PHOTONICS Research

Single-shot three-input phase retrieval for quantitative back focal plane measurement

Mengqi Shen,[1,†] Qi Zou,[1,†] Xiaoping Jiang,[1,2] Fu Feng,[1] and Michael G. Somekh[1,3,*]

[1]Nanophotonics Research Center, Shenzhen Key Laboratory of Micro-Scale Optical Information Technology & Institute of Microscale Optoelectronics, Shenzhen University, Shenzhen 518060, China
[2]Department of Electronics and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China
[3]Faculty of Engineering, University of Nottingham, Nottingham NG7 2RD, UK
*Corresponding author: mike.somekh@szu.edu.cn

Received 18 October 2021; revised 8 December 2021; accepted 15 December 2021; posted 16 December 2021 (Doc. ID 445189); published 1 February 2022

This paper presents quantitative measurements facilitated with a new optical system that implements a single-shot three-input phase retrieval algorithm. The new system allows simultaneous acquisition of three distinct input patterns, thus eliminating the requirement for mechanical movement and reducing any registration errors and microphonics. We demonstrate the application of the system for measurement and separation of two distinct attenuation measurements of surface waves, namely, absorption and coupling loss. This is achieved by retrieving the phase in the back focal plane and performing a series of virtual optics computations. This overcomes the need to use a complicated series of hardware manipulations with a spatial light modulator. This gives a far more accurate and faster measurement with a simpler optical system. We also demonstrate that phase measurements allow us to implement different measurement methods to acquire the excitation angle for surface plasmons. Depending on the noise statistics different methods have superior performance, so the best method under particular conditions can be selected. Since the measurements are only weakly correlated, they may also be combined for improved noise performance. The results presented here offer a template for a wider class of measurements in the back focal plane including ellipsometry.

https://doi.org/10.1364/PRJ.445189

1. INTRODUCTION

The use of phase retrieval has proved to be exceptionally important for imaging applications opening up possibilities such as ultrawide field of view [1,2], lensless imaging [3], and 3D imaging [4] to name just a few examples. Another great potential advantage of phase retrieval lies in the field of measurement. This subject has been addressed in far less depth than imaging although some potential applications are considered in Ref. [5]. Our group has been interested in the potential of making measurements in the back focal plane (BFP) of a high numerical aperture (NA) objective to extract quantitative information of plasmon propagation with applications in highly localized sensing [6,7]. In these papers, a spatial light modulator (SLM) was used to control the phase distribution on the sample. In the present paper, we show that by using phase retrieval this situation can be changed dramatically. By extracting the amplitude and phase of the BFP all the manipulations that were, up to now, performed with an SLM can be replaced by digital processing, which is both considerably less expensive and also fundamentally more accurate as once the phase retrieval is complete all the operations performed previously in hardware can be entirely replaced with numerical computation. Although the phase of the BFP may be recovered with interferometry [8,9] phase retrieval methods required less elaborate instrumentation; moreover, the phase retrieval process is an inherently common path so the signal degradation to environmental fluctuations and microphonics is far more benign. In a recent paper, we showed how the propagation path of a rich array of surface waves could be visualized by recovering the amplitude and phase in the BFP, filtering and propagating from the Fourier plane to an imaging plane [10]. In the present paper, we further advance the application of virtual optics to describe some new quantitative measurements from the BFP. In particular, we demonstrate measurements of the real and imaginary parts of the wave vector. Specifically, we showed previously that by performing various hardware manipulations with an SLM it was possible to separate two fundamental mechanisms of attenuation, namely, absorption loss and coupling loss [11]. Separating these signals provides considerable physical insight into the nature of the surface wave propagation as well as the quality and thickness of the metallic supporting layer. In the present paper, we improve on these results using computational

optics, which both simplifies the measurement and improves the reliability. A complementary theme of the present paper is to show that phase retrieval offers different possibilities for measurement of the real part of the surface wave k-vector, and it is often the case that one measurement procedure is not universally superior, so the virtual optics approach allows different measurement protocols to be used on the same data. One is then able to select the most precise method for the particular dataset. There is an additional advantage of being able to invoke multiple measurements. It may be shown by simulation and experiment that the noise from two different measurements often shows a low degree of correlation, in some cases even negative correlation, which provides additional opportunities for noise reduction.

2. PHASE RETRIEVAL ALGORITHM, MEASUREMENT SYSTEM, AND BFP RECONSTRUCTIONS

A. Phase Retrieval Algorithm and Measurement System

In Ref. [10], we adapted the phase retrieval algorithm of Allen and Oxley [12] to perform the phase reconstruction. This algorithm uses a few, typically three, defocus positions to reconstruct the phase using an iterative reconstruction algorithm. In the Allen and Oxley's paper no support constraint was explicitly used. We found, however, that even a loose support constraint for the domain of the reconstructed signal ensured much more reliable and rapid convergence. The details of the algorithm and the stopping criterion are discussed in Appendix A. Since the task is to reconstruct the amplitude and phase of the BFP distribution, the support is well known being determined by the maximum spatial frequency in the Fourier plane, which is, of course, determined by the NA of the objective lens. In Ref. [10] this algorithm was used with a single camera and the different defocuses were applied by moving the camera. Although this worked well, moving the camera increased the measurement time. Moreover, there was always a possibility of introducing registration errors between measurements. For this reason, the measurement system has been upgraded to a three-camera real-time data capture system, as shown in Fig. 1. The sample (SPR Kretschmann structure) is illuminated by linearly polarized He–Ne laser with $\lambda = 633$ nm; light reflected from the sample is collected by the objective lens (Nikon, CFI Apochromat TIRF, oil immersion, 100×, NA = 1.49). A polarizer is placed in the detection arm to select the co- or cross-polar components (E_x and E_y). Three detection arms were built to capture three transform images; these are placed close to the Fourier plane of the BFP with different defocus distances to capture the diversity required for the phase retrieval algorithm. To this end three cameras (Thorlabs, CS2100M,1 960 × 1080, 16 bits) were placed at three different negative defocus positions on their corresponding detection arms. Each camera provides one image for the reconstruction algorithm which is shown in Appendix A. To equalize the power to each camera BS2 is a 33:67 beam splitter [Thorlabs, BP133, 33:67 (R:T) split ratio at 635 nm], so one third of the power is directed to Camera 1 and the remaining power is split evenly by BS3 to Camera 2 and Camera 3. Although this ensures that the power to each camera is nearly equal, we added an additional assurance by normalizing the total power detected in Camera 2 and Camera 3 to the power in Camera 1. Another advantage of using the cameras displaced from the focal position is that the dynamic range of the returning signal is greatly reduced so the effects of camera quantization are mitigated significantly. Nevertheless, it is important to ensure that

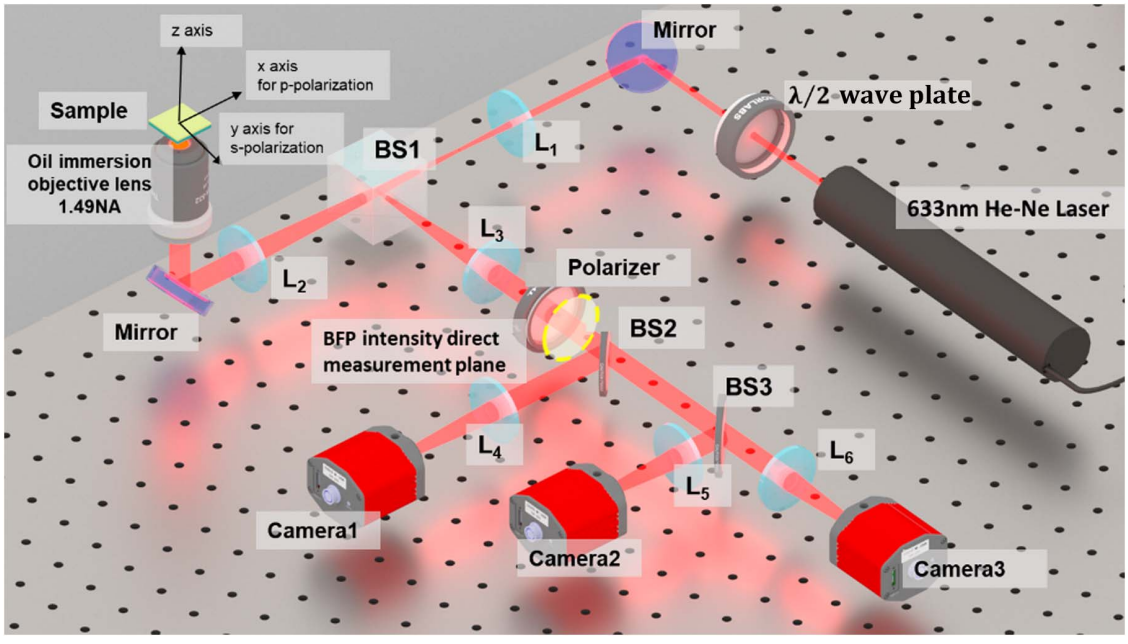

Fig. 1. Schematic of the system used to recover the complex field of the BFP, showing three cameras are placed at three corresponding detection arms with various negative defocus positions. One additional arm (the yellow dashed circle plane) was inserted for direct observation of the intensity in the back focal plane. This was not used in any of the reconstructions and simply used for comparison.

the camera quantization has no material effects on the measurement process. For this reason we implemented an automated high dynamic range stitching procedure described in detail in Appendix B. This operation was integrated into the data acquisition process.

The issue of matching the power into multiple planes by cascaded beam splitters is discussed [13]; however, the input data used in our algorithm is close to the Fourier plane of the reconstructed plane (the BFP). Provided sufficiently different defocus distances are used these planes are relatively uncorrelated providing the diversity needed for reliable reconstruction [10], so that the elaborate splitting approach discussed in Ref. [13] is not required.

It is useful to consider the advantages of the three-camera system compared to alternatives. The main advantage of the present system over the system using a single camera as described in Ref. [10] is the far more rapid data acquisition time. With the single-camera system, even when the camera was mounted on a motorized stage the image acquisition took approximately 2 s. In the present three-camera system all the data can be acquired in 42 ms, even with the high dynamic range stitching discussed in Appendix B. This clearly has advantages in monitoring dynamic samples whose properties change with time and also reduces problems with environmental fluctuations. Furthermore, the use of three cameras eliminates the possible problem with registration error of the camera position and reduces the effects of microphonics since all the measurements are taken in parallel rather than serially.

An alternative approach is to use an SLM as the probe. This was used in slightly different context in Ref. [14]. The use of an SLM can remove registration errors and can deliver relatively rapid data acquisition compared to mechanical movement of a camera. The main aim of the present system, however, is for quantitative measurements as discussed in subsequent sections of this paper, and the non-idealities of SLM discourage their use for reliable quantitative measurements. Such deviations from ideality include non-linear phase response, fill factor below 100%, and phase jitter in some commercial SLMs. While these problems are not necessarily insuperable, they introduce unnecessary and avoidable uncertainties into the measurement process.

B. BFP Reconstructions

It is well known that the BFP of a uniform sample maps the reflectivity of the sample with respect to radial position which is proportional to the sine of the incident angle. For linearly polarized light incident on the BFP along one direction (say horizontal) the signal is purely p-polarized (TM), whereas orthogonal to this direction (say vertical) the signal corresponds to pure s polarization (TE). Along other directions the signal results from interference between these components. Separating and extracting these components is necessary to reduce the noise in the measurement. This is discussed in Appendix C.

We carried out two groups of measurements on gold and silver layers, respectively. Five different thicknesses of gold layer were fabricated—35 nm, 41 nm, 47 nm, 53 nm, and 58 nm, respectively—and five different thicknesses of silver layer were also fabricated at 40 nm, 47 nm, 53 nm, 60 nm, and 66 nm. In addition, an 8 nm thick layer of Al_2O_3 was deposited to avoid the oxidation of silver. These sensors were fabricated using a sputtering machine (Kurt J Lesker, Nano36). A surface profiler was used to calibrate the sputtering machine by measuring the film thickness obtained at a known deposition time. The form of the phase of the reflection coefficient is strongly dependent on the thickness of the metal layer as this strongly controls the loss due to coupling [15]. Tables 1 and 2 show the retrieved phase, retrieved amplitude, and directly measured amplitude of BFPs for gold and silver sensors, respectively. The amplitude is measured by inserting a camera at a plane conjugate to the yellow circle in Fig. 1 for observation and comparison.

Table 1. Retrieved Phase, Retrieved Amplitude, and Measured Amplitude of BFPs for Five Different Thicknesses of Gold Layers

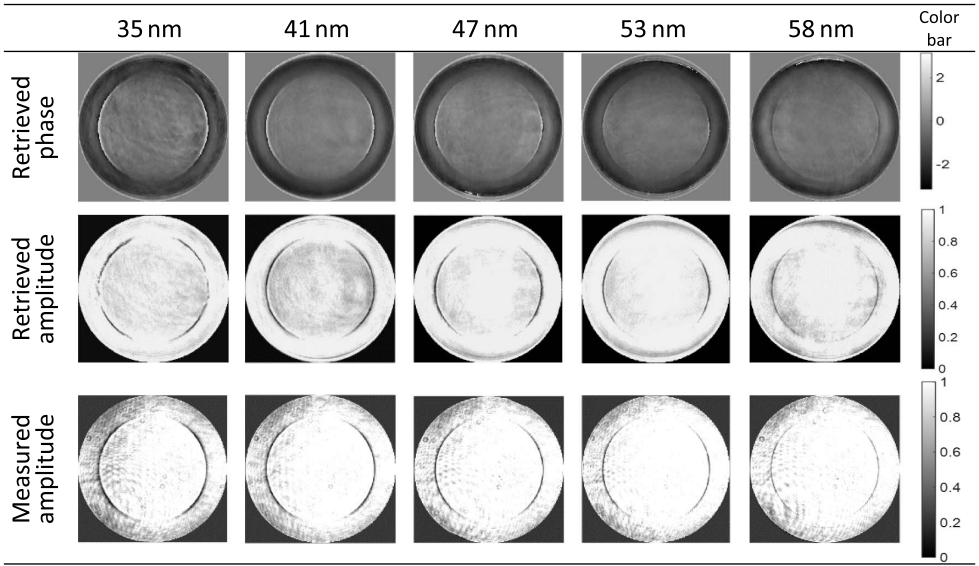

Table 2. Retrieved Phase, Retrieved Amplitude, and Measured Amplitude of BFPs for Five Different Thicknesses of Silver Layers

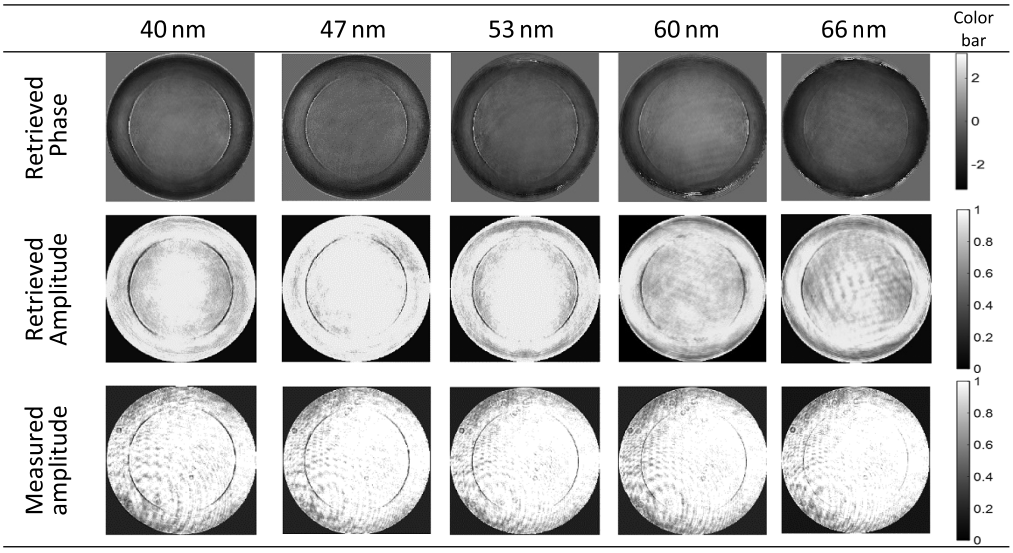

Figure 2 shows the line traces of the amplitude [Fig. 2(a)] and phase [Fig. 2(b)] of p polarization after noise reduction on gold samples of different thicknesses of 35 nm, 41 nm, 47 nm, 53 nm, and 58 nm, respectively. In Table 1, we can see that apart from the pure p-polarization information, there is also a great deal of information at other azimuthal angles, although along directions other than the horizontal (p polarization) and vertical (s polarization) the two polarization states interfere. We therefore apply a least square algorithm to utilize the available data effectively; this method is described in Appendix C. The same process is used to obtain the line trace of s-polarization information.

Table 2 shows the retrieved phase, retrieved amplitude, and measured amplitude of BFPs of silver layers of different thicknesses of 40 nm, 47 nm, 53 nm, 60 nm, and 66 nm, respectively. Figure 3 shows the line traces of the amplitude [Fig. 3(a)] and phase [Fig. 3(b)] of p polarization after noise reduced on silver samples of different thicknesses of 40 nm, 47 nm, 53 nm, 60 nm, and 66 nm, respectively. By comparing the gold sensor data and silver sensor data, it can be seen that, as expected, the silver layer sensor gives narrower dips and sharper phase transition compared to gold sensors.

3. APPLICATIONS OF THE BFP COMPLEX FIELD

A. Separating the Absorption and Coupling Attenuation Mechanisms

With the reconstructed complex BFP field distributions, we then can separate attenuation due to different mechanisms. When a surface plasmon (SP) is excited, there are two principal loss mechanisms: one is due to the absorption loss and the other arises from coupling loss. Reciprocity indicates that if a wave can be coupled from an external propagating mode it will also leak energy as it propagates. Clearly, surface wave energy lost to absorption is converted to heat, whereas leakage energy is converted to propagating light. Crucial to understanding the measurement of attenuation in a microscope system is to understand that in the confocal system the detected light appears to come from the focus so that, for sufficient defocus, the propagation distance $2x$ is related to the defocus by $x = z \tan \theta_p$ (see Fig. 4), where z is the defocus and θ_p is

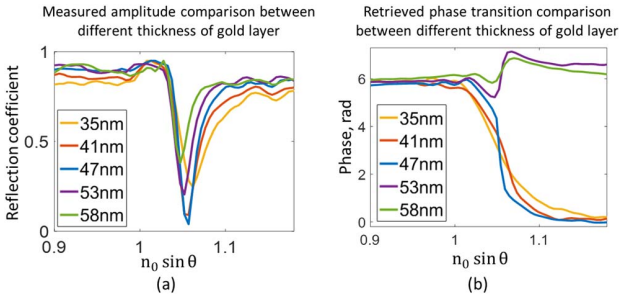

Fig. 2. (a) Measured amplitude comparison between various thicknesses of the gold layer along p polarization; (b) retrieved phase transition comparison between different thicknesses of the gold layer along p polarization. The thick layers show the characteristic phase inversion.

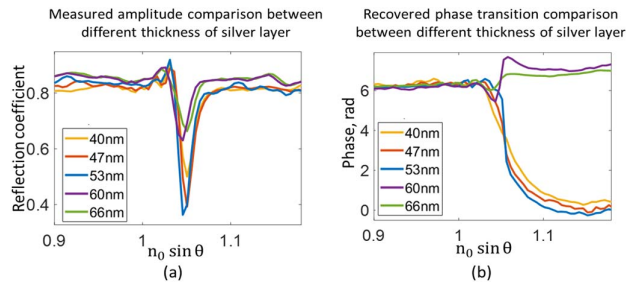

Fig. 3. (a) Measured amplitude comparison between various thicknesses of the silver layer along p polarization; (b) recovered phase transition comparison between different thicknesses of the silver layer along p polarization.

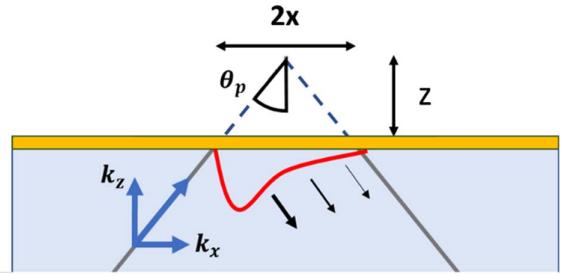

Fig. 4. Reradiated leakage propagating light from the surface plasmon.

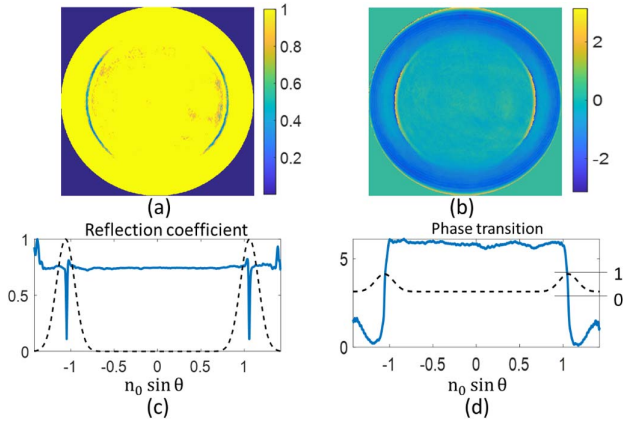

Fig. 5. (a) Retrieved BFP amplitude distribution for 41 nm thick gold sample; (b) retrieved BFP phase distribution for 41 nm thick gold sample; (c) blue line is the amplitude line trace of p polarization after noise reduction, and black dashed line is the pupil function applied; (d) blue line is the phase line trace of p polarization after noise reduction, and black dashed line is the pupil function applied.

the angle at which surface plasmons are excited. In a confocal microscope the pinhole will block out light that does not appear to come from the focus, so the confocal system precisely defines the path of the surface waves. This allows the attenuation to be measured. In the microscope system it is straightforward to measure the total attenuation. We only need to examine the decay of the wave as a function of defocus [16]. It is far more challenging to separate the two attenuation mechanisms, and this was addressed in Ref. [11] using an SLM to produce an "artificial" surface wave whose known properties could be used to calibrate the instrument. Let us consider a change in the SP signal modulus with defocus (which as mentioned above can be readily related to the signal change as it propagates along the sample). This can be represented by

$$|\mathrm{SP}(z)| = 2k''_{cz}\gamma e^{-(k''_{cz}+k''_{\Omega z})z}, \tag{1}$$

where γ is a parameter that describes the instrument; this depends on the pupil function of the microscope, the effective pinhole size of the microscope, and the sensitivity of the detectors. Essentially, this parameter bundles up all the instrumental parameters, k''_{cz} and $k''_{\Omega z}$ are the coupling attenuation and absorption attenuation along the z direction, respectively. Taking natural logs of Eq. (1), we have

$$\ln|\mathrm{SP}(z)| = \ln 2k''_{cz} + \ln\gamma - (k''_{cz} + k''_{\Omega z})z. \tag{2}$$

Therefore, the gradient gives the total attenuation, and the intercept gives the coupling attenuation once the instrumental parameter γ is available. Separating the two attenuations thus requires the determination of γ.

The procedures to separate the coupling attenuation and ohmic attenuation are listed as follows.

(1) Recover the complex BFP with three-input phase retrieval algorithm (Tables 1 and 2).

(2) Use the least square method to get noise reduction for pure p-polarization (r_p) information and pure s-polarization (r_s) information (see Appendix C).

(3) Apply a tapered window on the r_p to filter the region of the reflection coefficient within a few degrees around the plasmon angle and remove the normal incidence angle contribution. Here we use a tapered window (in Fig. 5) rather than an on–off window to avoid the diffraction ripples appearing on the transform images due to the sharp transition. This is unlike the previous work that used an arc or a ring on the BFP to select the mode information. The presence of the complex information gives a much cleaner separation of the two reflection coefficients allowing a complete isolation of r_p, and thus avoids the interference from the positions that have both p- and s-polarization information.

(4) From the complex BFP distribution one can calculate the complex output signal that one would expect if we performed a real experiment where the sample was physically defocused. The complex output $V(z)$ is given by

$$V(z) = \int_0^{\theta_{\max}} P(\theta) r_p(\theta) e^{-2ik\cos\theta z} \sin\theta \, \mathrm{d}\theta, \tag{3}$$

where θ represents the incident angles emerging from the objective lens, k is the wave vector of the light emerging from the objective given by the $2n_0\pi/\lambda$, where n_0 is the refractive index of the oil of the immersion lens and λ is the wavelength in free space (633 nm). These contributions are integrated between normal incidence and the maximum angle of the objective determined by the objective, $\theta_{\max}$. $P(\theta)$ is the pupil function of the objective lens. In this case filtering the recovered BFP with a suitable function allows one to apply a virtual pupil in software.

By choosing $P(\theta)$ to only be non-zero over a range of a few degrees around the plasmon angle, θ_p, as shown schematically in Figs. 5(c) and 5(d), it may be shown by simulation or stationary phase arguments [11] that after approximately 3 μm negative defocus (moving the sample toward the objective) that the magnitude of $V(z)$ closely follows, albeit with some deviations, the ideal form of $|\mathrm{SP}(z)|$ occurs. The curve of $|V(z)|$ should be straight on the log scale; following Eqs. (1) and (2), the total attenuation can thus be determined from the gradient. Figures 6(a) and 6(c) show the $|V(z)|$ curves of the gold and silver layers with various thicknesses. An offset of 0.2 has been applied in Figs. 6(a) and 6(c) for clearer illustration. Figures 6(b) and 6(d) show the same data on the log scale. The curves in Figs. 6(b) and 6(d) show that on the log scale the curves are approximately linear beyond 3 μm negative defocus. The deviations from linearity arise primarily from the branch point near the critical angle in $r_p(\theta)$. Nevertheless, measuring the gradient between −3 μm and −15 μm gives

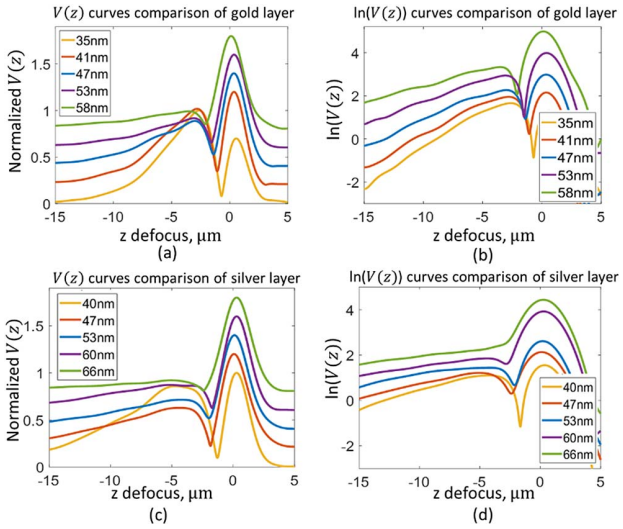

Fig. 6. (a) Normalized experimental $V(z)$ attenuation measurements for 35 nm (yellow curve), 41 nm (red curve), 47 nm (blue curve), 53 nm (purple curve), 58 nm (green curve) thick gold samples; (b) natural log scale of the $V(z)$ curves in (a); (c) normalized experimental $V(z)$ attenuation measurements for 40 nm (yellow curve), 47 nm (red curve), 53 nm (blue curve), 60 nm (purple curve), 66 nm (green curve) thick silver samples; (d) natural log scale of the $V(z)$ curves in (c).

consistent measurements that allow the total attenuation of different sensors to be obtained. An offset of 1 has been applied in Fig. 6(b) and an offset of 0.5 has been applied in Fig. 6(d) for clearer illustration at $z = -15$ μm. It may be readily seen that the quality of the measurements obtained in the present work is much better than that obtained previously with hardware manipulation.

(5) The next stage is to calculate the instrumental parameter, γ. We now use the s-polarized reflection coefficient, r_s. This is multiplied by a series of different phase profiles at the angle that corresponds to the same surface plasmon filter as used for r_p. The reason to do this is that different phase gradients correspond to different values of attenuation, as shown in Figs. 2(b) and 3(b). We are therefore converting r_s to a reflection coefficient that generates a virtual or "artificial" surface wave with known properties. Since r_s is generated with the same objective it will carry similar instrumental parameters to those affecting r_p. Furthermore, since there is no dip in the amplitude of r_s these gradients will correspond to coupling attenuation alone. We can therefore remove the ohmic loss from Eq. (4):

$$\ln|\mathrm{ASP}(z)| = \ln(2k''_{\mathrm{cz,ASP}}) + \ln(\gamma) - (k''_{\mathrm{cz,ASP}})z, \qquad (4)$$

where ASP is the "artificial" surface wave formed by applying a phase gradient to the recovered s-polarized reflection. Since the gradient of our "artificial" surface wave gives the coupling loss, we know the contribution of this loss to the intercept, so the remainder is the instrumental parameter, γ. This can be repeated with different phase profiles applied to r_s to get an averaged value of the instrumental parameter.

(6) A measure of the intercept with known γ then allows us to obtain k''_{cz} for each thickness. From these values of k''_{cz} we can

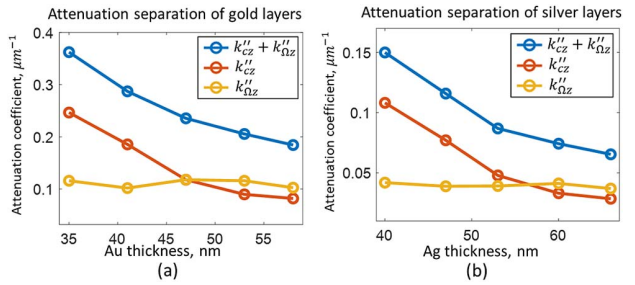

Fig. 7. (a) Attenuation coefficients due to coupling loss and ohmic loss with varying gold thickness; (b) attenuation coefficients due to coupling loss and ohmic loss with varying silver thickness, obtained computationally as opposed to manipulation of the spatial light modulator.

easily obtain $k''_{\Omega z}$, since the sum of these parameters is known from the measurement of total attenuation obtained from the gradient. The attenuation parameters obtained directly from the gradient give the attenuation for a change in defocus, z. This is converted to an attenuation with respect to propagation distance along the surface by noting that $x = z \tan\theta_p$.

The attenuation of the surface waves per micrometer propagation distance along the surface is presented in Fig. 7. The blue line is the recovered total attenuation for different metal thicknesses. The red and yellow lines are the coupling loss coefficient and ohmic loss coefficient, respectively. The result agrees with the predicted trends. First, the coupling loss decreases as the thickness of the film increases. Second, the ohmic loss is relatively insensitive to the variation of the thickness. For film thicknesses greater than the penetration depth of SP in the metal (about 30 nm in gold and 26 nm in silver) the absorption loss is almost constant with layer thickness. The results obtained for the absorption loss in the present experiment show far less variability than those obtained by hardware manipulation [11] reinforcing our expectation of the superior robustness and accuracy of the present virtual optics-based measurement. Third, for gold samples, the thickness value where the two attenuation mechanisms intersect is at approximately 46 nm, corresponding to the position where the reflection coefficient approaches zero. For silver samples, the two curves intersect at around 57 nm thickness where the loss due to absorption and coupling loss are equal. The intercept is consistent with the position of the phase inversion of Figs. 2 and 3 where the ohmic loss exceeds the coupling loss. This shows how computational postprocessing of the complex BFP may be used to recover new information; moreover, we reiterate that the noise on the attenuation measurements is considerably smaller than obtained using hardware manipulation of the wavefronts with the SLM.

B. Measurement of the Real Part of the Surface Wave *k*-Vector

Measurement of the real part of the SP k-vector, or the angle at which SPs are excited, is one of the fundamental measurement tasks applied in the sensing of binding and refractive index changes. The advantage, of course, of measurement with an objective lens as opposed to a prism coupler is that the objective

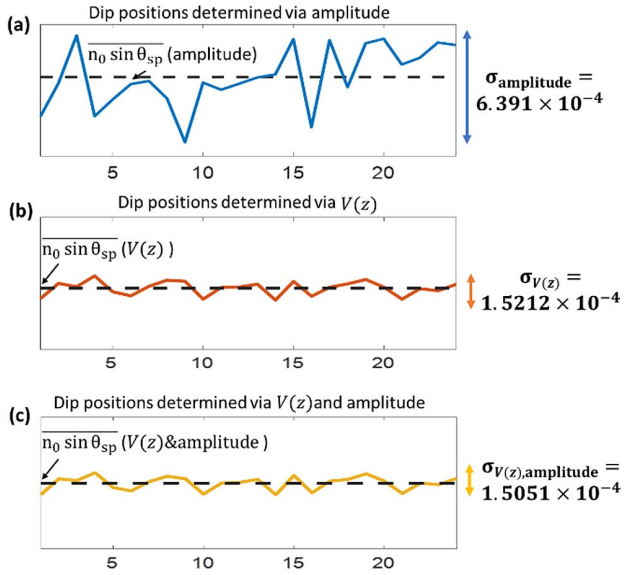

Fig. 8. (a) Dip positions calculated by amplitude; (b) dip positions calculated by $V(z)$; (c) dip positions calculated by combining $V(z)$ and amplitude.

lens examines the sample over much finer spatial location providing sensing operations on a microscopic scale.

A direct intensity measurement of the BFP can provide a measure of the k-vector by measuring the position of the intensity dips and relating this to the sine of the incident angle for SP excitation. Appendix C shows that different algorithms should be used for separation of the TM and TE components of the reflection coefficient when phase measurement is available compared to the case where only the intensity is measured. The use of the phase signal results in better conditioned reconstruction. Even allowing for the fact that in one case the amplitude of the reflection coefficient is reconstructed and in the other the intensity reflection coefficient is obtained, the overall noise in measuring the dip position is considerably better when the phase measurement is available.

The presence of the BFP phase information also allows virtual optics reconstruction of the k-vector. In the previous section we showed that with an appropriate pupil function, $P(\theta)$, that allows transmission of a wave around the angle for SP excitation and an appropriate defocus the attenuation could be evaluated from the modulus of the $V(z)$ where the curve follows the ideal form of Eq. (1). Similarly, if we include the phase variation of the same function, we have

$$\mathrm{SP}(z) = 2k''_{cz}\gamma e^{-(k''_{cz}+k''_{\Omega z})z} e^{\left(i\frac{4\pi n}{\lambda}\cos\theta_p z\right)}. \tag{5}$$

Since the phase of the BFP is available the phase of $V(z)$ is readily calculated, which can be equated to $\mathrm{SP}(z)$ in the negative defocus region away from the focus. Measuring the phase gradient over a suitable range of z allows the value of θ_p to be evaluated and hence the SP wave vector ($\frac{2n\pi}{\lambda}\sin\theta_p$). It is worth pointing out that the fit to a straight line corresponding to the phase is better than the one used for the attenuation. This is largely because the numerical value of the real part of the k-vector is larger than the imaginary part so deviations from linearity are less noticeable.

We therefore have two separate measures of the SP wave vector. The interesting thing is that the noise associated with the values recovered with the two methods is only weakly correlated so the two measures can be combined to improve the overall SNR. The correlation coefficient between these two measurements is only 0.0736. In the present case, the measurement of the $V(z)$ obtained from the complex BFP was much better than the dip measurement, so the benefit of combining the measurements was relatively small. These results are presented in Fig. 8. Simulated results which show different noise cases are discussed in Appendix D.

4. CONCLUSION

In this paper we have shown how non-interferometric phase retrieval can be used to perform quantitative measurements in the BFP. The measurements of attenuation and the differences between the gold and silver samples agree well with theory. Phase measurement also has a major advantage for measurement of the SP excitation angle, θ_p, as it allows different measurement protocols to be applied to the same data. This allows the better measurement for the particular dataset to be selected. Moreover, the relatively small correlation between different methods allows them to be combined statistically to reduce the noise.

We believe these measures of the phase of the BFP offer new capabilities in objective-based measurements, simplifying the hardware while improving the measurement performance. Further developments will allow the inclusion of reference regions to further enhance measurement precision. Finally, although the present measurements have concentrated on surface wave and SP measurement, the approach described may also be applied to other fields of metrology such as ellipsometry.

APPENDIX A: RECONSTRUCTION ALGORITHM

The process of phase reconstruction algorithm is shown in Fig. 9.

Input data: acquire images at three defocus planes and a rough estimate of the window function $W(k_r)$.

Computational operations on input data are as follows.

(1) Initial estimate of the phase and amplitude of the object, $O_{i,j}(k_r,\phi)$, can be random or uniform. In our experiments, a uniform initialization converges approximately 15% more quickly. This is because the reconstructed BFP is approximately constant apart from the regions where surface waves are excited. For other objects an initial random distribution may be better. It should be emphasized that reliable convergence was achieved with either starting condition. The subscript i is the ith transform image corresponding to the ith defocus position, and j represents the jth iteration.

(2) The exit-wave $\psi_{i,j}(k_r,\phi)$ is the product of object $O_{i,j}(k_r,\phi)$, the defocused wavefront $P_i(k_r)$, and the illumination window $W(k_r)$.

(3) Fourier transform $\psi_{i,j}(k_r,\phi)$, $\Psi_{i,j}(u) = \mathcal{F}[\psi_{i,j}(k_r,\phi)]$.

(4) Replace the modulus of the resulting computed inverse Fourier transform with the measured diffraction amplitude ($|\Psi_{\mathrm{measured}}(u)|$), $\Psi'_{i,j}(u) = |\Psi_{i,\mathrm{measured}}(u)|\angle\,\Psi_{i,j}(u)$.

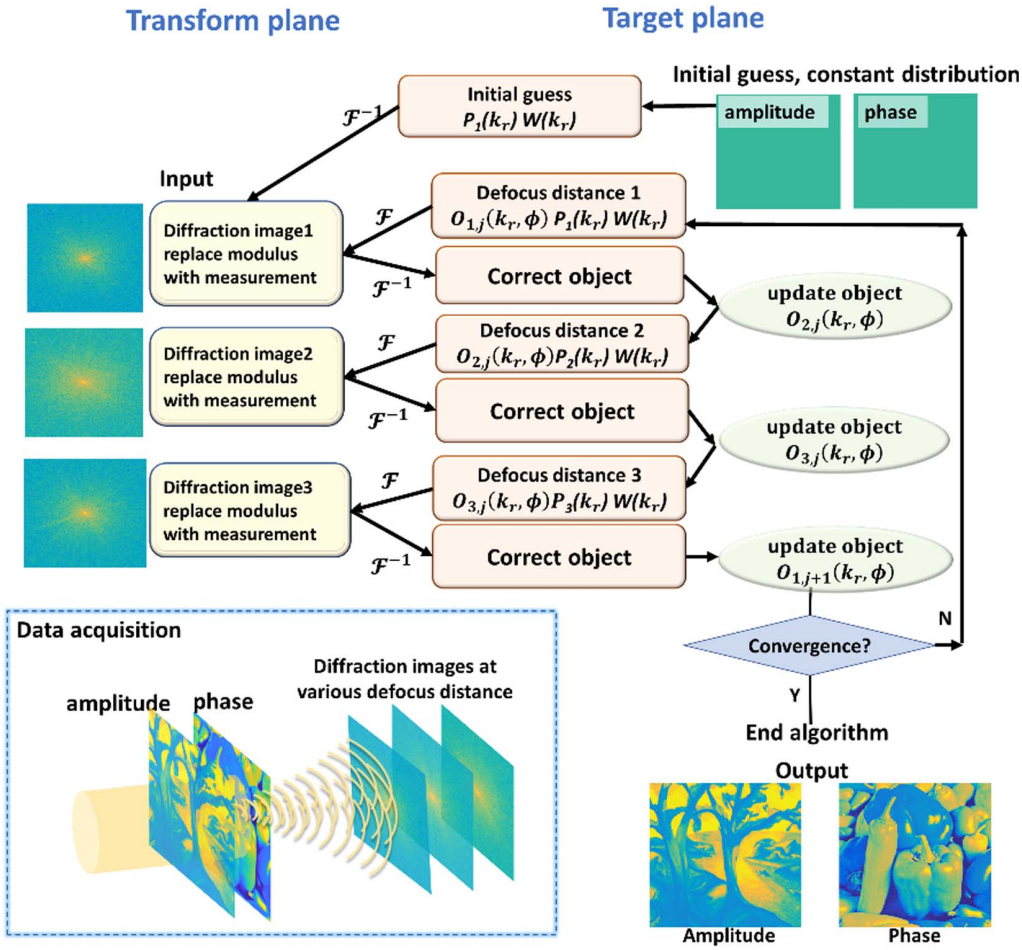

Fig. 9. Flow chart of three-input phase retrieval algorithm.

(5) Inverse Fourier transform $\Psi'_{i,j}(u)$, $\psi'_{i,j}(k_r,\phi) = \mathcal{F}^{-1}[\Psi'_{i,j}(u)]$.

(6) Update function: $O_{i+1,j}(k_r,\phi) = O_{i,j}(k_r,\phi) + P_i(k_r)\cdot[\psi'_{i,j}(k_r,\phi) - \psi_{i,j}(k_r,\phi)]$.

(7) EXIT when the appropriate criterion is reached.

(i) For simulations this is determined by the sum of squares error (SSE) value. If the target SSE is reached, EXIT. The SSE is defined as

$$\mathrm{SSE} = 10\log_{10}\left\{\frac{\sum[|O_j(k_r,\phi)| - |O(k_r,\phi)|]^2}{\sum[|O(k_r,\phi)|]^2}\right\}, \tag{A1}$$

where $|O(k_r,\phi)|$ is the measured amplitude of the target, which can be as a reference to evaluate the accuracy of the retrieved field. $\sum$ means the summation of points in target plane.

(ii) For experimental implementation where the SSE is not available, the algorithm is stopped when the intensity correction in stage (4) does not change. Specifically, this was when the integrated intensity over the measured field changed by less than 5×10^{-5} over the last five iterations.

(8) Return to (2).

APPENDIX B: IMAGE ACQUISITION AND PRE-PROCESSING

The images acquired from the optical system are close to the transform plane; therefore, the pattern will show a sharp peak at the center when close to the focal plane, thus using the available dynamic range of the camera, so that peak values are saturated or small values are swamped by quantization. Using substantial defocus corresponding to a few micrometers at the sample surface greatly relieves this problem; however, to ensure that the low signals are detected while avoiding saturation, multiple exposures are used to increase the dynamic range. This process was automated into the image acquisition process. The process to create high dynamic range (HDR) image is described as follows. Each image is shown in Figs. 10(c)–10(e).

(1) The coordinates of the beam center are located by summing up the most strongly focused image in both dimensions. The maximum indices are the coordinates of the beam center.

(2) A contour integral is calculated for each frame from the beam center to the edge of the image in order to work out the stitching position as in Eq. (B1), and the profiles are shown in Fig. 10(a):

$$A(r) = \int_0^{2\pi} I(r,\theta)\mathrm{d}\theta, \tag{B1}$$

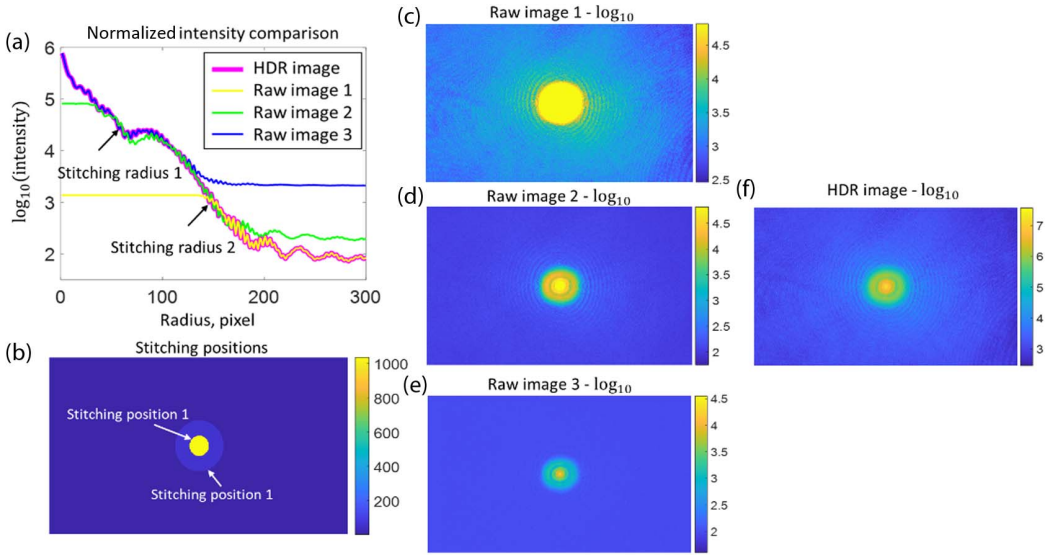

Fig. 10. (a) Normalized intensity comparison between three raw images and the HDR image; (b) shows the coefficients and stitching positions to produce the HDR image; (c) raw image 1 in $\log_{10}$ scale, this is a long exposure image that shows the low intensity values but saturates the center values; (d) raw image 2 in $\log_{10}$ scale; (e) raw image 3 in $\log_{10}$ scale, this is a short exposure image that shows the high intensity center but loses the low intensity regions away from the center; (f) the final HDR image.

where $I(r,\theta)$ is intensity 2D-interpolated from normalized images, and $A(r)$ is the accumulated intensity.

(3) Stitching position is worked out by matching the intensity and slope of the intensity. A coefficient is calculated at the stitching position to compensate for inaccuracy in exposure time. Figure 10(b) shows the coefficient and stitching position to produce the HDR image. Then the images could be stitched together to get an HDR image.

Once the coefficients are calculated for each camera, the process is integrated into the image acquisition sequence to generate the HDR images in real time. The total acquisition time including HDR image generation is less than 0.042 s. Normally, because we are using the defocused plane, two images are sufficient to satisfy the dynamic range requirement. Only extreme cases very close to the focus would require three images. To better illustrate the process, below we used three images to show one extreme case at the defocus of −0.3 μm.

Figure 10 illustrates the process. We can see in Fig. 10(a) that the blue curve representing the lowest exposure image recovers the peak well, but the low values do not show the signal variation due to quantization. Since the exposure time for this curve was shorter, the curve has been calibrated to allow for this. The center values for raw images 1 and 2 are not used as they are saturated, but they seamlessly stitched together, so exposure 2 is used to represent the mid-values and exposure 1 the high values.

APPENDIX C: SEPARATION OF NOISE OF DATA IN BACK FOCAL PLANE

Figure 11 shows the schematic of BFP for linear polarization. Consider linearly polarized illumination aligned along the x direction. The reflected field along the x direction is given by

$$E_x = r_p(\sin\theta)\cos^2\phi + r_s(\sin\theta)\sin^2\phi, \tag{C1}$$

where θ is the incident angle. The radial in the BFP position maps to $\sin\theta$, and the maximum value is determined by the NA of the objective. Dropping the explicit reference to $\sin\theta$, the intensity along the x direction, I_x, is given by (omitting a proportionality term)

$$I_x = |r_p|^2\cos^4\phi + |r_s|^2\sin^4\phi + \frac{1}{2}|r_p||r_s|\cos\beta\sin^2 2\phi, \tag{C2}$$

where β is the phase angle between r_p and r_s.

It can be seen from Fig. 11 that along $\phi = 0$ the signal is purely TM polarized and along $\phi = \frac{\pi}{2}$ the signal is pure TE. At other angles the signal is an average of the two; however, these angles contain a great deal of information, so we try to get a weighted average over ϕ to get an optimum signal-to-noise ratio. This process also allows separation of r_p and r_s, which

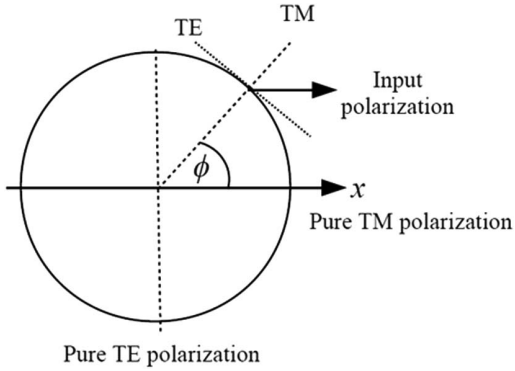

Fig. 11. Schematic of BFP for linear polarization.

proves extremely useful as demonstrated in the main body of the paper.

Consider the case where there is no phase information; in this case, we need to work with intensity as given by Eq. (C2).

To get the TM and TE signals, we use a least square approach analogous to that developed by Greivenkamp for phase stepping interferometry [17]. We measure the intensity distribution I_n along many azimuthal directions, ϕ_n. We wish to find the values of $|r_p|^2$, $|r_s|^2$, and $\frac{1}{2}|r_p||r_s|\cos\beta$ that minimizes the squared difference between the measured data and fitted data, i_n. We thus form a series of equations to minimize the residuals by setting the partial derivatives with respect to the fitting parameters to zero. Thus,

$$\frac{\partial\sum_n (I_n - i_n)^2}{\partial|r_p|^2} = 0, \qquad \frac{\partial\sum_n (I_n - i_n)^2}{\partial|r_s|^2} = 0,$$
$$\frac{\partial\sum_n (I_n - i_n)^2}{\partial|r_p||r_s|} = 0. \tag{C3}$$

From these relations standard manipulations lead to a matrix equation:

$$\begin{bmatrix} \sum_n \cos^8\phi_n & \sum_n \cos^4\phi_n \sin^4\phi_n & \sum_n \cos^4\phi_n \sin^2 2\phi_n \\ \sum_n \cos^4\phi_n \sin^4\phi_n & \sum_n \sin^8\phi_n & \sum_n \sin^4\phi_n \sin^2 2\phi_n \\ \sum_n \cos^4\phi_n \sin^2 2\phi_n & \sum_n \sin^4\phi_n \sin^2 2\phi_n & \sum_n \sin^4 2\phi_n \end{bmatrix}$$
$$\times \begin{bmatrix} |r_p|^2 \\ |r_s|^2 \\ \frac{1}{2}|r_p||r_s|\cos\beta \end{bmatrix} = \begin{bmatrix} \sum_n I_n \cos^4\phi_n \\ \sum_n I_n \sin^4\phi_n \\ \sum_n I_n \sin^2 2\phi_n \end{bmatrix}. \tag{C4}$$

We take many measurements distributed uniformly between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ to form the RHS; this is multiplied by the inverse of the 3×3 matrix to recover the desired parameters. This results in considerable noise reduction and separates the TM and TE components. This procedure was used in Ref. [18] for data recovery with intensity only BFP measurements.

If the phase of the signal is available, we do not need to work with Eq. (C2) but can work with Eq. (C1) directly. We can apply exactly the same procedure to the real and imaginary parts of the recovered complex field. This leads to a matrix equation for the real part of the reflection coefficients:

$$\frac{1}{2}\begin{bmatrix} \sum_n \cos^4\phi_n & \sum_n \cos^2\phi_n \sin^2\phi_n \\ \sum_n \cos^2\phi_n \sin^2\phi_n & \sum_n \sin^4\phi_n \end{bmatrix}\begin{bmatrix} r'_p \\ r'_s \end{bmatrix} = \begin{bmatrix} \sum_n E'_{xn} \cos^2\phi_n \\ \sum_n E'_{xn} \sin^2\phi_n \end{bmatrix}, \tag{C5}$$

where the single dash denotes the real part of the field. An identical equation is used for the imaginary part allowing the complex value of r_p and r_s to be recovered. This processing was used in our earlier work [10].

Both formulations allow separation of the TM and TE components and permit noise averaging. The noise reduction is more effective when the phase measurement is available because the condition number of the 2×2 matrix in Eq. (C5) is 2 for uniformly distributed values of ϕ, whereas the 3×3 matrix in Eq. (C4) has a condition number just over 5. This means that the SNR for the recovered reflection coefficients is better when the phase is known.

APPENDIX D: RELATION BETWEEN NOISE WITH DIFFERENT MEASUREMENT METHODOLOGIES

The fact that we have recovered the phase of the BFP means that we can use different methods to recover the angle for SP excitation θ_p. The advantage of this is, of course, if one measurement gives a better result, we can select that methodology. Here, we will also examine situations where we may use both measurements together to get a better result. We will illustrate the ideas in simulation to explain the process and examine the correlations between different methods.

To illustrate the effect of different noise statistics, we consider the following simple model.

(1) We generate a noiseless BFP distribution.
(2) A complex random distribution is generated.

This is filtered with different cutoff frequencies to provide different noise statistics.

The variance is set so that it is independent of the cutoff frequency.

(3) The noise reduction algorithm Eq. (C5) in Appendix C is used to produce an estimate of r_p (and r_s).
(4) The value of θ_p is estimated in two different ways.

The position of the minimum dip of r_p is estimated by fitting a quadratic around the dip minimum and differentiating the fitted curve.

The complex value of r_p is used to calculate SP(z), the linear region is fitted to a straight, and the calculated gradient is equated to $\frac{4\pi n}{\lambda}\cos\theta_p$, allowing θ_p to be recovered.

(5) These simulations are repeated 1000 times to get an estimate of the variance of the recovered k-vector. We also get an estimate of the correlation coefficient between the measurements.
(6) If one measurement is superior by a large factor, we simply use this measurement. The particular measurement that performs better depends on the statistics of the noise. When the two methods give comparable noise variance, they are combined taking account of the correlation coefficient between the methods.

We consider three cases and cutoff frequencies of 125, 35, and 12 noise cycles across the BFP. These illustrative examples are shown in Figs. 12(a), 12(b), and 12(c), respectively. Essentially, the only difference between the noise in each of these examples is the cutoff frequency; the variance in each case is the same. The trends are exactly the same over a wide range of practical noise variances, and the specific case shown illustrates the general trends. The key results are tabulated in Table 3.

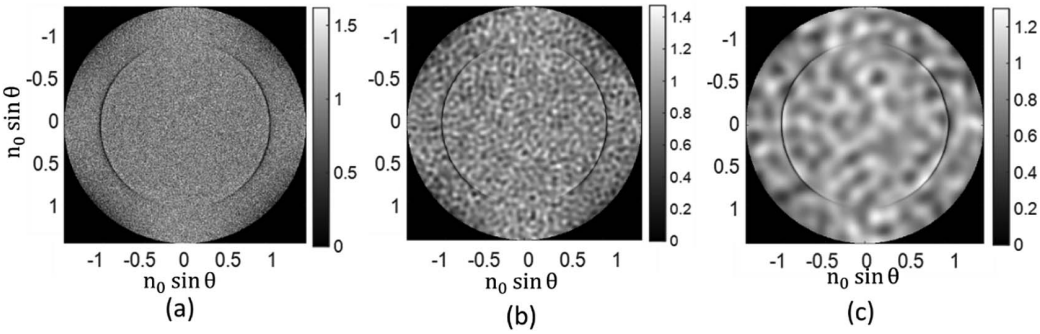

Fig. 12. (a) Simulated back focal plane with high-frequency noise case; (b) back focal plane with mid-frequency noise case; (c) back focal plane with low-frequency noise case.

Table 3. Summarizing the Noise from Different Noise Statistics in the BFP

Noise Case	Fig. 12(a)	Fig. 12(b)	Fig. 12(c)
Std of sin θ_p, dip	2.074×10^{-4}	3.203×10^{-4}	6.14×10^{-4}
Std of sin θ_p, $V(z)$	1.3×10^{-3}	2.888×10^{-4}	1.52×10^{-5}
Correlation	0.0530	0.0939	−0.0097
Dip signal proportion	>0.98	0.44	8.5×10^{-4}
Combined Std	2.062×10^{-4}	2.24×10^{-4}	1.519×10^{-5}

We can see for the high-frequency noise the measurement of the position of the dip gives a far better result compared to the $V(z)$. For the intermediate-frequency noise both methods give comparable noise, and for the low-frequency noise the $V(z)$ is clearly superior (this case seems to match our actual experimental results). For the first and last case we could just select the dip position and the $V(z)$, respectively. For the intermediate case we can combine the two measurements using the standard formulation for the sum of correlated variances:

$$\sigma_T^2 = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\rho\sqrt{\sigma_1\sigma_2}. \quad \textbf{(D1)}$$

σ_T^2 is the total variance; a is the fraction of signal 1; b is the fraction of signal 2; σ_1^2, σ_2^2 are the variances for signals 1 and 2, respectively; and ρ is the correlation between them. To get the correct mean signal it is necessary that $a + b = 1$.

Table 3 shows that for the high- and low-frequency signals the effect of the signal with higher noise can be ignored. For the intermediate case the fact that the noise from each method is quite similar and that they are only weakly correlated means that there is substantial improvement using both signals. The fact that a combination of 44% of the dip signal and 56% of the $V(z)$ comprises the optimum overall signal indicates the importance of both signals. It is also important to note that the very low correlation between the signals means that they are almost independent so can be used to efficiently improve the overall noise performance.

Funding. Shenzhen University (2019073); Science, Technology and Innovation Commission of Shenzhen Municipality (20200803150227003, KQTD20180412181324255); Natural Science Foundation of Guangdong Province (2020A1515010598); National Natural Science Foundation of China (61905147); Guangdong Provincial Pearl River Talents Program (2019JC01Y178).

Acknowledgment. Mengqi Shen led the experimental work and instrument design, analyzed the data, developed the phase retrieval code, and wrote the paper with Michael G. Somekh. Qi Zou carried out experiments and prepared the samples. Xiaoping Jiang carried out the experiments and developed the data acquisition process. Fu Feng analyzed the data and made suggestions to improve the paper. Michael G. Somekh conceived the project, developed the processing algorithms, and wrote the paper with Mengqi Shen. All authors reviewed the paper.

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

†These authors contributed equally to this work.

REFERENCES

1. X. Ou, R. Horstmeyer, G. Zheng, and C. Yang, "High numerical aperture Fourier ptychography: principle, implementation and characterization," Opt. Express **23**, 3472–3491 (2015).
2. J. Sun, C. Zuo, J. Zhang, Y. Fan, and Q. Chen, "High-speed Fourier ptychographic microscopy based on programmable annular illminations," Sci. Rep. **8**, 7669 (2018).
3. H. M. L. Faulkner and J. M. Rodenburg, "Movable aperture lensless transmission microscopy: a novel phase retrieval algorithm," Phys. Rev. Lett. **93**, 023903 (2004).
4. L. Tian and L. Waller, "3D intensity and phase imaging from light field measurements in an LED array microscope," Optica **2**, 104–111 (2015).
5. A. Ozcan and E. McLeod, "Lenless imaging and sensing," Annu. Rev. Biomed. Eng. **18**, 77–102 (2016).
6. T. W. K. Chow, B. Zhang, and M. G. Somekh, "Hilbert transform-based single-shot plasmon microscopy," Opt. Lett. **43**, 4453–4456 (2018).
7. S. Pechprasarn, B. Zhang, D. Albutt, J. Zhang, and M. Somekh, "Ultrastable embedded surface plasmon confocal interferometry," Light Sci. Appl. **3**, e187 (2014).
8. C. W. See, M. G. Somekh, and R. D. Holmes, "Scanning optical microellipsometry for pure surface profiling," Appl. Opt. **35**, 6663–6668 (1996).

9. G. D. Feke, D. P. Snow, R. D. Grober, P. J. de Groot, and L. Deck, "Interferometric back focal plane microellipsometry," Appl. Opt. **37**, 1796–1802 (1998).
10. M. Shen, T. W. K. Chow, H. Shen, and M. G. Somekh, "Virtual optics and sensing of the retrieved complex field in the back focal plane using a constrained defocus algorithm," Opt. Express **28**, 32777–32792 (2020).
11. S. Pechprasarn, T. W. K. Chow, and M. G. Somekh, "Application of confocal surface wave microscope to self-calibrated attenuation coefficient measurement by Goos-Hanchen phase shift modulation," Sci. Rep. **8**, 8547 (2018).
12. L. Allen and M. Oxley, "Phase retrieval from series of images obtained by defocus variation," Opt. Commun. **199**, 65–75 (2001).
13. A. M. S. Maallo and P. F. Almoro, "Power loss due to beam splitter cascade in the simultanous sampling of a volume speckle field for phase retrieval," Proc. SPIE **7155**, 71551J (2008).
14. P. F. Almoro, Q. D. Pham, D. I. Serrano-Garcia, S. Hasegawa, Y. Hayashi, M. Takeda, and T. Yatagai, "Enhanced intensity variation for multiple-plane phase retrieval using spatial light modulator as convenient tunable diffuser," Opt. Lett. **41**, 2161–2164 (2016).
15. H. Raether, *Surface Plasmon on Smooth and Rough Surfaces and on Gratings* (Springer-Verlag, 1986).
16. B. Zhang, S. Pechprasarn, and M. G. Somekh, "Quantitative plasmonic measurements using embedded phase stepping confocal interferometry," Opt. Express **21**, 11523–11535 (2013).
17. J. Greivenkamp, "Generalized data reduction for heterodyne interferometry," Opt. Eng. **23**, 350–352 (1984).
18. M. Shen, S. Learkthanakhachon, S. Pechprasarn, Y. Zhang, and M. G. Somekh, "Adjustable microscopic measurement of nanogap waveguide and plasmonic structures," Appl. Opt. **57**, 3453–3462 (2018).