with pitch (p_x, p_y) and centered at the origin $(0, 0)$, shown by the gray squares in the right panel in Fig. 5.2.

2. $W_M(\mathbf{r}_{mn}, \mathbf{r}_0)$ are points located on the periodic array defined by $\mathbf{r} = \mathbf{r}_{mn} - \mathbf{r}_0 = [mp_x - x_0, np_y - y_0]$ with pitch (p_x, p_y) and centered at $(-x_0, -y_0)$, shown by the blue squares in the right panel of Fig. 5.2.

3. $W_M(\mathbf{r}_0, \mathbf{r}_{m,n})$ are points located on the periodic array defined by $\mathbf{r} = \mathbf{r}_0 - \mathbf{r}_{mn} = [x_0 - mp_x, y_0 - np_y]$ with pitch (p_x, p_y) and centered at $(+x_0, +y_0)$, shown by the red squares in the right panel of Fig. 5.2.

The role played by the reference pinhole of the PAM is in analogy to that by the point perturbation in FTH [15], namely to create an interference between the light transmitted by the reference pinhole and by the measurement pinholes.

In analogy to holography, the first term of Eq. (5.4), which represents the auto-correlation of the measurement pinholes, is referred to as the quadratic term, while the other two terms are called the cross terms, which are due to the interference of the light transmitted by the reference pinhole and the measurement pinholes.

We illustrate the three terms of Eq. (5.4) in Fig. 5.2, which represent three periodic arrays with the same pitch but different center location. So the three terms are spatially separated. The condition for spatial separation requires the pitch (p_x, p_y) of the PAM to be at least three times larger than the width (w_x, w_y) of the pinhole:

$$\begin{cases} p_x \geq 3w_x \\ p_y \geq 3w_y \end{cases}. \tag{5.5}$$

In the case when the three terms are just separated spatially, the reference pinhole should locate at the center of the periodic array of the measurement pinholes:

$$\begin{cases} x_0 = \dfrac{p_x}{2} \\ y_0 = \dfrac{p_y}{2} \end{cases}. \tag{5.6}$$

The information about the correlation function of the incident light is carried by the two cross terms. Due to the Hermitian property of the MCF, the two cross terms are the complex-conjugate of each other. In Fig. 5.2, the two cross terms are illustrated by two periodic arrays located symmetrically about the origin. Take the cross term $W_M(\mathbf{r}_{m,n}, \mathbf{r}_0)$ as an example. We can retrieve $W_M(\mathbf{r}_{m,n}, \mathbf{r}_0)$ by applying spatial filtering to the inverse Fourier transform of the diffraction pattern $\mathcal{F}^{-1}[I(\mathbf{k})](\mathbf{r})$ given by Eq. (5.4). The spatial filter is given by

$$F_M(\mathbf{r}) = \sum_{mn} \delta(\mathbf{r} - \mathbf{r}_{mn} + \mathbf{r}_0). \tag{5.7}$$

The other cross term $W_M(\mathbf{r}_0, \mathbf{r}_{m,n}) = W_M(\mathbf{r}_{m,n}, \mathbf{r}_0)^*$ can be retrieved by using $F_M(-\mathbf{r})$ as the spatial filter. We remark that $W_M(\mathbf{r}_{m,n}, \mathbf{r}_0)$ and $W_M(\mathbf{r}_0, \mathbf{r}_{m,n})$ contain exactly identical information.

As a result, we retrieve the correlation function of the incident light $W_M(\mathbf{r}_{mn}, \mathbf{r}_0)$ in the PAM plane between the fields transmitted by the measurement pinholes at $\mathbf{r}_{mn}$ and by the reference pinhole at $\mathbf{r}_0$. Notice that $W(\mathbf{r}_{mn}, \mathbf{r}_0)$ is sampled by the

measurement pinholes. Therefore, the interval and the range of the sampling are given by the pitch and the size of the periodic array of the measurement pinhole, respectively.

We should not be confused about the sampling of $\mathbf{r}_{mn}$ and $\mathbf{r}$. According to the Shannon-Nyquist sampling theorem, the sampling of $\mathbf{r}$ is determined by the diffraction pattern and hence is ultimately determined by the camera. However, the sampling interval of $\mathbf{r}_{mn}$ is at least three times larger than the sampling interval of $\mathbf{r}$. Because the pinhole size cannot be smaller than the sampling interval of $\mathbf{r}$. In the extreme case when the pinhole width is equal to the sampling interval of $\mathbf{r}$, all three periodic arrays are connected but not overlapped.

5.2.2. Step 2: Reconstruction of object in the Object Plane

Now we consider the propagation of light from the object plane to the PAM plane by distance z in the Fresnel propagation approximation. We can write the relation between the MCF in the object and PAM plane by

$$W_M(\mathbf{r}_1,\mathbf{r}_2) = \iint\iint W_O(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2) O(\boldsymbol{\rho}_1) O(\boldsymbol{\rho}_2)^* \times \exp\left\{ i\frac{\pi}{\lambda z}\left[(\boldsymbol{\rho}_1-\mathbf{r}_1)^2-(\boldsymbol{\rho}_2-\mathbf{r}_2)^2\right]\right\} \mathrm{d}\boldsymbol{\rho}_1 \mathrm{d}\boldsymbol{\rho}_2 \tag{5.8}$$

where $W_O(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2)$ is the MCF of the illumination light beam. In Eq. (5.8), by setting $\mathbf{r}_1 = \mathbf{r}_{mn}$ and $\mathbf{r}_2 = \mathbf{r}_0$, we can obtain the expression for the correlation function $W(\mathbf{r}_{mn},\mathbf{r}_0)$ in the PAM plane:

$$W_M(\mathbf{r}_{mn},\mathbf{r}_0) = \iint\iint W_O(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2) O(\boldsymbol{\rho}_1) O(\boldsymbol{\rho}_2)^* \times \exp\left\{ i\frac{\pi}{\lambda z}\left[(\boldsymbol{\rho}_1-\mathbf{r}_{mn})^2-(\boldsymbol{\rho}_2-\mathbf{r}_0)^2\right]\right\} \mathrm{d}\boldsymbol{\rho}_1 \mathrm{d}\boldsymbol{\rho}_2. \tag{5.9}$$

By integrating Eq. (5.9) sequentially first over $\boldsymbol{\rho}_2$ and then over $\boldsymbol{\rho}_1$, we obtain:

$$W_M(\mathbf{r}_{mn},\mathbf{r}_0) = \int T(\boldsymbol{\rho}_1,\mathbf{r}_0) O(\boldsymbol{\rho}_1) \exp\left[i\frac{\pi}{\lambda z}(\boldsymbol{\rho}_1-\mathbf{r}_{mn})^2\right] \mathrm{d}^2\boldsymbol{\rho}_1, \tag{5.10}$$

where

$$T(\boldsymbol{\rho}_1,\mathbf{r}_0) = \int W_O(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2) O(\boldsymbol{\rho}_2)^* \exp\left[-i\frac{\pi}{\lambda z}(\boldsymbol{\rho}_2-\mathbf{r}_0)^2\right] \mathrm{d}\boldsymbol{\rho}_2. \tag{5.11}$$

As can be seen in Eq. (5.10) that $T(\boldsymbol{\rho}_1,\mathbf{r}_0)$ represents a two-dimensional function that depends on the MCF $W_O(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2)$ of the illumination and $O(\boldsymbol{\rho})$ of the object.

Eq. (5.10) shows that by propagating the correlation function $W_M(\mathbf{r}_{mn},\mathbf{r}_0)$ reversely from the PAM plane to the object plane using Fresnel propagation, we obtain $T(\boldsymbol{\rho},\mathbf{r}_0)O(\boldsymbol{\rho})$. Because both $W_M(\mathbf{r}_{mn},\mathbf{r}_0)$ and $T(\boldsymbol{\rho},\mathbf{r}_0)O(\boldsymbol{\rho})$ are 2-dimensional functions, the Fresnel propagation can be computed efficiently.

$T(\boldsymbol{\rho},\mathbf{r}_0)$ acts as a modulation to the transmission function $O(\boldsymbol{\rho})$ of the object. We remark that the modulation $T(\boldsymbol{\rho},\mathbf{r}_0)$ also depends on the object.

For spatially coherent illumination, the MCF $W_O(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2)$ of the illumination is a uniform and constant function and hence the modulation $T(\boldsymbol{\rho},\mathbf{r}_0)$, although still

depending on the object, is also a constant. Therefore eliminating $T(\boldsymbol{\rho}, \mathbf{r}_0)$ is not necessary. However, for spatially partially coherent illumination, $W_O(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2)$ is a 4-dimensional function which yields a 2-dimensional function $T(\boldsymbol{\rho}, \mathbf{r}_0)$.

In our method, we use a differential approach to eliminate the modulation $T(\boldsymbol{\rho}, \mathbf{r}_0)$. The differential approach needs two diffraction patterns with and without perturbation to the transmission function of the object at a particular point $\boldsymbol{\rho} = \boldsymbol{\rho}_p$, respectively. The point perturbation is achieved by changing either the amplitude or the phase of $O(\boldsymbol{\rho})$ in a small region in the vicinity of $\boldsymbol{\rho}_p$. We expressed the perturbed transmission function of the object as

$$\begin{aligned} O_p(\boldsymbol{\rho}) &= [O(\boldsymbol{\rho}) - O(\boldsymbol{\rho})\delta(\boldsymbol{\rho} - \boldsymbol{\rho}_p)] + CO(\boldsymbol{\rho})\delta(\boldsymbol{\rho} - \boldsymbol{\rho}_p) \\ &= O(\boldsymbol{\rho}) + C_p\delta(\boldsymbol{\rho} - \boldsymbol{\rho}_p), \end{aligned} \tag{5.12}$$

where C is the complex-valued constant of perturbation and $C_p = [(C-1)O(\boldsymbol{\rho}_p)$. Eq. (5.12) shows that at $\boldsymbol{\rho} = \boldsymbol{\rho}_p$, the transmission function of the object $O(\boldsymbol{\rho}_p)$ is changed by a constant factor C to obtain $CO(\boldsymbol{\rho}_p)$. Alternatively, we can also interpret that $O_p(\boldsymbol{\rho})$ consists of the transmission function of the original object $O(\boldsymbol{\rho})$ and an extra Dirac delta function $C_p\delta(\boldsymbol{\rho} - \boldsymbol{\rho}_p)$

Substituting $O(\boldsymbol{\rho})$ by the perturbed object $O_p(\boldsymbol{\rho})$ in Eq. (5.9), we obtain

$$\begin{aligned} W_{M,p}(\mathbf{r}_{mn}, \mathbf{r}_0) = \iint\iint &[O(\boldsymbol{\rho}_1) + C_p\delta(\boldsymbol{\rho}_1 - \boldsymbol{\rho}_p)][O(\boldsymbol{\rho}_2) + C_p\delta(\boldsymbol{\rho}_2 - \boldsymbol{\rho}_p)]^* \\ &\times W_O(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2)\exp\left\{i\frac{\pi}{\lambda z}\left[(\boldsymbol{\rho}_1 - \mathbf{r}_{mn})^2 - (\boldsymbol{\rho}_2 - \mathbf{r}_0)^2\right]\right\}\mathrm{d}\boldsymbol{\rho}_1\mathrm{d}\boldsymbol{\rho}_2. \end{aligned} \tag{5.13}$$

Expanding the brackets, we obtain:

$$\begin{aligned} W_{M,p}(\mathbf{r}_{mn}, \mathbf{r}_0) &= |C_p|^2 W_O(\boldsymbol{\rho}_p, \boldsymbol{\rho}_p) + W_M(\mathbf{r}_{mn}, \mathbf{r}_0) \\ &+ \iint [C_p W_O(\boldsymbol{\rho}_p, \boldsymbol{\rho})]O(\boldsymbol{\rho})^* \exp\left\{i\frac{\pi}{\lambda z}\left[(\boldsymbol{\rho}_p - \mathbf{r}_{mn})^2 - (\boldsymbol{\rho} - \mathbf{r}_0)^2\right]\right\}\mathrm{d}\boldsymbol{\rho} \\ &+ \iint [C_p W_O(\boldsymbol{\rho}_p, \boldsymbol{\rho})]^* O(\boldsymbol{\rho}) \exp\left\{i\frac{\pi}{\lambda z}\left[(\boldsymbol{\rho} - \mathbf{r}_{mn})^2 - (\boldsymbol{\rho}_p - \mathbf{r}_0)^2\right]\right\}\mathrm{d}\boldsymbol{\rho}. \end{aligned} \tag{5.14}$$

We can further derive that

$$\begin{aligned} W_{M,p}(\mathbf{r}_{mn}, \mathbf{r}_0) &= |C_p|^2 W_O(\boldsymbol{\rho}_p, \boldsymbol{\rho}_p) + W_M(\mathbf{r}_{mn}, \mathbf{r}_0) \\ &+ \exp\left[i\frac{\pi}{\lambda z}(\boldsymbol{\rho}_p - \mathbf{r}_{mn})^2\right]\iint [C_p W_O(\boldsymbol{\rho}_P, \boldsymbol{\rho})]O(\boldsymbol{\rho})^* \exp\left[-i\frac{\pi}{\lambda z}(\boldsymbol{\rho} - \mathbf{r}_0)^2\right]\mathrm{d}\boldsymbol{\rho} \\ &+ \exp\left[-i\frac{\pi}{\lambda z}(\boldsymbol{\rho}_p - \mathbf{r}_0)^2\right]\iint [C_p W_O(\boldsymbol{\rho}_p, \boldsymbol{\rho})]^* O(\boldsymbol{\rho}) \exp\left[i\frac{\pi}{\lambda z}(\boldsymbol{\rho} - \mathbf{r}_{mn})^2\right]\mathrm{d}\boldsymbol{\rho}. \end{aligned} \tag{5.15}$$

Subtracting the unperturbed correlation function $W_M(\mathbf{r}_{mn}, \mathbf{r}_0)$ from the perturbed correlation function $W_{M,p}(\mathbf{r}_{mn}, \mathbf{r}_0)$ yields

$$\begin{aligned} W_{M,p}(\mathbf{r}_{mn}, \mathbf{r}_0) - W_M(\mathbf{r}_{mn}, \mathbf{r}_0) = |C_p|^2 W_O(\boldsymbol{\rho}_p, \boldsymbol{\rho}_p) + \alpha \exp\left[i\frac{\pi}{\lambda z}(\boldsymbol{\rho}_p - \mathbf{r}_{mn})^2\right] \\ + \beta \iint W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)O(\boldsymbol{\rho}) \exp\left[i\frac{\pi}{\lambda z}(\boldsymbol{\rho} - \mathbf{r}_{mn})^2\right]\mathrm{d}^2\boldsymbol{\rho}, \end{aligned} \tag{5.16}$$

where

$$\alpha = \iint [C_p W_O(\boldsymbol{\rho}_P, \boldsymbol{\rho})] O(\boldsymbol{\rho})^* \exp\left[-i\frac{\pi}{\lambda z}(\boldsymbol{\rho} - \mathbf{r}_0)^2\right] \mathrm{d}\boldsymbol{\rho}, \tag{5.17}$$

and

$$\beta = C_p^* \exp\left[-i\frac{\pi}{\lambda z}(\boldsymbol{\rho}_p - \mathbf{r}_0)^2\right]. \tag{5.18}$$

As can be seen in Eq. (5.16), the first term is a constant, the second term can be regarded generated by a point source located at $\boldsymbol{\rho}_0$ in the object plane, and the third term is given by Fresnel propagating the product of $O(\boldsymbol{\rho})$ and $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ from the object plane to the PAM plane.

Finally, $O(\boldsymbol{\rho})W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ can be reconstructed by reversely Fresnel propagating $W_{M,p}(\mathbf{r}_{mn}, \mathbf{r}_0) - W_M(\mathbf{r}_{mn}, \mathbf{r}_0)$ from the PAM plane to the object plane. $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ represents the correlation function of the illumination in the object plane between the fields at the perturbation point $\boldsymbol{\rho}_p$ and all locations $\boldsymbol{\rho}$. Notice that $O(\boldsymbol{\rho})W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ at the location of the perturbation point $\boldsymbol{\rho}_p$ cannot be reconstructed.

$W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ is primarily determined by the illumination. However, the location of the perturbation point $\boldsymbol{\rho} = \boldsymbol{\rho}_0$ also plays a role. Usually, $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ is simply a Gaussian function, for example when using the Gaussian-Schell model beam for illumination. When the coherence structure of the illumination is complicated and unknown, we need to calibrate $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ by performing the reconstruction using an empty window as the object, which allow us to reconstruct only $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$. Then we divide the reconstructed product $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)O(\boldsymbol{\rho})$ by $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ and finally obtain the transmission function of the object $O(\boldsymbol{\rho})$ alone.

This means that in the worst scenario we need three measurements for reconstructing the object:

1. measurement with the non-perturbed object,
2. measurement with the perturbed object,
3. measurement without any object.

We remark that the main reason for propagating the MCF from the PAM plane to the object plane is because the sampling interval in the PAM, given by the pitch of the PAM, is too large. By Fresnel propagation we can scale MCF to achieve a sufficiently small sampling interval in the object plane.

5.3. Results and Discussions

In the experiment we use SPC beams with two types of correlations for illumination: the Gaussian Schell-model (GSM) beam and the Laguerre-Gaussian Schell-model (LGSM) beam. We validate our method for each type of illumination.

The experimental setup for generating the GSM beam is shown in Fig. 5.3. A coherent laser beam at a wavelength of $\lambda = 523$ nm is expanded by a beam expander (BE) and then focused on a rotating ground-glass disk (RGGD) by lens L1. The light scattered by the RGGD is spatially partially coherent because the RGGD introduces fluctuation of both amplitude and phase and hence destroys the correlation between the fields at any pair of points.

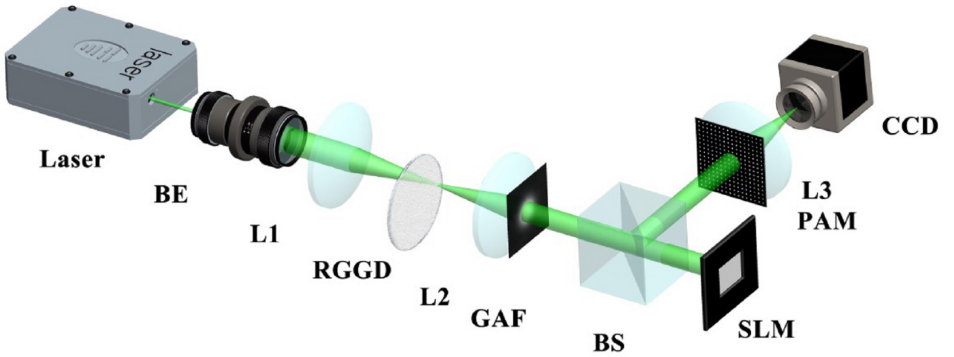

Figure 5.3: Experimental setup for Gaussian Schell-model beam generation and for diffractive imaging. BE: beam expander, RGGD: rough ground glass disc, GAF, Gaussian amplitude filter, BS: beam splitter, SLM: spatial light modulator, PAM: pinhole array mask, and L1,L2,L3 are thin lenses.

The scattered light is collimated by lens L2 and then passed through a Gaussian amplitude filter (GAF). Therefore, in the object plane, both the intensity distribution and the structure of the correlation obey the Gaussian distribution. For GSM beam, the MCF in the object plane is given by

$$W_0(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2) = \exp(-\frac{\boldsymbol{\rho}_1^2+\boldsymbol{\rho}_2^2}{w^2})\exp(-\frac{(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2)^2}{2\sigma^2}), \tag{5.19}$$

where w and σ are the width of the Gaussian intensity distribution and the Gaussian correlation function, respectively.

For the generation of the LGSM beam, we need to insert a spiral phase plate between the BE and the focusing lens L1 in the experimental setup shown in Fig. 5.3. The spiral phase plate produces a dark hollow focal spot on the RGGD. The order n of the LGSM beam is determined by the topological charge of the spiral. When $n = 0$, the spiral phase plate has a constant phase and the LGSM beam becomes the GSM beam. When $n \neq 0$, the MCF of the LGSm beam and the GSM beam have the same amplitude but different phase. We can express the MCF of the LGSM beam in the object plane as

$$W_0(\boldsymbol{\rho}_1,\boldsymbol{\rho}_2) = \exp(-\frac{\boldsymbol{\rho}_1^2+\boldsymbol{\rho}_2^2}{w^2})\exp(-\frac{(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2)^2}{2\sigma^2})L_n^0(\frac{(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2)^2}{2\sigma^2}), \tag{5.20}$$

where $L_n^0(\boldsymbol{\rho})$ is the Laguerre polynomial of order n and $m = 0$. The experimental generation of the LGSM beam has been reported in [24, 25].

The width w of the Gaussian intensity distribution is determined by the GAF and is set to be 0.85 mm, while the width σ of the Gaussian correlation function, also known as the degree of coherence, is determined by the size of the focal spot on the RGGD. We can control σ by translating back-and-forth the focusing lens L1. We calibrate the degree of coherence σ using the method proposed in [26].

In the experiment for diffractive imaging, we use a phase object which has uniform amplitude and binary phase (0.1π and 0.9π) in the shape of a panda. The

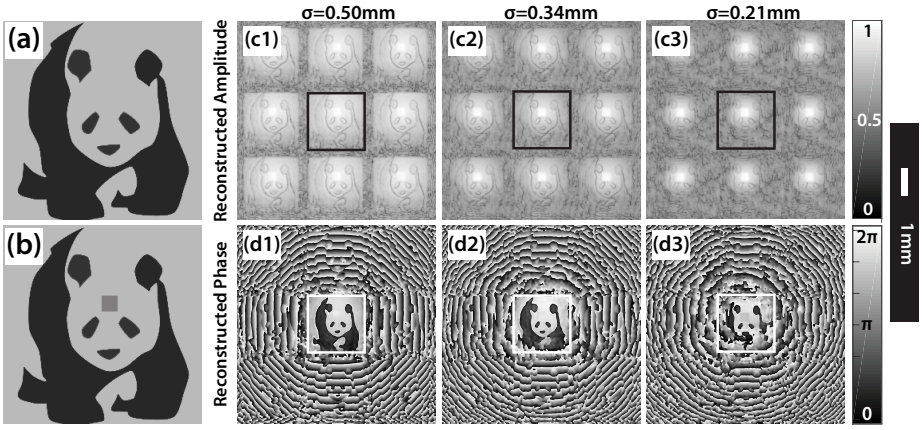

Figure 5.4: The unperturbed and the perturbed object transmission function and the experimental results using GSM beam illumination with various degree of spatial coherence σ. (a,b): the phase of the unperturbed (a) and the perturbed (b) object. The perturbation is at the head of the panda. (c,d): the amplitude (c1-c3) and the phase (d1-d3) of the retrieved product of the transmission function of the object and the correlation function of the illumination.

5

phase object is displayed on a reflective phase spatial light modulator (SLM). By applying phase tilt to the SLM, we deviate the beam that is incident on the area inside and outside the support of the object to the direction of the BS, which then propagates to the PAM and the camera, and to another direction, respectively.

Finally, we measure the far-field diffraction pattern of the light transmitted by the PAM using a CCD camera. The PAM and the CCD camera are placed in the front and the back focal plane of the Fourier transform lens L3 with a focal length of 150 mm, respectively. We set the pitch of the PAM and the size of the pinhole to be $p_x = p_y = 270$ μm and $w_x = w_y = 54$ μm, respectively. The object on the SLM is defined to be with a size 240 pixel $\times$ 240 pixel and with a resolution 8 μm $\times$ 8 μm. The propagation distance between the object and the PAM is $z = 1170mm$.

5.3.1. Experimental results using GSM beam illumination

As indicated by Eq. (5.16), our method requires two diffraction patterns for the object with and without the perturbation, respectively. This is because the GSM beam has a uniform phase. By inverse Fourier transform the two measurements, we retrieve $W_{M,p}(\mathbf{r}_{mn}, \mathbf{r}_0)$ and $W_M(\mathbf{r}_{mn}, \mathbf{r}_0)$, respectively. By reversely propagating the difference from the PAM plane to the object plane, we retrieve the product $O(\boldsymbol{\rho})W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$, where $O(\boldsymbol{\rho})$ is the transmission function of the object and $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ is the correlation function with respect to $\boldsymbol{\rho}_p$, the location of the perturbation point.

In Fig. 5.4, we illustrated the retrieved product $O(\boldsymbol{\rho})W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ for illumination with various degrees of coherence. The perturbation point is placed on the head of the panda and is shown by the gray square. Because for the GSM beam $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ has a uniform phase, the retrieved phase is contributed only by the phase object, in which the panda shape is clearly visible. The retrieved amplitude follows the

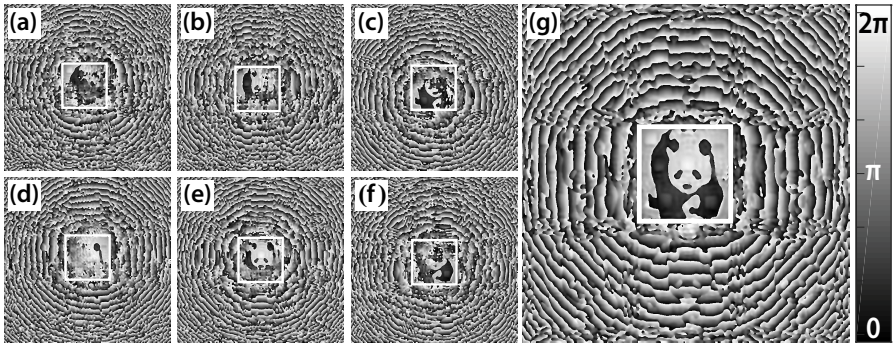

Figure 5.5: The experimental results using GSM beam illumination with the perturbation point at various locations of the object. (a-f): The phase of the experimental reconstruction results. Each result shows a reduced FOV in the vicinity of the location of the perturbation point. (g): The combination of the results in (a-f), which shows a clear panda in the entire FOV.

Gaussian distribution given by the amplitude of $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$.

The panda shape in the amplitude is due to the discontinuity of the phase. We can observe in Fig. 5.4 that for a lower degree of spatial coherence σ, the amplitude of the correlation function $W_0(\boldsymbol{\rho}, \boldsymbol{\rho}_0)$ decreases faster as the distance $\boldsymbol{\rho}_0 - \boldsymbol{\rho}$ increases, and hence the FOV of object's transmission function $O(\boldsymbol{\rho})$ is smaller.

As can be seen in Fig. 5.4 that the degree of coherence σ determines the spread of the amplitude and hence determines the FOV of the phase. We remark that the phase is lost at locations where the amplitude is corrupted by noise.

Fig. 5.4 indicates that to increase the FOV, we can either increase the degree of coherence σ or decrease the noise level. In Fig. 5.5 we demonstrate that by placing the perturbation point at different locations, we can retrieve different parts of $O(\boldsymbol{\rho})$. Therefore we can still retrieve $O(\boldsymbol{\rho})$ in the entire FOV in the case of lowest degree of coherence ($\sigma = 0.21$mm).

However, the approach requires repeating the measurement and the retrieval for each location of the perturbation point. By combining the retrieved $O(\boldsymbol{\rho})$ using low σ illumination, we can obtain $O(\boldsymbol{\rho})$ in the entire FOV as if using high σ illumination.

5.3.2. Experimental results using LGSM beam illumination

In Fig. 5.4 and 5.5, the phase of the retrieved product $O(\boldsymbol{\rho})W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ is contributed by only the phase object because the correlation function of the GSM beam has uniform phase. However, for the LGSM beam, the phase of the correlation function is not uniform. In Fig. 5.6(a) we show the retrieved product $O(\boldsymbol{\rho})W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ when using the LGSM beam for illumination. We can observe that the panda shape is not visible due to the modulation by the phase of $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$.

As a result, we need to calibrate $W_0(\boldsymbol{\rho}_0, \boldsymbol{\rho})$. The calibration can be done by applying our method to an empty window, which allows the retrieval of only $W_O(\boldsymbol{\rho}, \boldsymbol{\rho}_p)$ as shown in Fig. 5.6(b).

Finally, in Fig. 5.6(c) we exhibit the transmission function of the object $O(\boldsymbol{\rho})$ ob-

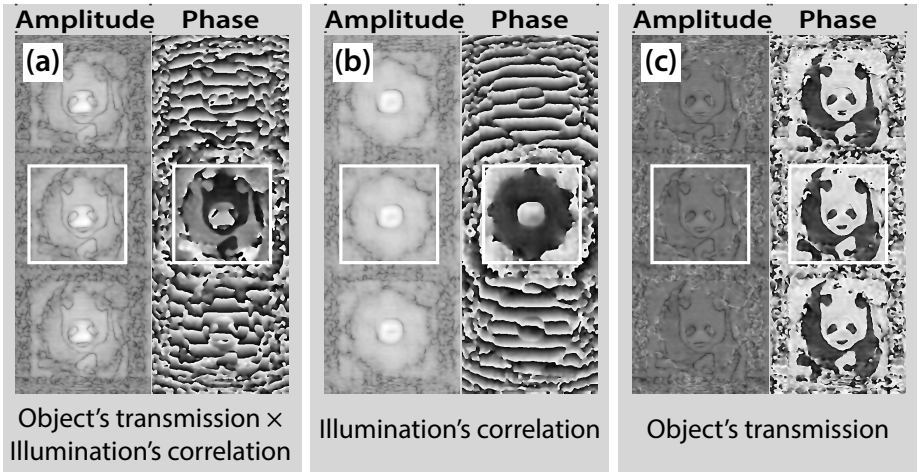

Figure 5.6: The experimental results using LGSM beam for illumination. (a): the reconstructed product of the object's transmission and the illumination's correlation. (b): the calibration result of the the illumination's correlation using an empty window as object. (c): the object's transmission obtained by dividing the results in (a) by the result in (b).

tained by dividing the reconstructed product $O(\boldsymbol{\rho})W_0(\boldsymbol{\rho}_0,\boldsymbol{\rho})$ by the calibrated correlation function $W_0(\boldsymbol{\rho}_0,\boldsymbol{\rho})$. Now the panda shape in the phase of the reconstructed object is clearly visible. This example validates that our method can be applied to object reconstruction in cases when the MCF of the illumination beam is unknown a prior.

5.4. Conclusion

In summary, we developed and validated a non-iterative method to reconstruct the complex-valued transmission function of an object illuminated by spatially partially coherent beam using a pinhole array mask placed in between the object and the camera. Our method overcomes several challenges of conventional iterative CDI algorithms and holographic methods. In particular, our method does not depend on the mode decomposition of the MCF of the SPC beam, and can freely choose the location of the perturbation point, which is beneficial for achieving a large FOV when using a low degree of spatial coherence in the illumination.

Moreover, we demonstrate that our method can be used to calibrate the MCF of an arbitrary SPC beam. On one hand, the calibration allows the reconstruction of the object almost as accurate as if using complete spatially coherent illumination. On the other hand, the calibration provides a novel approach for spatial coherence property characterization, which is needed for applications like the measurement of an optical coherence singularity [27, 28]. Finally, our method is wavelength independent and hence can be applied to a wide range of wavelengths, from X-ray to infrared light.

References

[1] X. Lu, Y. Shao, C. Zhao, S. Konijnenberg, X. Zhu, Y. Tang, Y. Cai, and H. P. Urbach, *Noniterative spatially partially coherent diffractive imaging using pinhole array mask,* Advanced Photonics **1**, 016005 (2019).

[2] J. Miao, Y. Nishino, Y. Kohmura, B. Johnson, C. Song, S. H. Risbud, and T. Ishikawa, *Quantitative image reconstruction of gan quantum dots from oversampled diffraction intensities alone,* Physical review letters **95**, 085503 (2005).

[3] C. Song, H. Jiang, A. Mancuso, B. Amirbekian, L. Peng, R. Sun, S. S. Shah, Z. H. Zhou, T. Ishikawa, and J. Miao, *Quantitative imaging of single, unstained viruses with coherent x rays,* Physical review letters **101**, 158101 (2008).

[4] J. Miao, P. Charalambous, J. Kirz, and D. Sayre, *Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens,* Nature **400**, 342 (1999).

[5] R. W. Gerchberg, *A practical algorithm for the determination of phase from image and diffraction plane pictures,* Optik **35**, 237 (1972).

[6] J. R. Fienup, *Phase retrieval algorithms: a comparison,* Applied optics **21**, 2758 (1982).

[7] J. R. Fienup, *Reconstruction of a complex-valued object from the modulus of its fourier transform using a support constraint,* JOSA A **4**, 118 (1987).

[8] H. Faulkner and J. Rodenburg, *Movable aperture lensless transmission microscopy: a novel phase retrieval algorithm,* Physical review letters **93**, 023903 (2004).

[9] J. Rodenburg, A. Hurst, A. Cullis, B. Dobson, F. Pfeiffer, O. Bunk, C. David, K. Jefimovs, and I. Johnson, *Hard-x-ray lensless imaging of extended objects,* Physical review letters **98**, 034801 (2007).

[10] A. M. Maiden and J. M. Rodenburg, *An improved ptychographical phase retrieval algorithm for diffractive imaging,* Ultramicroscopy **109**, 1256 (2009).

[11] N. Nakajima, *Noniterative phase retrieval from a single diffraction intensity pattern by use of an aperture array,* Physical review letters **98**, 223901 (2007).

[12] C.-S. Guo, K. Liang, X.-T. Zhang, and H.-T. Wang, *Real-time coherent diffractive imaging with convolution-solvable sampling array,* Optics letters **35**, 850 (2010).

[13] A. Konijnenberg, X. Lu, L. Liu, W. Coene, C. Zhao, and H. Urbach, *Non-iterative method for phase retrieval and coherence characterization by focus variation using a fixed star-shaped mask,* Optics express **26**, 9332 (2018).

[14] A. Konijnenberg, W. Coene, and H. Urbach, *Non-iterative phase retrieval by phase modulation through a single parameter,* Ultramicroscopy **174**, 70 (2017).

[15] Y. Shao, X. Lu, S. Konijnenberg, C. Zhao, Y. Cai, and H. P. Urbach, *Spatial coherence measurement and partially coherent diffractive imaging using self-referencing holography,* Optics express **26**, 4479 (2018).

[16] L. Whitehead, G. Williams, H. Quiney, D. Vine, R. Dilanian, S. Flewett, K. Nugent, A. G. Peele, E. Balaur, and I. McNulty, *Diffractive imaging using partially coherent x rays,* Physical review letters **103**, 243902 (2009).

[17] P. Thibault and A. Menzel, *Reconstructing state mixtures from diffraction measurements,* Nature **494**, 68 (2013).

[18] D. Parks, X. Shi, and S. Kevan, *Partially coherent x-ray diffractive imaging of complex objects,* Physical Review A **89**, 063824 (2014).

[19] P. Li, T. Edo, D. Batey, J. Rodenburg, and A. Maiden, *Breaking ambiguities in mixed state ptychography,* Optics express **24**, 9038 (2016).

[20] I. McNulty, J. Kirz, C. Jacobsen, E. H. Anderson, M. R. Howells, and D. P. Kern, *High-resolution imaging by fourier transform x-ray holography,* Science **256**, 1009 (1992).

[21] S. Eisebitt, J. Lüning, W. Schlotter, M. Lörgen, O. Hellwig, W. Eberhardt, and J. Stöhr, *Lensless imaging of magnetic nanostructures by x-ray spectroholography,* Nature **432**, 885 (2004).

[22] L.-M. Stadler, C. Gutt, T. Autenrieth, O. Leupold, S. Rehbein, Y. Chushkin, and G. Grübel, *Hard x ray holographic diffraction imaging,* Physical review letters **100**, 245503 (2008).

[23] P. Gao, B. Yao, I. Harder, N. Lindlein, and F. J. Torcal-Milla, *Phase-shifting zernike phase contrast microscopy for quantitative phase measurement,* Optics letters **36**, 4305 (2011).

[24] Y. Chen and Y. Cai, *Generation of a controllable optical cage by focusing a laguerre–gaussian correlated schell-model beam,* Optics letters **39**, 2549 (2014).

[25] Y. Chen, F. Wang, C. Zhao, and Y. Cai, *Experimental demonstration of a laguerre-gaussian correlated schell-model vortex beam,* Optics express **22**, 5826 (2014).

[26] F. Wang and Y. Cai, *Experimental observation of fractional fourier transform for a partially coherent optical beam with gaussian statistics,* JOSA A **24**, 1937 (2007).

[27] D. Palacios, I. Maleev, A. Marathay, and G. Swartzlander Jr, *Spatial correlation singularity of a vortex field,* Physical review letters **92**, 143905 (2004).

[28] W. Wang, Z. Duan, S. G. Hanson, Y. Miyamoto, and M. Takeda, *Experimental study of coherence vortices: local properties of phase singularities in a spatial coherence function,* Physical review letters **96**, 073902 (2006).

6

Conclusion

In this thesis we try to provide novel solutions to key problems related to imaging and imaging system.

In chapter 2 we study the problem of aberration retrieval for spatially incoherent imaging system. Our approach is based on images with phase diversity. In particular, we used images with defocus diversity in the experiment. Our approach requires the imaging process to be spatially inchoerent. Nature scenes, for example objects illuminated by nature light or objects in the sky such as stars, will satisfy the requirement.

The phase diversity approach was first proposed by Paxman in 1992, which is only a theoretical work. In chapter 2, we present the theoretical framework for solving the inverse problem of aberration retrieval using iterative optimization. Further, we discuss the implementation of the approach in detail. Techniques such as window filtering and Tikhonov regularization are crucial for the image restoration.

We show that a blurred image can be restored to a diffraction-limited image using our approach. Only the information of the diversity is required. In the visible wavelength range, our approach offers an alternative to the existing methods such as wavefront sensor and shearing interferometer. However, for applications like scanning electron microscope (SEM), our approach exhibits the potential to become the standard for aberration metrology.

In chapter 3 we study the problem of measuring the aberrations of a lithographic imaging system, which has both high resolution and large FOV. The aberrations vary spatially over its entire FOV.

We propose a novel measurement scheme using a pair of periodic masks. Our method allows the measurement of the PSF-like images at as many FOV locations as the number of the camera pixels in parallel. The aberrations at each FOV location are then retrieved from the corresponding PSF-like image by iterative optimization.

Our method offers great ability of parallelizing the process of not only the measurement but also the retrieval. It can be further improved by accelerating the retrieval using advanced computational techniques such as machine learning and

neural network. Therefore, it is suitable for large FOV imaging applications e.g. optical lithography.

There are three ways of using the proposed measurement scheme:

1. Real-time alignment.
2. Measurement of distortion, defocus (field-curvature), and telecentricity.
3. Measurement of the aberrations in terms of the Zernike polynomials.

1 and 2 have been implemented and validated by Mikhail Loktev at Liteq B.V. [1], and 3 has been studied at TUDelft based on the experimental data used in [1]. The proposed measurement scheme provides a complete industrial level solution for the measurement spatially-varying aberrations. A patent has been filed and we are now looking for companies and institutions who are interested in acquiring it.

In chapter 4 and 5, we developed two methods for measuring the MCF of an arbitrary light beam, respectively. Both methods use the concept of holography. The key is to split the light beam into two parts keeping the correlation. We guarantee that the two parts of the light beam can form an interference pattern. The rest of the story is to extract the two cross-terms from the interference pattern.

Our method in chapter 4 is believed to be one of the most efficient method for measuring the complete complex-valued MCF: we measure a 2-dimensional slice of the 4-dimensional MCF at a time. Our method in chapter 5 is a non-iterative method for diffractive imaging. We validate that the transmission/reflection function of a sample can be reconstructed (imaged) almost independent of the spatial coherence of the illumination beam.

References

[1] M. Loktev and Y. Shao, *Projection lens testing with moiré effect,* in *Metrology, Inspection, and Process Control for Microlithography XXXI*, Vol. 10145 (International Society for Optics and Photonics, 2017) p. 101452S.

Curriculum Vitæ

Yifeng SHAO

30-05-1989 Born in Wuhan, China.

Education

2007–2011 Bachelor degree in Science
Sun Yat-sen University
Guangzhou, China

2011–2012 Master in Optics, Matter & Plasma
Institut d'Optique de l'Université Paris-Saclay
co-accredited by Ecole Polytechnique
Palaiseau, France

2012 – 2013 Master in Applied Physics / Ingenieur
Delft University of Technology
Delft, Netherlands

2013 Ph.D
Delft University of Technology
Delft, Netherlands

Thesis:	Inverse Problem on Imaging and Imaging System
Promotor:	Prof.dr. H.P. Urbach
Copromotor:	Dr. F. Bociort

List of Publications

6. **Y. Shao,**, M. Loktev, Y. Tang, F. Bociort, & H. P. Urbach, *Spatially varying aberration calibration using a pair of matched periodic pinhole array masks*, Optics Express, 27(2), 729-742 (2019).

5. X. Lu, **Y. Shao**, C. Zhao, S. Konijnenberg, X. Zhu, Y. Tang, & H. P. Urbach, *Noniterative spatially partially coherent diffractive imaging using pinhole array mask*, Advanced Photonics, 1(1), 016005 (2019). (Equal contribution first author)

4. **Y. Shao**, X. Lu, S. Konijnenberg, C. Zhao, Y. Cai, & H. P. Urbach, *Spatial coherence measurement and partially coherent diffractive imaging using self-referencing holography*, Optics Express, 26(4), 4479-4490 (2018)

3. **Y. Shao**, & S. Konijnenberg, *Introduction to computational Phase retrieval and its applications*, Photonics Magazine (publisher: PhotonicsNL), 42e jaargang nummer 3 (2017)

2. M. Strauch, **Y. Shao**, F. Bociort, & H. P. Urbach, *Study of surface modes on a vibrating electrowetting liquid lens*, Applied Physics Letters, 111(17), 171106 (2017)

1. **Y. Shao**, N. Doelman, S. F. Pereira, & H. P. Urbach, *Extended object recon-struction from image intensities blurred by unknown aberrations*, arXiv preprint arXiv:1510.06254 (2015)

Conference Proceedings

3. M. Strauch, S. Konijnenberg, **Y. Shao**, & H. P. Urbach, *Wavefront shaping with an electrowetting liquid lens using surface harmonics*, Adaptive Optics and Wavefront Control for Biological Systems III Vol. 10073, p. 1007304 (2017, April).

2. M. Loktev, & **Y. Shao**, *Projection lens testing with Moiré effect*, Metrology, Inspection, and Process Control for Microlithography XXXI Vol. 10145, p. 101452S (2017, March)

1. I. Livshits, Z. Hou, P. van Grol, **Y. Shao**, M. van Turnhout, H. P. Urbach, & F. Bociort, *Using saddle points for challenging optical design tasks*, Current Developments in Lens Design and Optical Engineering XV Vol. 9192, p. 919204 (2014, September)

Conference Presentations and Posters

6. *Direct measurement of complex-valued mutual coherence function*, EOS topical meeting on Diffractive Optics 2017

5. *Non-iterative diffraction imaging method using partially coherent illumination*, SPIE, Digital Optical Technologies 2017 (cancelled due to illness)

4. *Non-iterative filed reconstruction method for partially coherent illumination*, Light Conference 2016 (best poster prize)

3. *Aberrations retrieval and restoration of diffraction-limited image from two measured image intensi-ties by computational imaging*, 10th International Conference on Optics-Photonics Design and Fabrication, 2016

2. *Aberration Retrieval by Using Extended Nijboer-Zernike Theory*, 9th International Conference on Optics-photonics Design and Fabrication, 2014

1. *Demonstration of aberration retrieval by using extended Nijboer-Zernike theory*, Fraunhofer IISB Lithography Simulation Workshop , 2014