Analysis of numerical diffraction calculation methods: from the perspective of phase space optics and the sampling theorem

Article *in* Journal of the Optical Society of America A · October 2020

DOI: 10.1364/JOSAA.401908

CITATION
1

READS
134

4 authors, including:

Wenhui Zhang
Tsinghua University
32 PUBLICATIONS **79** CITATIONS
SEE PROFILE

Hao Zhang
Tsinghua University
74 PUBLICATIONS **1,423** CITATIONS
SEE PROFILE

Colin Sheppard
University of Wollongong
714 PUBLICATIONS **18,043** CITATIONS
SEE PROFILE

Journal of the Optical Society of America A
OPTICS, IMAGE SCIENCE, AND VISION

Analysis of numerical diffraction calculation methods: from the perspective of phase space optics and the sampling theorem

WENHUI ZHANG,[1] HAO ZHANG,[1,*] COLIN J. R. SHEPPARD,[2,3] AND GUOFAN JIN[1]

[1]State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instrument, Tsinghua University, Beijing 100084, China
[2]School of Chemistry, University of Wollongong, Wollongong, NSW 2522, Australia
[3]Nanoscopy and NIC@IIT, Istituto Italiano di Tecnologia, Via Enrico Melen, 83 Edificio B, 16152 Genova, Italy
*Corresponding author: haozhang274@tsinghua.edu.cn

Received 2 July 2020; revised 21 September 2020; accepted 21 September 2020; posted 22 September 2020 (Doc. ID 401908); published 13 October 2020

Diffraction calculations are widely used in applications that require numerical simulation of optical wave propagation. Different numerical diffraction calculation methods have their own transform and sampling properties. In this study, we provide a unified analysis where five popular fast diffraction calculation methods are analyzed from the perspective of phase space optics and the sampling theorem: single fast Fourier transform-based Fresnel transform, Fresnel transfer function approach, Fresnel impulse response approach, angular spectrum method, and Rayleigh–Sommerfeld convolution. The evolutions of an input signal's space-bandwidth product (SBP) during wave propagation are illustrated with the help of a phase space diagram (PSD) and an ABCD matrix. It is demonstrated that all of the above methods cannot make full use of the SBP of the input signal after diffraction; and some transform properties have been ignored. Each method has its own restrictions and applicable range. The reason why different methods have different applicable ranges is explained with physical models. After comprehensively studying and comparing the effect on the SBP and sampling properties of these methods, suggestions are given for choosing the proper method for different applications and overcoming the restrictions of corresponding methods. The PSD and ABCD matrix are used to illustrate the properties of these methods intuitively. Numerical results are presented to verify the analysis, and potential ways to develop new diffraction calculation methods are also discussed.

https://doi.org/10.1364/JOSAA.401908

1. INTRODUCTION

Diffraction is an important feature of optical waves. It is challenging and essential to understand the diffraction phenomenon in physical optics [1,2]. In the past two centuries, optical diffraction has been studied by a large number of scientists, and many theories have been developed to help people understand and calculate the diffraction fields [1–5]. Based on either the Rayleigh–Sommerfeld theory or angular spectrum theory, both strict solutions to the free-space wave equation, the diffraction field is an integral transform of the input aperture [1]. However, only a few simple apertures have their diffraction fields in an analytical form. Therefore, in most cases, the diffraction field can be obtained only by means of discrete computation. In such cases, it is essential to correctly sample the input field and the transform kernel to avoid aliasing errors. Meanwhile, it should be guaranteed that the discrete operation does not change the true diffraction field predicted by analytical theories.

Recently, with the rise of computational imaging, which relies directly on numerical diffraction calculation to reconstruct images [6–9], to understand and utilize the numerical diffraction calculation well becomes more and more important. In addition, numerical diffraction calculation is also the key process for digital holography [10–12], computer holography [13,14], beam shaping [15,16], and diffraction tomography [17,18], etc. Due to the importance of this topic, numerical diffraction calculation has already been extensively studied and discussed in different ways. Voelz *et al*. analyzed scalar optical diffraction from the view of chirp function sampling [7,19]. An ideal sampling distance z_c was defined; and it was suggested that the impulse response method was better when propagation distance $z > z_c$, and the transfer function method was better when $z < z_c$. Kelly has also concluded similar remarks for the numerical calculation analysis of the Fresnel transform [20]. Goodman added a new chapter titled "Computational

diffraction and propagation" to discuss the sampling issue of different numerical diffraction methods in the fourth edition of *Introduction to Fourier Optics* [2].

The task of numerical diffraction calculation is to quickly obtain the accurate diffraction field. Besides, because of the use of fast Fourier transform (FFT) algorithm, the sampling pitch of the diffraction field is fixed by the calculation parameters. Thus, to flexibly adjust the sampling pitch is also a demand. Therefore, high calculation accuracy, high calculation efficiency, and high calculation flexibility are our goals. Unfortunately, these three goals are contradictory. Generally, massive computing resources are required to get a high calculation accuracy, which would sacrifice the calculation efficiency. The ordinary FFT cannot be directly used if high calculation flexibility is desired, which would also decrease the calculation efficiency. During the last two decades, to get a high calculation accuracy or flexibility without decreasing the efficiency, some techniques have been proposed for different numerical diffraction calculation methods.

Aiming to overcome the low calculation accuracy of the angular spectrum method (ASM) in the far field, Matsushima *et al*. proposed band-limited ASM [21]; Kim *et al*. proposed wide-range ASM [22], and we proposed band-extended ASM [23]. Both of the first two methods use only the nonaliased transfer function of ASM to calculate the diffraction field, and the last method extends the effective bandwidth of the nonaliased transfer function. By these methods, calculation errors caused by an aliased transfer function can be avoided to get a high accuracy. Aiming to overcome the fixed sampling pitch in ASM, Shimobaba *et al*. proposed scaled ASM to adjust the sampling pitch of the diffraction field by nonuniform FFT [24]. Inspired by the single FFT-based Fresnel transform, Tomasz *et al*. proposed a compact space-bandwidth product (SBP) method to make the sampling pitch of the diffraction field vary with propagation distance. In this way the calculation accuracy of ASM is high in the large propagation distance range, while the flexibility of adjusting the sampling pitch is limited because it is mainly determined by the propagation distance [25]. Note that high accuracy in [25] is determined only for the diffraction amplitude, and the phase is not considered. Aiming to overcome the fixed sampling pitch in the single FFT-based Fresnel transform, which is determined by propagation distance, Zhang *et al*. proposed two-step Fresnel propagation to adjust the sampling pitch according to the ratio between these two propagation distances [26]. Scaled Fresnel diffractions were used in computer holography and digital holography to implement scale operations on the diffraction field based on chirp-z transform [27,28]. Due to aliasing errors in the scaled Fresnel method, the calculation accuracy would decrease in the near field. For this problem, Shimobaba *et al*. used a band-limited function to reduce the aliasing errors for high accuracy [29]. Other concerns in numerical diffraction calculation, such as oblique illumination [30], shifted observation window [31], propagation between tilt planes [32], and high-order dispersion caused by phase aliasing [33], have also been studied. In addition, sampling in the fractional Fourier domain [34–36] and phase space [37] has been studied. By observing the Fresnel diffraction field at a reference spherical surface with fractional Fourier transform, a natural sampling grid, both transverse and longitudinal, was defined to express the structure of Fresnel diffraction. Based on a phase space analysis, a generalized sampling method was proposed [38]. Such a method could realize a resolution enhancement digital holography at the cost of large detector size [39], which is similar to the sampling analysis in [40] when the object is space limited. All these mentioned methods have successfully solved one problem in numerical diffraction calculation and pushed the research forward. However, there lacks a general analysis of the existing commonly used diffraction calculation methods to answer the questions about the properties of different methods, how to choose one proper method for different applications, and how to realize a generalized method which can efficiently, accurately, and flexibly calculate the diffraction field. This unified analysis is essential and would help establish a generalized framework for understanding the diffraction calculation in the numerical diffraction calculation community and related research areas.

In this work, we have comparatively analyzed five popular diffraction calculation methods: single FFT-based Fresnel transform, Fresnel transfer function (Fres-TF) approach, Fresnel impulse response (Fres-IR) approach, ASM, and Rayleigh–Sommerfeld convolution (RSC). The effect on the signal's SBP and the sampling properties of each method have been studied in detail through the phase space diagram (PSD) and the sampling theorem. Compared with the analysis in single-space domain or single-frequency domain, joint analysis in space-frequency domain with PSD is a more effective method because it gives a complete signal status. Such a unified analysis provides a new perspective: to study whether the mathematics of diffraction calculation methods (examined by the sampling theorem) is consistent with the physics (represented in phase space). With the results obtained, we provide solutions to overcome the restrictions of each method, and give suggestions on how to choose the proper method, or adapt one, for different applications. Finally, we suggest an outlook on development of new diffraction methods.

This paper is organized as following: Section 1 is the introduction and motivation of this work; Section 2 introduces the PSD and ABCD matrix that are the tools used in our analysis; Section 3 analyzes five methods in detail; Section 4 provides solutions to overcome the restrictions of each method and gives suggestions on how to choose or adapt one method for specific applications; and Section 5 summarizes this work and presents an outlook for new diffraction calculation methods.

2. PSD AND ABCD MATRIX

A. PSD

An optical signal can be viewed in the space domain (x) or the spatial frequency domain (f_x). If it is viewed simultaneously from these two domains (x, f_x), it is represented in a phase space. Analytically, the PSD of a signal (which we refer as the SBP distribution in phase space) can be given by the Wigner distribution function (WDF).

For a one-dimensional signal $u(x)$, its WDF can be expressed as [41,42]

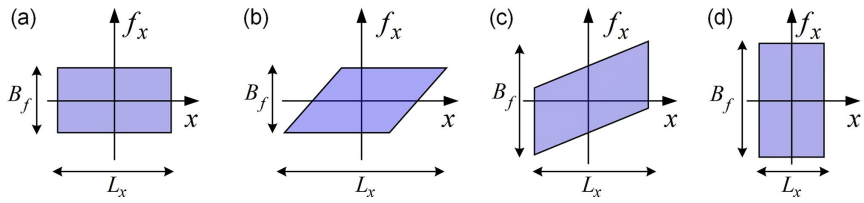

Fig. 1. (a) PSD of a discretized signal $u(n\Delta_x)$; (b) PSD after paraxial Fresnel propagation; (c) PSD after chirp modulation; (d) PSD after Fourier transform.

$$W_u(x,\, f_x) = \int_{-\infty}^{+\infty} u\left(x + \frac{x'}{2}\right) u^*\left(x - \frac{x'}{2}\right) \times \exp(-i2\pi\, f_x x')\mathrm{d}x', \tag{1}$$

where $*$ is a complex conjugation. Because WDF doubles the number of the dimensions of the input signal, a one-dimensional signal has a two-dimensional WDF, and a two-dimensional signal has a four-dimensional WDF that cannot be illustrated intuitively in three-dimensional space. Therefore, we take a one-dimensional signal for our example to conduct the analysis, it being easy to extend to the two-dimensional case. WDF is a powerful analytical tool; however, it is not intuitive. Therefore, we use PSD to represent the signal's SBP distribution in phase space, as shown in Fig. 1. Note that PSD used here is a joint space-spatial frequency diagram to intuitively show how the SBP distributes, which is more like a representation in the physical scene. As shown in Fig. 1, the SBP distribution of a signal is a specific area in the space-spatial frequency domain $(x,\, f_x)$, which is defined by the location (x) and by the range of spatial frequency (f_x) within which the signal is nonzero [41].

In numerical diffraction calculations, the signal $u(x)$ is uniformly sampled as $u(n\Delta_x)$ where $n = -N/2, -N/2+1, \ldots 0, \ldots N/2-1$, with N being the sampling number (assumed to be even) and Δ_x being the sampling pitch. The discrete sampling operation should obey the sampling theorem that the sampling rate should be no less than twice the signal's highest frequency. In this way, the signal is bounded within a finite effective region in both the space domain and the spatial frequency domain. Therefore, the PSD of the discretized signal $u(n\Delta_x)$ is a rectangle, as shown in Fig. 1(a). The side length is $L_x = N\Delta_x$ and $B_f = 1/\Delta_x$ along the x axis and f_x axis, respectively. The enclosed area represents the maximum SBP value of the signal, which is $L_x B_f = N$. Signal energy is negligible outside this region. Note that the PSD of discrete signals is periodic [37], but here we only consider one period located at the origin. As long as there is no overlap between neighboring copies, this operation is valid. The request of no overlapping is guaranteed by the sampling theorem, which is used to analyze the sampling request of different methods.

B. ABCD Matrix

As introduced above, viewing the signal in phase space is convenient. However, it cannot tell us how the signal changes after diffraction. Fortunately, the ABCD matrix is a mathematical form that can be used to analyze the evolution properties of the signal during paraxial propagation [43,44]. Collins has written the lens-system diffraction integral under the paraxial approximation in terms of matrix optics [45]. And Tan *et al.* have proven that the ABCD matrix method can be extended into the nonparaxial regime [46]. Therefore, the PSD and the ABCD matrix can be perfectly combined to analyze how different calculation methods affect the signal's SBP and how to satisfy their sampling requirements.

Diffraction propagation can be regarded as an operator P on the signal. Take paraxial Fresnel diffraction as example and we can get the diffraction field,

$$\tilde{u}(X) = P\{u(x)\} = \frac{\exp(i2\pi z/\lambda)}{\sqrt{i\lambda z}} \times \int u(x) \exp\left(\frac{i\pi}{\lambda z}(X-x)^2\right) \mathrm{d}x, \tag{2}$$

where λ is the wavelength and z is the propagation distance. The effect of the operator P on WDF is

$$W(x,\, f_x) \to W(ax + bf_x,\, cx + df_x) = W(X,\, f_X), \tag{3}$$

where

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{4}$$

is the ABCD matrix. Equation (3) can be rewritten as a matrix multiplication,

$$\begin{bmatrix} X \\ f_X \end{bmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{bmatrix} x \\ f_x \end{bmatrix}. \tag{5}$$

Equation (5) tells us the new position $(X,\, f_X)$ of the point in the phase space after propagation from a point that is originally located at $(x,\, f_x)$. The parameters in matrix M are related to the transform property of P. For paraxial Fresnel diffraction [42,45],

$$M = \begin{pmatrix} 1 & \lambda z \\ 0 & 1 \end{pmatrix}. \tag{6}$$

The PSD of the signal after paraxial Fresnel propagation is shown in Fig. 1(b). As we can see, a horizontal shearing of the PSD along the x axis is caused in this case. For the nonparaxial case, the PSD behaves quite differently. A spherical aberration is introduced, and therefore, the shearing of the rectangle is not a simple shear, but the sides of the figure become curved [47–49].

Furthermore, we deal with four different operations in Section 3: (1) chirp modulation of the signal, $u(x)\exp(i\pi x^2/\lambda z)$; (2) fast Fourier transformation of the

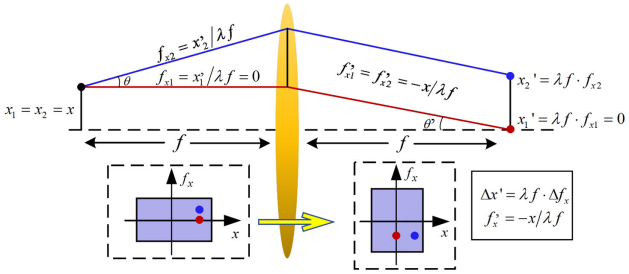

Fig. 2. Diagram of the optical Fourier transform.

signal, FFT$\{u(x)\}$; (3) chirp modulation of the Fourier-transformed signal, $U(f_x)\exp(-i\pi\lambda z f_x^2)$; and (4) optical Fourier transformation of the signal, OFT$\{u(x)\}$. The corresponding ABCD matrices of these operations are given below:

$$M_{(1)} = \begin{pmatrix} 1 & 0 \\ 1/\lambda z & 1 \end{pmatrix}, \quad M_{(2)} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

$$M_{(3)} = \begin{pmatrix} 1 & 0 \\ -\lambda z & 1 \end{pmatrix}, \quad M_{(4)} = \begin{pmatrix} 0 & \lambda z \\ -1/\lambda z & 0 \end{pmatrix}. \quad \textbf{(7)}$$

There are two points requiring attention. First, FFT is a fast algorithm for implementing Fourier transform and should be applied to discrete variables. For simplicity, we use $u(x)$ to represent its discrete version $u(n\Delta_x)$, and FFT is applied to the dimensionless array u_n at the interval of Δ_x. This convention also applies to other variables when FFT is used in the paper. Second, the so-called optical Fourier transform is different from FFT but implemented by FFT. For a signal $u(x)$, we refer to $\int u(x)\exp[-i2\pi Xx/(\lambda z)]\mathrm{d}x$ as its optical Fourier transform [42]. Therefore, operation (2) and operation (4) have different ABCD matrices. FFT is a purely mathematical or algorithmic transformation, and it has no relationship with the physical propagation distance z. In contrast, the optical Fourier transform has physical meaning, as explained and associated with Fig. 2.

Two rays indicated by the red line and the blue line have the same spatial location $x_1 = x_2 = x$ at the front focal plane of the lens. Their spatial frequencies are different: the red line is parallel to the optical axis having $f_{x1} = 0$, and there is an angle θ between the blue line and the optical axis giving $f_{x2} = \sin\theta/\lambda \approx \tan\theta/\lambda = x_2'/\lambda f$. The approximation is valid in paraxial optics. Based on geometrical optics, these two rays would be parallel after passing through the lens and have the same spatial frequency $f_{x1}' = f_{x2}' \approx \tan\theta'/\lambda = -x/\lambda f$. Their spatial locations at the rear focal plane of the lens can be easily obtained as $x_1' = \lambda f \cdot f_{x1} = 0$ and $x_2' = \lambda f \cdot f_{x2}$. The above analysis can be intuitively viewed by the change of PSD shown in Fig. 2. It is evident that the optical Fourier transform not only rotates the PSD but scales the rotated PSD with a factor λf. Therefore, the optical Fourier transform can be regarded as the matrix multiplication of a FFT and a scaling matrix,

$$\begin{pmatrix} 0 & \lambda f \\ -1/\lambda f & 0 \end{pmatrix} = \begin{pmatrix} \lambda f & 0 \\ 0 & 1/\lambda f \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \quad \textbf{(8)}$$

Changing f to z, we can get the ABCD matrix of the optical Fourier transform in free-space diffraction, as given in Eq. (7). The corresponding PSDs of operation (1) and operation (2) are shown in Figs. 1(c) and 1(d), respectively. It is shown that chirp modulation of the signal $u(x)$ causes a vertical shearing of the PSD along the f_x axis, and FFT causes a rotation of PSD by $\pi/2$ rad [41]. Note that these PSDs are determined in two steps: find the corner points' coordinates after transformation, and then connect them in order by straight lines.

3. NUMERICAL DIFFRACTION CALCULATION ANALYSIS

In this section, we analyze five popular numerical diffraction methods: three paraxial Fresnel diffraction methods, which are single FFT-based Fresnel transform, Fres-TF, Fres-IR, and two rigorous diffraction methods without paraxial approximation, which are ASM and RSC. As introduced above, the goal of all the methods is to quickly get the accurate diffraction field $\tilde{u}(X)$ in the destination plane of the input signal $u(x)$, as shown in Fig. 3. The calculation flexibility will be discussed in Section 4. Before getting into the calculation, we would like to introduce a little about the diffraction theory used. The first three methods are based on the Fresnel approximation, where a spherical wave is approximated by a parabolic wave. According to how the approximation is applied, single FFT-based Fresnel transform is most suitable for calculation of a diffracted field over a spherical (or parabolic) surface [50]. RSC is based on the first Rayleigh–Sommerfeld integral (RSI) theory, and ASM is its corresponding Fourier form [2]. Please note that some methods have been proposed to improve the accuracy of the Fresnel approximation [51], and to model the Rayleigh–Sommerfeld diffraction in k space [52], but here we mainly focus on the commonly used calculation methods.

Given the input signal $u(x)$ with the sampling number N and sampling pitch Δ_x, it is necessary to calculate the spatial range L_X of the diffraction field $\tilde{u}(X)$. Only inside the range $[-L_X/2, L_X/2]$, is it possible to get an accurate result; otherwise there is no physical meaning because L_X denotes the reach of diffracted light. According to the simple geometry shown in Fig. 3, it is easy to calculate such a range as $L_X = L_x + 2z\tan\alpha$ with $\alpha = \arcsin[\lambda/(2\Delta_x)]$. From Fig. 3, we can see that the signal $u(x)$ would spread over a larger spatial range after free-space diffraction, which can be also observed by comparing Figs. 1(a) and 1(b). However, the methods we introduce below cannot make full use of the spreading SBP without any extra cost—for example, computational complexity.

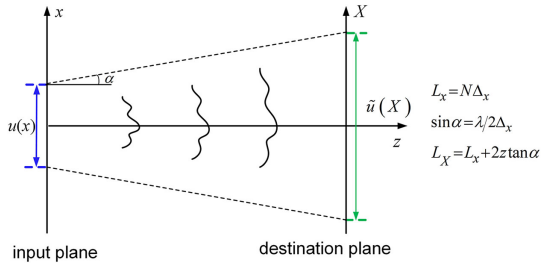

Fig. 3. Diffraction scenario between two planes in the free space.

In single FFT-based Fresnel transform, the observation window changes with the propagation distance because the sampling pitch of the diffraction field changes. The other four methods are all convolution-based, and the observation window is the same size as the input window. The difference among the convolution-based methods is that the transfer functions are directly sampled in the spatial frequency domain in Fres-TF and ASM, while in Fres-IR and RSC, the spatial impulse response functions are first sampled and then Fourier transformed to get the transfer function. In analytical form, the transfer function and the impulse response function are equivalent. But in discrete form, they have different sampling requirements, as analyzed below. Since sampling is essential during discretization, here we would like to mention a little about the sampling theorem in the numerical diffraction calculation. There are integral kernels in the diffraction calculations, and these kernels should be sampled correctly during discretization. The principle is that the sampling rate should be no less than twice the maximum local spatial frequency of the kernels' phases. For different methods, the kernels are different, and we analyze them one by one in the following parts of this paper.

Let us start with the single FFT-based Fresnel transform, since it is also called the "direct method" [19].

A. Single FFT-Based Fresnel Transform

Equation (2), describing paraxial Fresnel propagation, can be rewritten as

$$\tilde{u}(X)=\frac{\exp(i2\pi z/\lambda)}{\sqrt{i\lambda z}}\exp\left(\frac{i\pi}{\lambda z}X^2\right)\int\left[u(x)\exp\left(\frac{i\pi}{\lambda z}x^2\right)\right]\times\exp\left(-i2\pi\frac{X}{\lambda z}x\right)\mathrm{d}x. \tag{9}$$

Ignoring the constant term, we can see that there are three steps to get the diffraction field $\tilde{u}(X)$ from the input signal $u(x)$: (1) chirp modulation of $u(x)$; (2) (optically) Fourier transform the chirp-modulated signal; and (3) chirp modulation of the Fourier transformed signal. Step (2) is implemented by a FFT in practice by regarding $f_X=X/\lambda z$, but it has physical meaning. Please note that we call the chirp function in (1) the inner chirp function and that in (3) the outer chirp function. If we use the ABCD matrix to model these steps, we can get

$$\begin{bmatrix}X\\ f_X\end{bmatrix}=\begin{pmatrix}1 & 0\\ 1/\lambda z & 1\end{pmatrix}\begin{pmatrix}0 & \lambda z\\ -1/\lambda z & 0\end{pmatrix}\begin{pmatrix}1 & 0\\ 1/\lambda z & 1\end{pmatrix}\begin{bmatrix}x\\ f_x\end{bmatrix}. \tag{10}$$

PSDs after each step are shown in Fig. 4 to help find which step needs attention in sampling to avoid aliasing errors. The first step is the chirp modulation of the signal $u(x)$. As we can see, the bandwidth becomes larger and the spatial length holds constant, which leads to a larger sampling SBP. By the sampling SBP, we refer to the area of the circumscribed rectangle of the parallelogram. To make sure that the chirp-modulated signal is correct, the chirp function has to be sampled correctly, which is

$$\left|\frac{1}{2\pi}\frac{\partial\varphi}{\partial x}\right|_{\max}\le\frac{1}{2\Delta_x}, \tag{11}$$

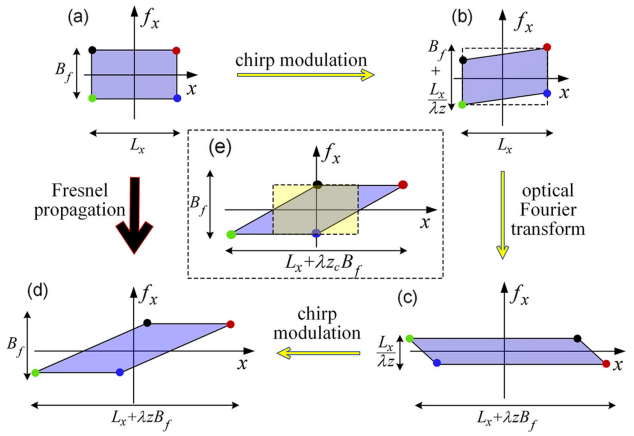

Fig. 4. PSD of the signal before (a) and after (b)–(d) transformations. PSD after $z=z_c$ diffraction propagation is shown in (e).

where $\varphi=\pi x^2/\lambda z$ is the phase of the inner chirp function. From inequality (11) we can get

$$z\ge\frac{2\Delta_x}{\lambda}|x|_{\max}. \tag{12}$$

Substituting $|x|_{\max}=L_x/2=N\Delta_x/2$, we get

$$z\ge\frac{N\Delta_x^2}{\lambda}. \tag{13}$$

This means that the inner chirp function can be only sampled correctly within this distance range, and the propagation distance should be as large as possible to satisfy its sampling requirement. This conclusion is also consistent with the trend of signal bandwidth after diffraction: the larger the distance z, the smaller the increment of bandwidth $L_x/\lambda z$, the smaller the sampling SBP, and the easier to satisfy its sampling requirement. In conventional methods, it is thought that as long as the inner chirp function is sampled without aliasing, the optical Fourier transform can be operated correctly because it satisfies the sampling SBP. However, it is not rigorous but, nevertheless, a correct estimate. As we stated above, the chirp modulation (multiplication with a chirp function) would cause a bandwidth increase, as shown in Fig. 4(b). That is to say, the product has a larger bandwidth. Therefore, the sampling theorem should be applied to the product rather than only the chirp function. For small-bandwidth signal and a small-bandwidth increase, to consider the chirp function separately turns out to provide a correct estimate.

The other operation needing attention is the outer chirp modulation to the Fourier transformed signal. Similarly, we find the sampling requirement by solving

$$\left|\frac{1}{2\pi}\frac{\partial\Phi}{\partial X}\right|_{\max}\le\frac{1}{2\Delta_X}, \tag{14}$$

where $\Phi=\pi X^2/\lambda z$ is the phase of the outer chirp function. Before solving the inequality for z, we need to calculate the sampling pitch Δ_X of the diffraction field. As shown in Fig. 4(d), the spatial length of the diffraction field is

$L_X = L_x + \lambda z B_f = N\Delta_x + \lambda z/\Delta_x$. Therefore, we have $\Delta_X = L_X/N = \Delta_x + \lambda z/N\Delta_x$ and $|X|_{\max} = L_X/2 = N\Delta_x/2 + \lambda z/2\Delta_x$. Substituting the expressions of Δ_X and $|X|_{\max}$ into inequality (14), we have

$$\frac{\lambda^2}{N\Delta_x^2} z^2 + \lambda z + N\Delta_x^2 \le 0. \tag{15}$$

There is no meaningful solution to inequality (15) because $\lambda^2 - 4 \cdot (\lambda^2/N\Delta_x^2) \cdot N\Delta_x^2 < 0$, which means the sampling requirement of the outer chirp function over the whole diffraction field cannot be satisfied. Fortunately, due to the property of FFT, the actual spatial length of $\tilde{u}(X)$ is $L_X = \lambda z B_f = \lambda z/\Delta_x$ [7]. In this situation, the new sampling pitch of the diffraction field is $\Delta_X = L_X/N = \lambda z/N\Delta_x$. Substituting the new expressions of Δ_X and $|X|_{\max}$ into inequality (14), we have

$$z \le \frac{N\Delta_x^2}{\lambda}. \tag{16}$$

This means that the outer chirp function can be only sampled correctly within this distance range, and the propagation distance should be as small as possible to satisfy its sampling requirement. This conclusion is also consistent with the trend of spatial length of the signal after diffraction: the smaller the distance z, the smaller the increment of spatial length $\lambda z B_f$, the smaller the sampling SBP, and the easier to satisfy its sampling requirement.

Combining inequality (13) and (16), only one propagation distance $z_c = N\Delta_x^2/\lambda$ can simultaneously satisfy the sampling requirements of these two chirp functions, which severely limits the application of the single FFT-based Fresnel transform. Defining the Fresnel number $N_F = (N\Delta_x/2)^2/\lambda z$, $z = z_c$ is equivalent to $N_F = 0.25$. Therefore, N_F can be also used to express the distance range. In this paper, we directly give the distance range, and it can be translated to N_F if needed. At $z = z_c$, we can calculate the spatial length of the diffraction field and its bandwidth, which are $L_X(z = z_c) = \lambda \cdot (N\Delta_x^2/\lambda)/\Delta_x = N\Delta_x$ and $B_f(z = z_c) = 1/\Delta_X = N\Delta_x/(\lambda \cdot N\Delta_x^2/\lambda) = 1/\Delta_x$. The PSD of the diffraction field in this case is shown in Fig. 4(e). The yellow rectangle surrounded by the dashed lines is the calculated SBP inside the range $L_X(z = z_c) \cdot B_f(z = z_c)$, while the effective SBP after propagation is inside the parallelogram. Therefore, the effective SBP is not fully obtained with the original N sampling points.

By revisiting these two chirp functions, it is easy to notice that the outer chirp function only affects the phase distribution of the diffraction field; therefore, if only amplitude (or intensity) is of concern, one only needs to satisfy inequality (13).

To verify the above analysis, the diffraction field of a rectangle aperture, $u(x) = \mathrm{rect}(x/l)$, where $l = 2.5$ mm is the width of the aperture, is calculated by single FFT-based Fresnel transform under different propagation distances. The parameters of this calculation are listed in Table 1. With these parameters, $z_c = N\Delta_x^2/\lambda = 50$ mm, and we set two propagation distances as $z_1 = 50$ mm and $z_2 = 100$ mm.

Figure 5 shows the results. Figures 5(a) and 5(b) show the diffraction amplitude and phase distributions with $z_1 = 50$ mm, respectively. Figure 5(c) shows the unwrapped phase distribution of the outer chirp function. The results with $z_2 = 100$ mm are shown in Figs. 5(d)–5(f) in the same order. In addition, the RSI results are also presented as reference. RSI is calculated by the point-by-point integral without fast calculation algorithms. Therefore, it can be used as a reference as long as the sampling of the input signal and diffraction field are correct.

Table 1. Parameters Used for Single FFT-Based Fresnel Transform

Parameters	Values
Sampling number of the signal	$N = 1000$
Sampling pitch of the signal	$\Delta_x = 5$ μm
Wavelength	$\lambda = 0.5$ μm

From Fig. 5, we can see that when the propagation distance $z_1 = z_c$, the amplitude and phase results given by the single FFT-based Fresnel transform are correct. The little difference between the results calculated by single FFT-based Fresnel transform and RSI is because there is no bandwidth limitation in RSI, while there is in the other. However, when the propagation distance $z_2 > z_c$, the phase of the diffraction field is no longer correct, while the amplitude is still correct, which is consistent with our analysis. By further analyzing the phase profile, we notice that it is correct for small X values, as indicated by the red rectangle in Fig. 5(g). That is to say, the sampling of the outer chirp in these parts is correct, which is consistent with the observation from Fig. 5(h), where the inner part is correct and the outer parts are aliased. We could calculate this effective region by solving inequality (14) but regarding X as the variable,

$$|X_{\text{effective}}| \le \frac{N\Delta_x}{2}. \tag{17}$$

This means that the calculated diffraction field is correct within this region. The small difference with that calculated by the RSI is due to the limited bandwidth of the single FFT-based Fresnel transform and approximation in the sampling process. Of course, one can do zero padding to the outer chirp to make the whole diffraction field correct, which will be discussed in Section 4.

In the single FFT-based Fresnel transform, the size of the diffraction field is automatically scaled with the propagation distance in the form of $L_X = \lambda z/\Delta_x$, not as indicated in Fig. 4(d) $L_X = N\Delta_x + \lambda z/\Delta_x$, because of the property of FFT. Besides, this method does not require zero padding to the input signal, which is needed in the convolution-based methods, to avoid circular convolution errors. However, it is only applicable at one specific propagation distance to ensure the correctness of the calculated diffraction field, if the phase is required, and this correctness depends on the sampling. Actually, zero padding can make more spatial frequency components transferred in this method because $f_X = X/\lambda z$ would increase with X.

B. Fres-TF and Fres-IR

There is another understanding of Eq. (2): that the diffraction field $\tilde{u}(X)$ is the linear convolution of the input signal and the convolution kernel,

$$h(X) = \frac{\exp(i2\pi z/\lambda)}{\sqrt{i\lambda z}} \exp\left(\frac{i\pi X^2}{\lambda z}\right), \tag{18}$$

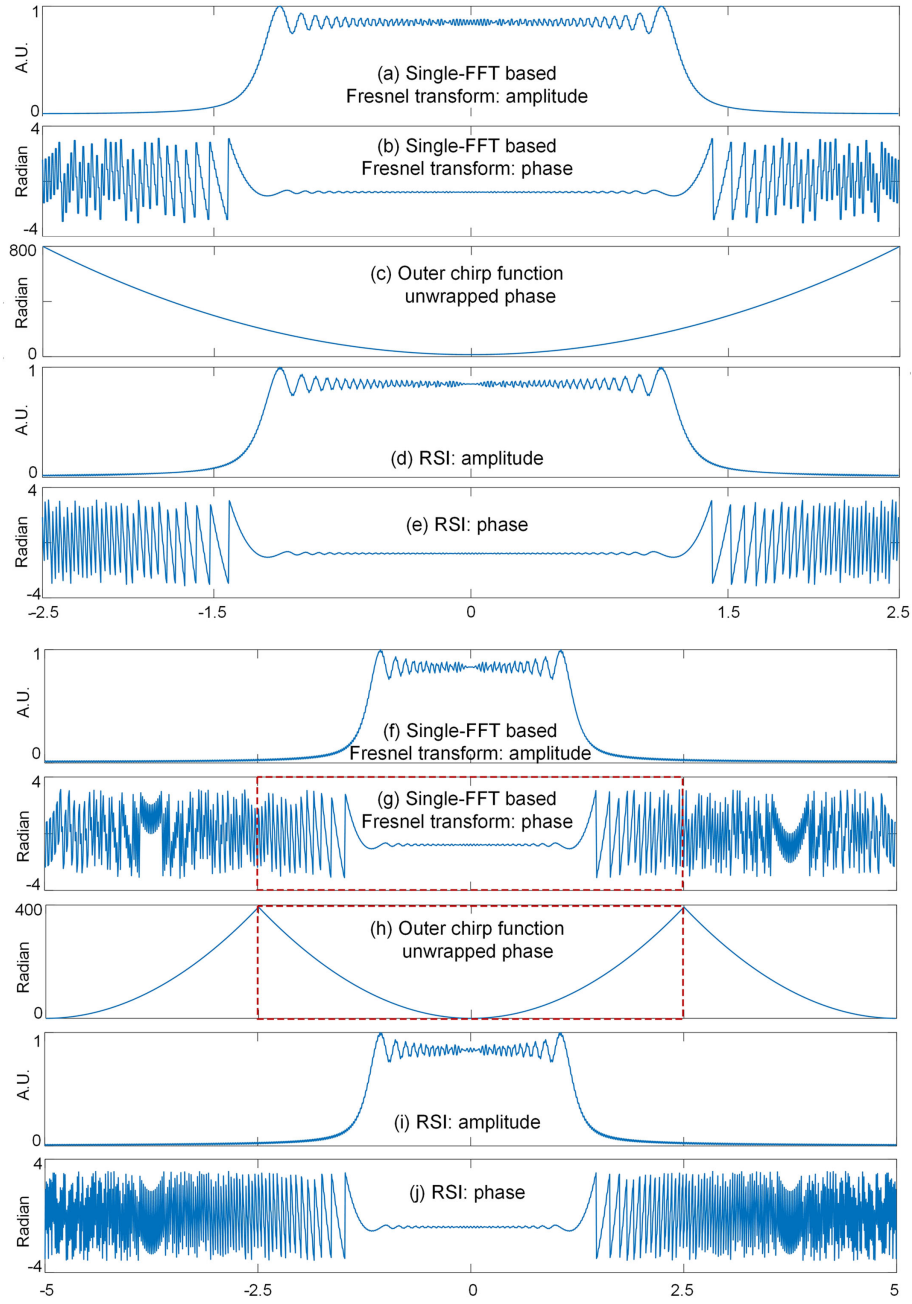

Fig. 5. Diffraction fields calculated by the single FFT-based Fresnel transform and RSI with (a)–(e) $z_1 = 50$ mm; and (f)–(j) $z_2 = 100$ mm . The unwrapped phases of the outer chirp function [(c) and (h)] are also shown in these two cases to illustrate its correct part.

which is modeled as

$$\tilde{u}(X) = \int u(x)h(X-x)\mathrm{d}x. \tag{19}$$

Usually, Eq. (19) is represented by a convolution symbol as

$$\tilde{u}(X) = u(X) \otimes h(X), \tag{20}$$

where $\otimes$ denotes linear convolution. Please note that the input signal is expressed as $u(X)$ here due to the mathematical conventions on the convolution formula. Actually, x and X are just coordinate symbols, used to indicate the spatial locations of the signal before and after diffraction propagation; in the convolution-based methods, x and X are located at the same positions because $\Delta_X = \Delta_x$. To get fast calculation of Eq. (20), based on the convolution theorem, multiplication of the Fourier transform of $u(X)$ and the Fourier transform of $h(X)$ is used, rather the convolution because the computational complexity of multiplication is much cheaper than that of convolution. We can obtain the analytic expression for the Fourier transform of $h(X)$ which is

$$H(f_X) = \frac{\exp(i2\pi z/\lambda)}{\sqrt{i\lambda z}} \exp(-i\pi\lambda z f_X^2), \tag{21}$$

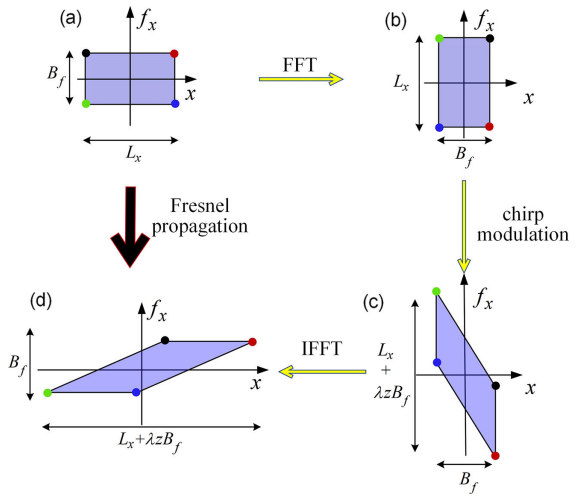

Fig. 6. PSDs of the signal (a) before and (b)–(d) after transformations.

where f_X is the spatial frequency. Mathematically, $h(X)$ and $H(f_X)$ are equivalent to model the diffraction. However, they have totally different sampling properties during discrete calculation. Based on how the transfer function is modeled, the convolution-based paraxial Fresnel diffraction can be divided into two approaches: Fres-TF and Fres-IR. In Fres-TF, the transfer function is directly modeled by $H(f_X)$, while in Fres-IR, the transfer function is obtained as the Fourier transform of $h(X)$. In these two approaches, there are also three steps from $u(X)$ to $\tilde{u}(X)$: (1) Fourier transform $u(X)$ to get $U(f_X)$ by FFT; (2) obtain $A(f_X) = U(f_X) \cdot H(f_X)$ or $A(f_X) = U(f_X) \cdot \mathrm{FFT}\{h(X)\}$; (3) inverse Fourier transform $A(f_X)$ to get the diffraction field $\tilde{u}(X)$ by IFFT,

$$\tilde{u}(X) = \mathrm{IFFT}\{A(f_X)\}. \tag{22}$$

If we use the ABCD matrix to model these steps, we can get

$$\begin{bmatrix} X \\ f_X \end{bmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\lambda z & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{bmatrix} x \\ f_x \end{bmatrix}. \tag{23}$$

Similarly, we draw the PSDs after each step.

Note that in Fig. 6, L_x and B_f are only values without units. The units are given by the domains (space or spatial frequency) where they are drawn. Because FFT does not change the sampling SBP, there is only one step needing attention to ensure correct sampling: the sampling of $H(f_X)$ in Fres-TF and $h(X)$ in Fres-IR.

Before discussing the sampling of $H(f_X)$ or $h(X)$, it is necessary to discuss the difference between circular convolution and linear convolution, since the convolution theorem for discrete signals is based on circular convolution. The convolution theorem states that the convolution of two signals is equivalent to the inverse Fourier transform of the product of their Fourier transforms. In discrete calculation, the convolution is circular convolution. However, Eq. (19) is a linear convolution. Therefore, to make the convolution theorem suitable for Eq. (19), zero padding to $u(X)$ and $h(X)$ is required [53] and the number of padded zeros is $N-1$, at least. Suppose the

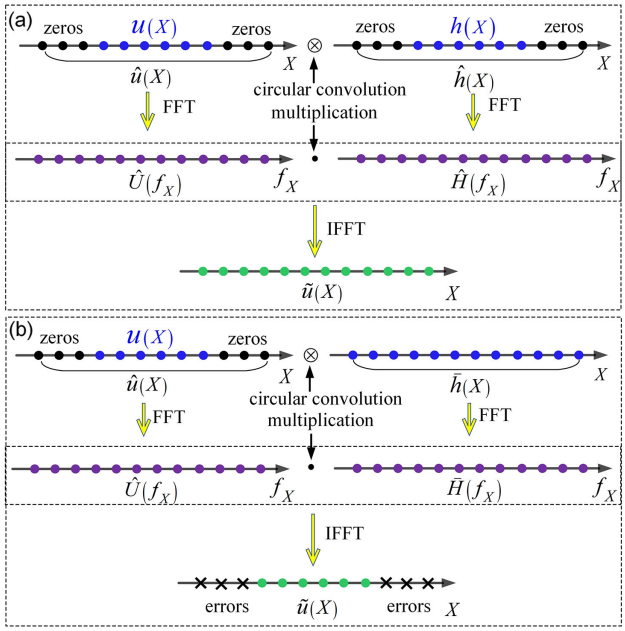

Fig. 7. Two strategies to avoid circular convolution errors. (a) The standard process, which does zero padding to both $u(X)$ and $h(X)$, and (b) the popular process, which only does zero padding to $u(X)$ and models $h(X)$ with 2 N points to get $\bar{h}(X)$.

N-zero padded $u(X)$ and $h(X)$ are $\hat{u}(X)$ and $\hat{h}(X)$, respectively: the circular convolution of $\hat{u}(X)$ and $\hat{h}(X)$ is equivalent to the linear convolution of $u(X)$ and $h(X)$ [53].

However, in practice, no one uses the above strategy. The reason is that since we have to use 2 N sampling points, why not fully use them to sample the impulse response function but only use N points surrounded by N zeros? The popular process is just that: (1) zero padding the N point sampled $u(X)$ to $\hat{u}(X)$; (2) using 2 N points to sample the impulse response function or the transfer function to get $\bar{h}(X)$ or $\bar{H}(f_X)$, respectively; (3) inverse Fourier transform the product of their Fourier transforms $\mathrm{IFFT}\{\bar{A}(f_X)\} = \mathrm{IFFT}\{\mathrm{FFT}\{\hat{u}(X)\} \cdot \mathrm{FFT}\{\bar{h}(X)\}\}$ or $\mathrm{IFFT}\{\bar{A}(f_X)\} = \mathrm{IFFT}\{\mathrm{FFT}\{\hat{u}(X)\} \cdot \bar{H}(f_X)\}$ to get the diffraction field. The difference between this strategy and the first strategy is shown in Fig. 7.

Figure 7(a) shows the standard process to transform linear convolution to circular convolution by zero padding. All the 2 N points of $\tilde{u}(X)$ are correct. However, the high spatial frequencies of $\hat{h}(X)$ are lost because they are sampled by zeros. In contrast, Fig. 7(b) shows the popular process in practice, where all the 2 N points are employed to sample the impulse response function $\bar{h}(X)$. In this way, the high spatial frequencies of $\bar{h}(X)$ are fully used. The cost of this process is that only N points of $\tilde{u}(X)$ are correct because $\bar{h}(X)$ is not obtained by zero padding. Note that Fig. 7(b) shows the Fres-IR case; for Fres-TF, $\bar{H}(f_X)$ is directly sampled by 2 N points. The process shown in Fig. 7(b) can be regarded as that: to fast calculate the linear convolution of N point sampled $u(X)$ and 2 N point sampled $\bar{h}(X)$, N zeros are padded to $u(X)$ for the 2 N point sampled $\hat{u}(X)$, and the circular convolution of $\hat{u}(X)$ and $\bar{h}(X)$ is carried out, which can be accelerated by FFT. Therefore, only N points are correct [53]. In this way, the spatial length of the correct diffraction

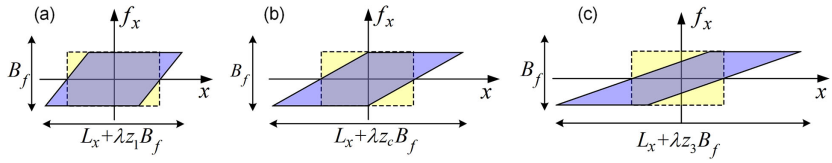

Fig. 8. PSDs with (a) $z_1 < \hat{z}$, (b) $z_2 = \hat{z}$, and (c) $z_3 > \hat{z}$.

field $\hat{u}(X)$ is the same as that of the input signal $u(X)$, which is $N\Delta_x$. The padded zeros are used only to guarantee the correct sampling in the spatial frequency domain.

Now, let us analyze the sampling requirement of $\bar{H}(f_X)$ or $\bar{h}(X)$, since they are used in practice. They have the same analytical expressions as $H(f_X)$ or $h(X)$, respectively, and the difference is just that they are sampled by 2 N points not N points.

For $\bar{H}(f_X)$, the first term, $\exp(i2\pi z/\lambda)/\sqrt{i\lambda z}$, is a constant that does not induce sampling aliasing; the second one, $\exp(-i\pi\lambda z f_X^2)$, does, because it varies with the spatial frequency f_X. According to the sampling theorem,

$$\left|\frac{1}{2\pi}\frac{\partial\varphi(f_X)}{\partial f_X}\right|_{\max} \le \frac{1}{2\Delta_f}, \tag{24}$$

where $\varphi(f_X) = -\pi\lambda z f_X^2$ is the phase of $\bar{H}(f_X)$ and $\Delta_f = B_f/2N = 1/2N\Delta_x$ is the sampling pitch of $\bar{H}(f_X)$. From inequality (24), we have

$$z \le \frac{2N\Delta_x^2}{\lambda}, \tag{25}$$

which means that $\bar{H}(f_X)$ can be correctly sampled only within this distance range.

Similarly for $\bar{h}(X)$, the first term, $\exp(i2\pi z/\lambda)/\sqrt{i\lambda z}$, is a constant that does not induce sampling aliasing; the second one, $\exp(i\pi X^2/\lambda z)$, does, because it varies with X. According to the sampling theorem,

$$\left|\frac{1}{2\pi}\frac{\partial\varphi(X)}{\partial X}\right|_{\max} \le \frac{1}{2\Delta_X}, \tag{26}$$

where $\varphi(X) = \pi X^2/\lambda z$ is the phase of $\bar{h}(X)$ and $\Delta_X = \Delta_x$ is the sampling pitch of $\bar{h}(X)$. From inequality (26), we have

$$z \ge \frac{2N\Delta_x^2}{\lambda}, \tag{27}$$

which means that $\bar{h}(X)$ can be correctly sampled only within this distance range.

By comparing inequalities (25) and (27), we can see that both Fres-TF and Fres-IR can be used when $z = \hat{z} = 2N\Delta_x^2/\lambda$. PSDs with $z_1 < \hat{z}$, $z_2 = \hat{z}$ and $z_3 > \hat{z}$ are shown in Fig. 8. The yellow rectangles surrounded by dashed line are the calculated SBPs, which do not fully cover the effective SBP inside the parallelogram. It is evident that with small z values, the utilization rate of the effective SBP, which is defined as the ratio of the covered effective SBP and the whole effective SBP, is high.

To verify the above analysis, we carried out numerical simulations with parameters listed in Table 1, and the results are shown in Fig. 9. The propagation distances are chosen as $z_1 = 30$ mm, $z_2 = 100$ mm, $z_3 = 300$ mm, and the critical distance in this case is $\hat{z} = 2N\Delta_x^2/\lambda = 100$ mm. RSI results are also given as reference.

From Fig. 9, we can conclude these points: (1) with $z_1 < \hat{z}$, Fres-TF gives correct results, while aliasing errors appear in Fres-IR results; (2) with $z_2 = \hat{z}$, both results given by Fres-TF and Fres-IR are correct, and they are identical; (3) with $z_3 > \hat{z}$, Fres-IR gives correct results, while high-frequency noises appear in Fres-TF results. One more observation is that with z increasing, the diffraction field inside the observation window, which is [−2.5 mm, 2.5 mm], is becoming sparse. "Sparse" means the diffraction field is spreading over a larger area, and the part inside the observation window becomes smooth. This is also consistent with the above analysis: that with larger z values, the SBP utilization rate would decrease because the covered effective SBP becomes smaller.

The reason why Fres-IR and Fres-TF have opposite and complementary applicable distance ranges can be explained physically. A spherical wave model is used in Fres-IR and a plane wave model is used in Fres-TF, as shown in Fig. 10. Because the input signal is sampled by pitch Δ_x, the highest spatial frequency $1/2\Delta_x$ determines the nonaliased biggest diffraction angle, which is $\alpha_{\max} = \arcsin(\lambda/2\Delta_x)$. This case occurs precisely when $z = \hat{z}$ both in Fres-IR and Fres-TF, as shown in Figs. 10(a) and 10(b). However, when $z < \hat{z}$, the biggest diffraction angle in Fres-IR is bigger than $\alpha_{\max}$, as line ③ indicates in Fig. 10(a), which would lead to aliasing error, while when $z > \hat{z}$, the biggest diffraction angle in Fres-IR is smaller than $\alpha_{\max}$, as line ① indicates in Fig. 10(a), which would give correct results. As for Fres-TF, there is no worry about the diffraction angle exceeding $\alpha_{\max}$ because the highest spatial frequency is always set to be $1/2\Delta_x$. However, the problem is whether the observation window could receive such high spatial frequencies. As shown in Fig. 10(b), when $z < \hat{z}$, all the spatial frequencies can be received by the observation window, providing correct results, while when $z > \hat{z}$, some high spatial frequencies exceed the observation window, as line ③ indicates, which would lead to high-frequency noises in the diffraction field. In addition, we can see another property of the plane-wave-model methods. For small propagation distances, the diffracted light from any point inside the source window cannot cover the full observation window. Therefore, the calculated results may be slightly different from that calculated by RSI, which can achieve a full-bandwidth calculation, as shown in Fig. 9.

C. ASM and RSC

The above three methods are all based on the Fresnel approximation, which is applicable in paraxial optics. Rigorous numerical diffraction calculation methods are often needed, such as in the

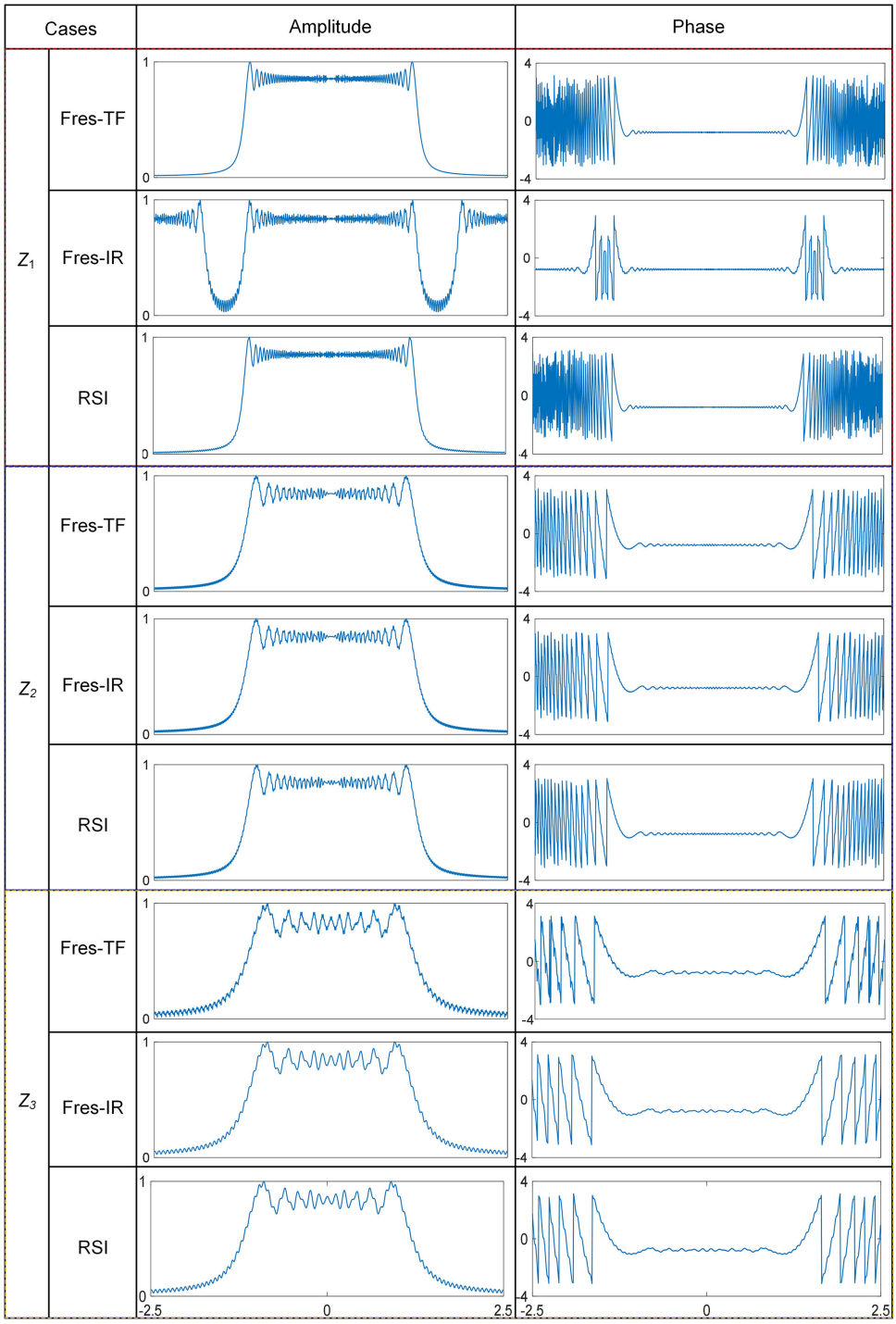

Fig. 9. Diffraction fields calculated by Fres-TF, Fres-IR, and RSI with $z_1 = 30$ mm, $z_2 = 100$ mm, and $z_3 = 300$ mm.

case of high numerical aperture. ASM and RSC are strict solutions to the diffraction theory without paraxial approximation, which can be used in such cases. Like Fres-TF and Fres-IR, ASM and RSC also model diffraction with plane waves and spherical waves, respectively. That is, ASM models diffraction by the transfer function, and RSC models diffraction by the impulse response function, which are given as below,

$$H_{\mathrm{ASM}}(f_X,\ f_Y) = \exp\left(i\frac{2\pi}{\lambda}z\sqrt{1-(\lambda f_X)^2-(\lambda f_Y)^2}\right), \quad \textbf{(28)}$$

$$h_{\mathrm{RSC}}(X,\ Y) = \frac{1}{2\pi}\frac{z}{r}\left(\frac{1}{r}-\frac{i2\pi}{\lambda}\right)\frac{\exp(i2\pi r/\lambda)}{r}, \quad \textbf{(29)}$$

where $r = \sqrt{X^2+Y^2+z^2}$. Mathematically, $H_{\mathrm{ASM}}(f_X,\ f_Y)$ and $h_{\mathrm{RSC}}(X,\ Y)$ are a Fourier transform pair [2]. However, again, they have opposite applicable distance ranges when discretely sampled. The calculation process is similar to that introduced above, except the nonapproximated functions, and

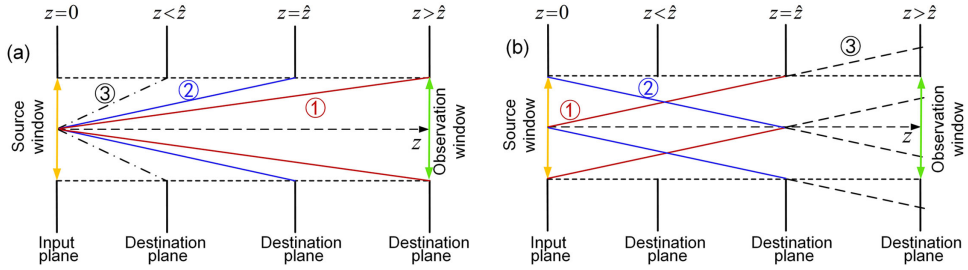

Fig. 10. Diagrams to physically explain why Fres-IR and Fres-TF have opposite and complementary applicable distance ranges.

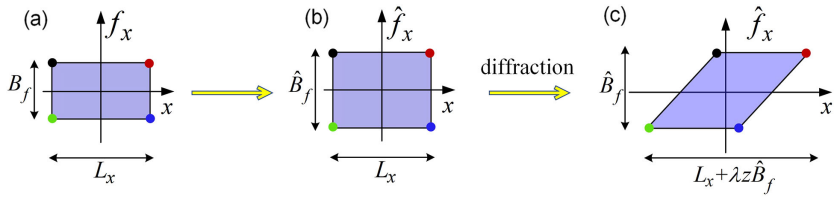

Fig. 11. (a) PSD shown in x, f_x coordinate system; pseudo-PSD shown in x, $\hat{f}_x$ coordinate system (b) before and (c) after diffraction.

give $\tilde{u}(X,Y)=\mathrm{IFFT}\{\mathrm{FFT}\{\hat{u}(X,Y)\}\cdot\overline{H_{\mathrm{ASM}}}(f_X,f_Y)\}$ or $\tilde{u}(X,Y)=\mathrm{IFFT}\{\mathrm{FFT}\{\hat{u}(X,Y)\}\cdot\mathrm{FFT}\{\overline{h_{\mathrm{RSC}}}(X,Y)\}\}$.

When it comes to obtaining the PSD through the ABCD matrix, we should be careful because the ABCD matrix is mainly used in paraxial optics. Although it has been proved that the ABCD matrix can be extended to nonparaxial optics, that would lead to a very complicated form [46]. This is because $\sin\alpha\approx\tan\alpha$ no longer holds in this case; and the polynomial expansion of trigonometric functions is complex. Physically, $X=x+z\tan\alpha$, where α is the direction angle and $f_X=f_x=\sin\alpha/\lambda$. When α is small, i.e., f_x is small, we have $X=x+z\tan\alpha\approx x+z\sin\alpha=x+\lambda z f_x$; in this way we can get a concise ABCD matrix. However, when α is large, i.e., f_x is large, $\tan\alpha=\sin\alpha/\cos\alpha=\lambda f_x/\sqrt{1-(\lambda f_x)^2}\neq\sin\alpha$; we do not have a concise ABCD matrix anymore if the vector $[x,f_x]$ is still used [54]. Therefore, in order to continue using the concise ABCD matrix form to describe the diffraction phenomenon, we define a new parameter named the pseudo-spatial frequency $\hat{f}_x=f_x/\sqrt{1-(\lambda f_x)^2}$ and use the ABCD matrix to find the relationship of the vector $[x,\hat{f}_x]$ before and after diffraction. In this way, we still have the concise form $X=x+\lambda z\hat{f}_x$,

$$\begin{bmatrix} X \\ \hat{f}_X \end{bmatrix}=\begin{pmatrix} 1 & \lambda z \\ 1 & 1 \end{pmatrix}\begin{bmatrix} x \\ \hat{f}_x \end{bmatrix}. \tag{30}$$

Please note that this is just for convenience and the actual spatial frequency is always f_x rather $\hat{f}_x$. Therefore, the SBP distribution in such a domain $(x,\hat{f}_x)$ is not a PSD anymore, but a pseudo one, which we define as a pseudo-PSD. Because we use the pseudo-spatial frequency $\hat{f}_x$, it is complicated to draw the PSDs after each step, especially the multiplication with the transfer function $\exp\left(i\frac{2\pi}{\lambda}z\sqrt{1-(\lambda f_x)^2}\right)$. Therefore, we directly show the pseudo-PSD after diffraction and skip the intermediate steps in the x, $\hat{f}_x$ coordinate, as shown in Fig. 11.

In the x, $\hat{f}_x$ coordinate system, the bandwidth $\hat{B}_f$ is larger than B_f because $\hat{f}_x=f_x/\sqrt{1-(\lambda f_x)^2}>f_x$. By using $\hat{f}_x$, the pseudo-PSD after diffraction is always a parallelogram, no matter whether in paraxial or nonparaxial optics. This is because $\hat{f}_x=\tan\alpha/\lambda$ and $X=x+z\tan\alpha$ describe the real physical scene. However, PSD is described in the phase space where the spatial frequency f_x is used. Therefore, the pseudo-PSDs shown in Figs. 11(b) and 11(c) are not real PSDs but the sketched version because $\hat{f}_x$ is obtained by sketching the corresponding f_x, and the sketching rate $1/\sqrt{1-(\lambda f_x)^2}$ is related with f_x. Mathematically, after considering the pseudo-PSD in the x, $\hat{f}_x$ coordinate system, the PSD in the x, f_x coordinate system can be obtained by sketching the pseudo-PSD with the rate $\sqrt{1-(\lambda f_x)^2}$. It can be expected that the difference between pseudo-PSD and PSD would be large in the nonparaxial cases and would be become small in the paraxial cases. We simulate these two cases with the parameters shown in Table 2.

Table 2. Parameters Used in the Simulation to Show the Relationship of Pseudo-PSDs in the x, $\hat{f}_x$ Coordinate System and PSDs in the x, f_x Coordinate System

Parameters	Values	Units
Length of input signal	$L_x=10$	mm
Bandwidth of input signal	$B_f=3$ and 1	$\mu\mathrm{m}^{-1}$
Wavelength	$\lambda=0.5$	μm
Propagation distance	$z=10$	mm

When f_x is small, i.e., the bandwidth is small, $\hat{f}_x=f_x/\sqrt{1-(\lambda f_x)^2}\approx f_x$, the pseudo-PSDs in x, $\hat{f}_x$ coordinate system would tend to PSDs in the x, f_x coordinate system. As can be seen from Fig. 12, with decreasing bandwidth B_f, the difference between $\hat{f}_x$ and f_x becomes smaller and PSDs in the x, f_x coordinate system are becoming parallelograms. It is also evident that even with large B_f, the central parts of pseudo-PSDs in the x, $\hat{f}_x$ coordinate system and the central parts of the PSDs in the x, f_x coordinate system are nearly coincident, which is consistent with the fact that $\sin\alpha\approx\tan\alpha$ when f_x is small.

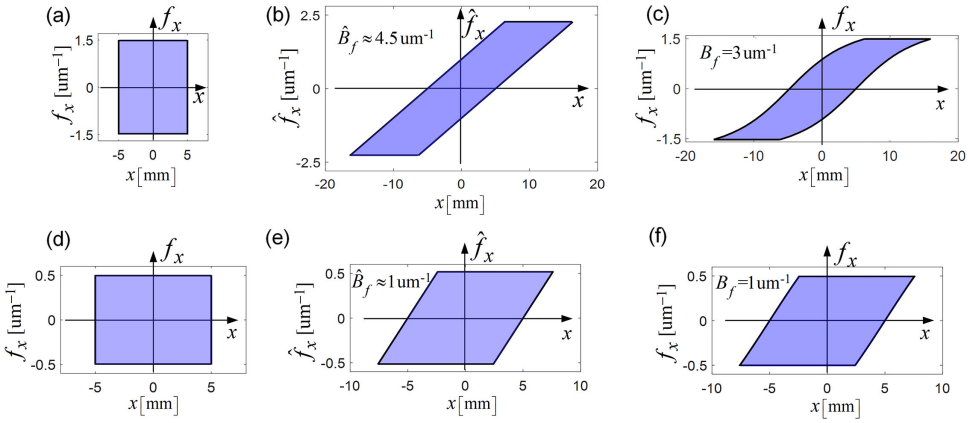

Fig. 12. PSDs in the x, f_x coordinate system and the pseudo-PSDs in the x, $\hat{f}_x$ coordinate system of large-bandwidth signal and small-bandwidth signal. (a) PSD of a large-bandwidth (3 μm^{-1}) signal before diffraction; (b) pseudo-PSD of the signal after diffraction; (c) PSD of the signal after diffraction; (d) PSD of a small-bandwidth (1 μm^{-1}) signal before diffraction; (e) pseudo-PSD of the signal after diffraction; (f) PSD of the signal after diffraction.

Figure 12 shows the PSDs/pseudo-PSDs given by ASM and RSC for both ASM and RSC. With these two methods, the observation window has the same size as the spatial length of the input signal, which is 10 mm in this case. This is because the sampling pitch in the destination plane is the same as that of the input plane. That is to say, even with these two nonparaxial methods, the SBP cannot be fully used, either, because the diffraction field spreads on a larger spatial range.

Next, let us analyze the sampling properties of $\overline{H_{\mathrm{ASM}}}(f_X)$ and $\overline{h_{\mathrm{RSC}}}(X)$.

In ASM, the key to getting the correct diffraction field is to sample a nonaliased transfer function $\overline{H_{\mathrm{ASM}}}(f_X)$, which is guaranteed by

$$\left|\frac{1}{2\pi}\frac{\partial\varphi(f_X)}{\partial f_X}\right|_{\max} \le \frac{1}{2\Delta_f}, \tag{31}$$

where $\varphi(f_X)=\frac{2\pi}{\lambda}z\sqrt{1-(\lambda f_X)^2}$ is the phase of $\overline{H_{\mathrm{ASM}}}(f_X)$ and $\Delta_f = B_f/2N = 1/2N\Delta_x$ is the sampling pitch of $\overline{H_{\mathrm{ASM}}}(f_X)$. From inequality (31), we get

$$z \le \frac{2N\Delta_x^2}{\lambda}\sqrt{1-\left(\frac{\lambda}{2\Delta_x}\right)^2}, \tag{32}$$

which means that $\overline{H_{\mathrm{ASM}}}(f_X)$ can be correctly sampled only within this distance range. Similarly, we can get the distance range where $\overline{h_{\mathrm{RSC}}}(X)$ is applicable by calculating

$$\left|\frac{1}{2\pi}\frac{\partial\varphi(X)}{\partial X}\right|_{\max} \le \frac{1}{2\Delta_X}, \tag{33}$$

where $\varphi(X)=\frac{2\pi}{\lambda}\sqrt{X^2+z^2}$ is the phase of $\overline{h_{\mathrm{RSC}}}(X)$ and $\Delta_X=\Delta_x$ is the sampling pitch of $\overline{h_{\mathrm{RSC}}}(X)$. From inequality (33), we get

$$z \ge \frac{2N\Delta_x^2}{\lambda}\sqrt{1-\left(\frac{\lambda}{2\Delta_x}\right)^2}, \tag{34}$$

which means that $\overline{h_{\mathrm{RSC}}}(X)$ can be correctly sampled only within this distance range.

By comparing inequalities (32), (34) and inequalities (25), (27), we can find that the relationship between ASM and RSC is very similar to that between Fres-TF and Fres-IR. The only difference is the factor $\sqrt{1-(\lambda/2\Delta_x)^2}$. When $\Delta_x \gg \lambda$, ASM and RSC will degenerate into Fres-TF and Fres-IR, respectively; since $\sqrt{1-(\lambda/2\Delta_x)^2}\approx 1$ in that case. From another perspective, $\Delta_x \gg \lambda$ means the bandwidth of the input signal $1/\Delta_x$ is much smaller compared to $1/\lambda$, wherefore $\frac{2\pi}{\lambda}z\sqrt{1-(\lambda f_x)^2}\approx\frac{2\pi}{\lambda}z-\pi\lambda z{f_x}^2$ and $\frac{2\pi}{\lambda}\sqrt{X^2+z^2}\approx\frac{2\pi}{\lambda}z+\pi x^2/\lambda z$. The phases of ASM's transfer function and RSC's impulse response function degenerate into that of Fres-TF's transfer function and Fres-IR's impulse response function, respectively, while for large numerical aperture cases, ASM or RSC cannot be approximated as Fres-TF or Fres-IR.

Table 3. Parameters Used for ASM and RSC

Parameters	Values
Sampling number of the signal	$N=1000$
Sampling pitch of the signal	$\Delta_x = 1$ μm
Wavelength	$\lambda = 0.5$ μm

To verify the above analysis, we carry out numerical simulations with the parameters listed in Table 3; the results are shown in Fig. 13. The propagation distances are chosen as $z_1 = 1$ mm, $z_2 = 3.87$ mm, $z_3 = 10$ mm, and the critical distance with these parameters is $\hat{z} = 2N\Delta_x^2/\lambda\cdot\sqrt{1-(\lambda/2\Delta_x)^2} = 3.87$ mm. RSI results are also given as reference. What we can conclude from Fig. 13 for ASM and RSC is similar to that concluded from Fig. 9 for Fres-TF and Fres-IR. One more thing to notice is that Fres-TF can be correctly used in the applicable range of ASM, and Fres-IR can be correctly used in the most applicable range of RSC. Here, "correctly" is in the sense of sampling. However, this does not mean that in the situations where ASM or RSC is available, Fres-TF or Fres-IR is also available. The criterion of judgment is the ratio between the bandwidth of the signal $1/\Delta_x$

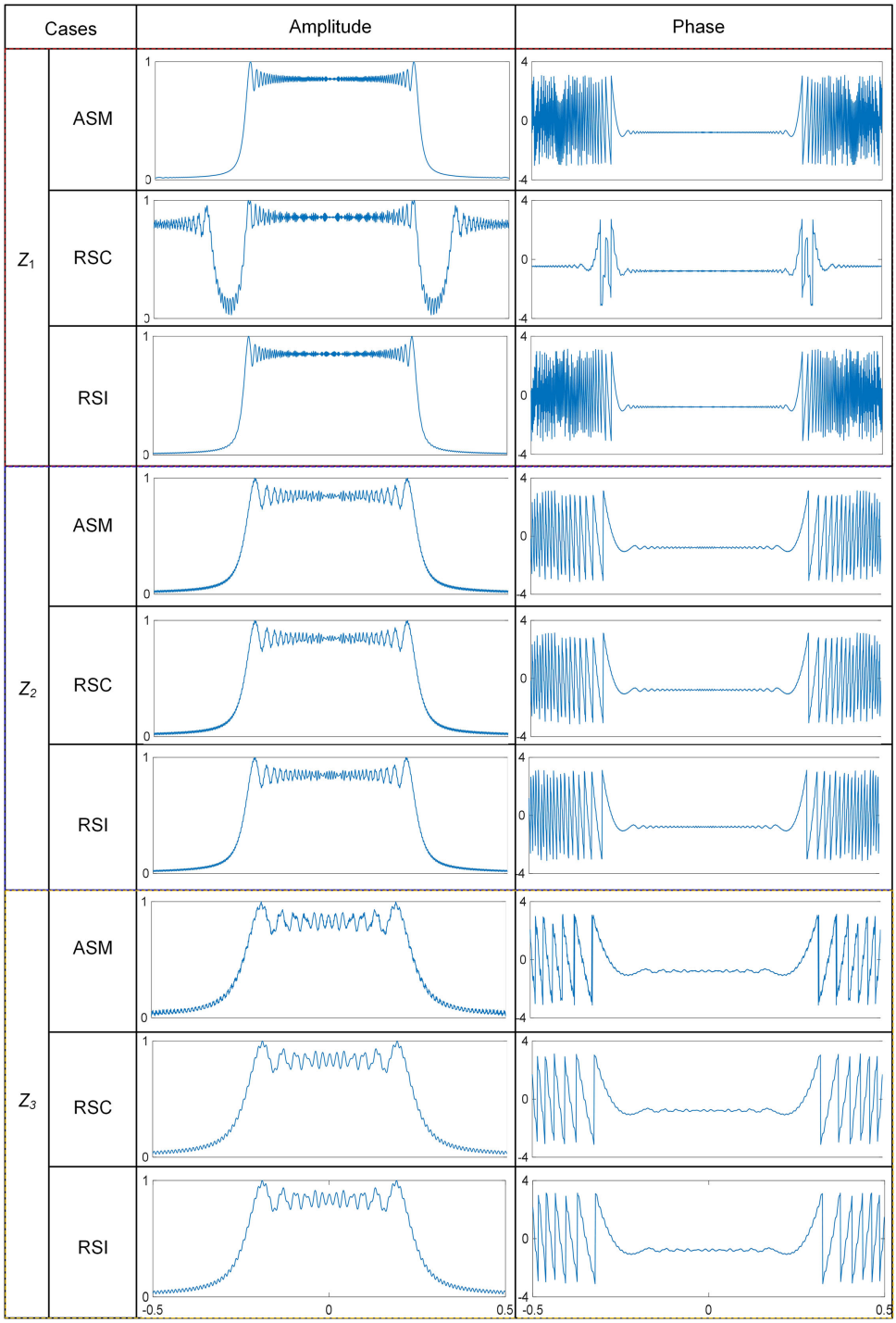

Fig. 13. Diffraction fields calculated by the ASM, RSC, and RSI with $z_1 = 1$ mm, $z_2 = 3.87$ mm, and $z_3 = 10$ mm.

and $1/\lambda$: if it is so small that $\sqrt{1-(\lambda/2\Delta_x)^2} \approx 1$, Fres-TF and Fres-IR can be used.

Up until now, all the five methods have been analyzed from the perspective of PSD and the sampling theorem. In Section 4, we would like to compare these methods and give some suggestions on how to choose one proper method for different applications, or adapt one of these methods for a specific purpose.

4. METHODS COMPARISON, AND IMPROVEMENT AND SELECTION SUGGESTIONS

Among these five methods, single FFT-based Fresnel transform, Fres-TF, Fres-IR, ASM, and RSC, the last four methods are all convolution-based methods, which regard the diffraction integral as a convolution. Fres-TF and ASM model the diffraction in the spatial frequency domain by the transfer function, and

Fres-IR and RSC model the diffraction in the space domain by the impulse response function. There is, in each case, a function to describe the response of the system during free-space diffraction propagation. All these four methods are based on the FFT algorithm. The input signal is transformed to the spatial frequency domain by FFT and then multiplied with the transfer function or the Fourier transform of the impulse response function; then the product to the space domain is transformed by inverse FFT to get the diffraction field. Due to the property of the FFT algorithm, the observation window has the same size as the spatial length of the input signal. This is because the sampling pitch in the destination plane is the same as that in the input plane, $\Delta_X = \Delta_x$. On the contrary, the first method, the single FFT-based Fresnel transform, does not obey $\Delta_X = \Delta_x$, but $\Delta_X = \lambda z/N\Delta_x$, which changes with the propagation distance. This is because single FFT-based Fresnel transform is not a convolution-based method and what the FFT algorithm achieves is actually an optical Fourier transform, which can be seen from Eq. (9).

Whether the sampling pitch in the destination plane changes or not is an important criterion to choose among diffraction calculation methods. In applications where the calculated diffraction field is as large as the input signal, such as reconstructing an image in digital holographic microscopy [12], the convolution-based methods are preferable, while in applications where the calculated diffraction field is larger than the input signal, such as holographic projection from computer-generated holograms [55,56], the single FFT-based Fresnel transform is preferable. As for how to choose one convolution-based method, ASM and RSC can be used in their applicable distance ranges, no matter whether they are paraxial or nonparaxial; Fres-TF and Fres-IR can be only used in their applicable distance ranges in paraxial optics. Therefore, ASM and RSC may be better. On the other hand, because the SBP utilization rate would decrease with the propagation distance, it is better to design the application case where ASM is applicable for higher SBP utilization, for example, to put the object as near to the sensor as possible in digital holography. A diagrammatic comparison among these four convolution-based methods is shown in Fig. 14.

Another dimension of comparison is the calculation efficiency. ASM and Fres-TF use FFT twice; RSC and Fres-IR use FFT 3 times; and single FFT-based Fresnel transform uses FFT once. If any method can be used for an application and high calculation efficiency is desired, especially in the iterative optimizations, single FFT-based Fresnel transform, ASM, and Fres-TF should be prioritized.

Although these methods enable us to calculate the diffraction field, there are still some functions these methods cannot achieve, for example, scaling the sampling pitch in the destination plane based on needs, obtaining the region of interest rather the whole diffraction field, etc. We call these two functions scaling and zooming, respectively. There are also some restrictions of each method, and here we provide some solutions to alleviate these restrictions and achieve these two functions.

Let us first alleviate the restrictions of each method.

For single FFT-based Fresnel transform, one restriction is that it works only at a specific propagation distance when both amplitude and phase are of concern, as analyzed in Section 3.A. This is because the two chirp functions have opposite applicable distance ranges. By comparing inequalities (13) and (16), we notice that we can use different sampling numbers to sample these two chirp functions based on needs, i.e., N for the inner chirp function and $\hat{N}$ for the outer chirp function. As long as $\hat{N} > N$, the applicable distance range of these two chirp functions would overlap, and the overlap region depends on how much is $\hat{N}$ bigger than N. In this way, we get the applicable distance range of these two chirp functions,

$$z \geq \frac{N\Delta_x^2}{\lambda}, \tag{35}$$

$$z \leq \frac{\hat{N}\Delta_x^2}{\lambda}. \tag{36}$$

Therefore, single FFT-based Fresnel transform can be used at an extended distance range. Sampling number $\hat{N}$ can be calculated by inequality (36) as $\hat{N} > [\lambda z/\Delta_x^2]$, where $[\cdot]$ represents a rounding operation and z is the propagation distance in practical applications.

The corresponding algorithm can be divided into three steps: (1) Multiply the N point sampled signal $u(x)$ with the inner chirp function $\exp(i\pi x^2/\lambda z)$ that is also N point sampled; (2) apply $\hat{N}$ point FFT to this product $u(x)\exp(i\pi x^2/\lambda z)$;

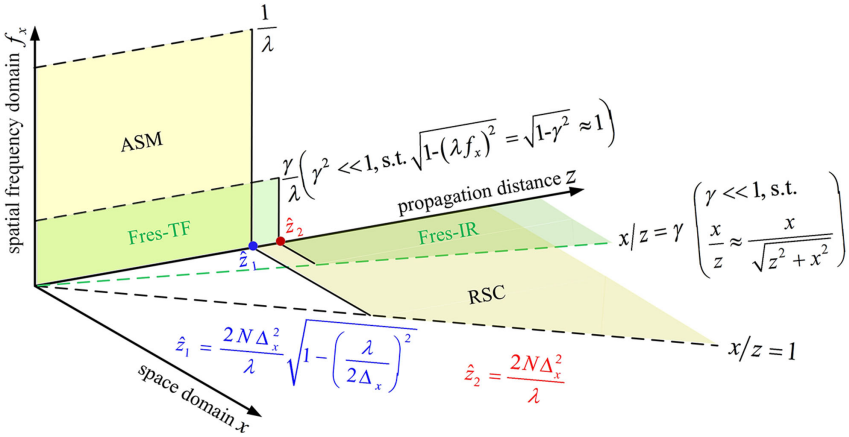

Fig. 14. Diagrammatic comparison among these four convolution-based methods.

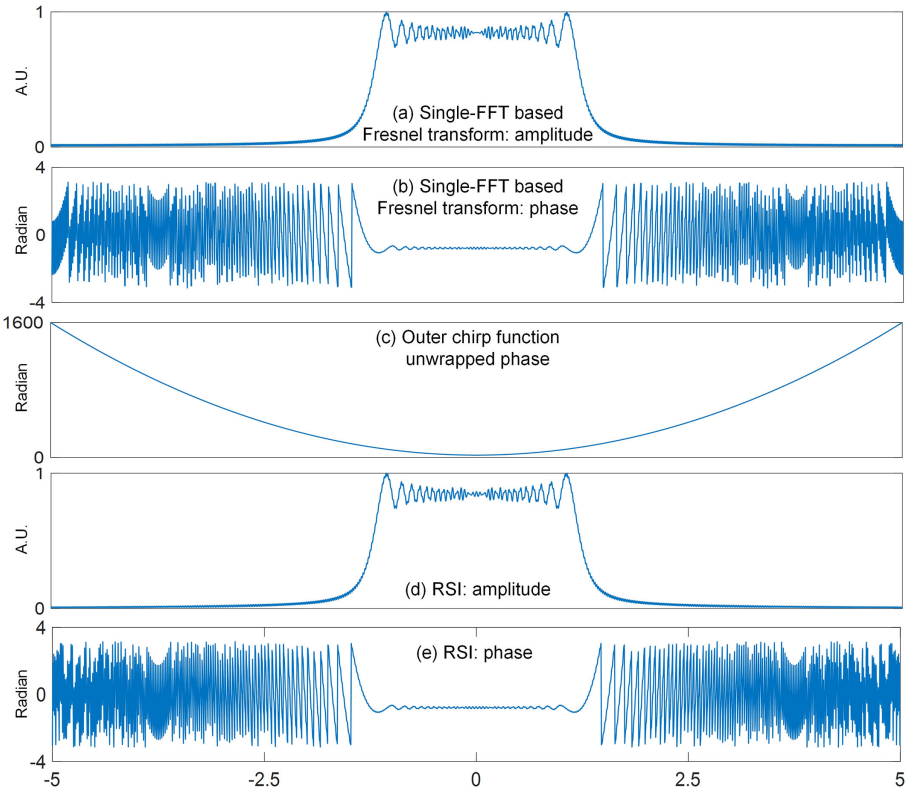

Fig. 15. Diffraction fields calculated by the adapted single FFT-based Fresnel transform and RSI with $z_2 = 100$ mm.

and (3) multiply $\hat{N}$ point sampled outer chirp function $\exp(i\pi X^2/\lambda z)$ with the Fourier transform obtained in (2). The constant term $\exp\left(i\frac{2\pi}{\lambda}z\right)/\sqrt{i\lambda z}$ can be simultaneously multiplied in step (3).

In Section 3.A, we made numerical simulations to demonstrate that the phase distribution given by single FFT-based Fresnel transform would be aliased if the propagation distance is larger than $N\Delta_x^2/\lambda z$. Here, we repeat this simulation while using the adapted version when $z_2 = 100$ mm; the results are shown in Fig. 15. In this simulation, $\hat{N} = 2000$, and other parameters are the same as shown in Table 1.

It is clear that the outer chirp function has been sampled without aliasing and the phase distribution of the calculated diffraction field by the adapted single FFT-based Fresnel transform is correct. From the perspective of PSD, when the propagation distance becomes large, the required sampling SBP is also becoming large, which means more sampling points should be used for correct calculation. This is the reason why increasing sampling points ($\hat{N} > N$) could give correct results.

For the convolution-based methods, ASM and Fres-TF can be used only with a small propagation distance; and RSC and Fres-IR can be used only with a large propagation distance, as shown in Fig. 14. The reason for these restrictions is because aliasing errors would appear if they are used beyond their applicable distance ranges. Therefore, to make ASM and Fres-TF applicable for large propagation distances or RSC and Fres-IR applicable for small propagation distances, the key is to avoid the aliasing errors.

For ASM, many methods have been proposed to make it applicable for large propagation distances. Here, a brief review associated with some analysis is given. Recall the phase of ASM's transfer function, $\varphi(f_X) = \frac{2\pi}{\lambda}z\sqrt{1-(\lambda f_X)^2}$. This time, we do not calculate the distance range z but the spatial frequency f_X. Based on inequality (31), we can get

$$|f_X| \le f_c = \frac{1}{\lambda\sqrt{1+z^2/N^2\Delta_x^2}}, \tag{37}$$

which means that the transfer function only inside $[-f_c, f_c]$ can be sampled without aliasing. From inequality (37), it is clear that f_c would decrease with z and increase with N. The highest spatial frequency associated with the sampling pitch Δ_x is $1/2\Delta_x$. Within the range given by inequality (32), $f_c = 1/2\Delta_x$ and f_c would decrease with z beyond that range. Therefore, there are two typical solutions to sample the ASM's transfer function without aliasing in large distances: (1) increasing sampling number N; and (2) forcing the components of ASM's transfer function beyond $[-f_c, f_c]$ to be zero. The first solution makes $f_c = 1/2\Delta_x$ always hold by increasing N, that is the zero-padding method [57]. The number of padded zeros increases dramatically with large propagation distance. The second solution does not increase sampling number but limits the effective bandwidth of the transfer function. Since not all the bandwidth $[-1/2\Delta_x, 1/2\Delta_x]$ can be correctly sampled, the aliased part is forced to be zero to avoid aliasing errors, which is the key point in this method, called band-limited ASM [21]. Inspired by this method, we have proposed band-extended ASM, which significantly extends the effective bandwidth of ASM's transfer

function without increasing sampling number. In our method, band extension is achieved by rearranging the sampling points in the spatial frequency domain. Recalling inequality (31), we use a new sampling pitch $\Delta_{f\_\text{new}} = 2f_{x\_\text{extend}}/2N$ rather than $\Delta_f = 1/2N\Delta_x$ to solve for the effective bandwidth $[-f_{X\_\text{extend}}, f_{X\_\text{extend}}]$. After some simplifications, we get

$$f_{X\_\text{extend}} = \sqrt{\frac{N}{2\lambda z}}. \tag{38}$$

It is easy to verify that f_{X_extend} is much larger than f_c with large z values. Thanks to the band extension, band-extended ASM can be used over a much wider distance range compared with ordinary ASM or band-limited ASM. Because we changed the sampling pitch in the spatial frequency domain, ASM's transfer function $\overline{H_{\text{ASM}}}(f_X)$ is sampled at these spatial frequencies $n\Delta_{f\_\text{new}}$, where $n = -N, -(N-1) \ldots 0 \ldots N-1$. However, the input signal $\hat{u}(X)$ is sampled at the spatial locations of $n\Delta_X$. FFT cannot transform $\hat{u}(X)$ to $\hat{u}(f_X)$, where f_X is the same as that of $\overline{H_{\text{ASM}}}(f_X)$, because $\Delta_{f\_\text{new}} \neq \Delta_f$. Therefore, we employed nonuniform FFT to achieve this operation. For details of this method, please refer to [23]; the link for the MATLAB code can be found there. Fres-TF can be processed in the same way. As for the RSC and Fres-IR, the methods of limiting or extending the effective bandwidth of the impulse response function can be similarly used.

Next, we focus on how to achieve scaling and zooming.

Scaling means observing the diffraction field with adjustable sampling pitch; and zooming means observing a region of interest, not the whole diffraction field. Also, there are some methods proposed to realize these two functions, such as [58]. Here, we take ASM as an example to illustrate how to achieve these two functions. In ASM, there are two key steps, which are Fourier transformation of the signal $\hat{u}(X)$, and inverse Fourier transformation of the product $\hat{U}(f_X)\overline{H_{\text{ASM}}}(f_X)$. Conventionally, the Fourier transform and inverse Fourier transform are implemented by the FFT algorithm, which fixes the relationship between the sampling pitches in the space domain and the spatial frequency domain as $\Delta_f = 1/2N\Delta_x$. Therefore, we need to find another tool to achieve scaling and zooming. Here, we use a scaled FFT, usually called the chirp-z transform [59,60]. We introduce how to use the chirp-z transform to realize scaling and zooming in principle, and the algorithm implementation can then be done easily.

Suppose (N, Δ_x) and (M, Δ_f) are the sampling parameters in the space domain and spatial frequency domain, respectively. Please note that they are independent, and the relationship $\Delta_f = 1/2N\Delta_x$ no longer holds. We write the discrete Fourier transform of the input signal $u(x)$ in the following form:

$$U(f_x) = U(m\Delta_f) = \sum_{n=0}^{N-1} u(n\Delta_x) \exp\Big[-i2\pi \times (f_0 + m\Delta_f)(x_0 + n\Delta_x)\Big], \quad m = 0, 1, \ldots M-1, \tag{39}$$

where f_0 and x_0 are the coordinates of the beginning points. $(f_0 + m\Delta_f)(x_0 + n\Delta_x)$ can be expressed by the identity

$$(f_0 + m\Delta_f)(x_0 + n\Delta_x) = x_0(m\Delta_f + f_0) + \frac{m^2}{2}\Delta_x\Delta_f + nf_0\Delta_x + \frac{n^2}{2}\Delta_x\Delta_f - \frac{(m-n)^2}{2}\Delta_x\Delta_f. \tag{40}$$

Substituting Eq. (40) into Eq. (39), we get

$$U(f_x) = U(m\Delta_f) = \exp\left[-i2\pi\left(x_0(m\Delta_f + f_0) + \frac{m^2}{2}\Delta_x\Delta_f\right)\right] \times \sum_{n=0}^{N-1} u(n\Delta_x)\exp\left[-i2\pi\left(nf_0\Delta_x + \frac{n^2}{2}\Delta_x\Delta_f\right)\right] \times \exp\left(i2\pi\frac{(m-n)^2}{2}\Delta_x\Delta_f\right), \quad m = 0, 1, \ldots M-1. \tag{41}$$

Equation (41) is precisely a convolution of the two sequences a_n and b_n defined by

$$a_n = u(n\Delta_x)\exp\left[-i2\pi\left(nf_0\Delta_x + \frac{n^2}{2}\Delta_x\Delta_f\right)\right], \tag{42}$$

$$b_n = \exp\left(i2\pi\frac{n^2}{2}\Delta_x\Delta_f\right), \tag{43}$$

with the output of the convolution multiplied by a phase term b_m,

$$b_m = \exp\left[-i2\pi\left(x_0(m\Delta_f + f_0) + \frac{m^2}{2}\Delta_x\Delta_f\right)\right]. \tag{44}$$

That is,

$$U(f_x) = U(m\Delta_f) = b_m\sum_{n=0}^{N-1} a_n b_{m-n}, \quad m = 0, 1, \ldots M-1. \tag{45}$$

Equation (45) can be calculated efficiently by FFT because it is in the form of a convolution. In the same way, we can calculate the inverse chirp-z transform. Therefore, we can choose the coordinates of the original points, the number of sampling points and sampling pitches. Benefitting from these flexibilities, scaling and zooming can be achieved in the diffraction calculation. The only cost of this method is that the computational complexity may increase because the Fourier transform is implemented using convolution. According to the convolution theorem, three FFTs are required to calculate Eq. (45). There are also other mathematical tools, such as nonuniform FFT, that can be used for scaling and zooming [61,62].

5. SUMMARY AND OUTLOOK

To summarize, we have analyzed five popular fast diffraction calculation methods: the single FFT-based Fresnel transform, the Fres-TF approach, the Fres-IR approach, the ASM, and the RSC, from the perspective of PSD and the sampling theorem. The sampling property and the effect on the signal's SBP for each method have been studied in detail and solutions to improve these methods are given.

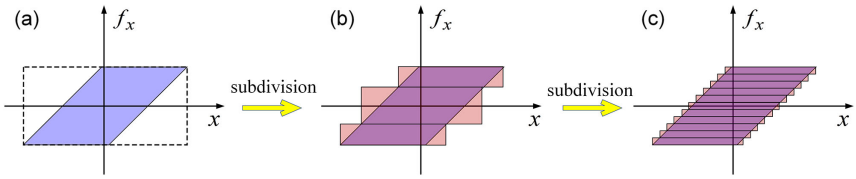

Fig. 16. (a)–(c) The sampling SBP can be gradually reduced down to the signal's SBP by subdivision along the f_x axis.

If the final phase is required, single FFT-based Fresnel transform is only applicable at one specific propagation distance, due to the opposite applicable distance ranges of two chirp functions. To overcome this restriction, we propose to apply FFT to the product of the object and the inner chirp with more sampling numbers, to satisfy the sampling requirement of the outer chirp function. In addition, the bandwidth increase caused by chirp modulation is ignored, which would cause aliasing with large-bandwidth signals. This problem should be quantitatively studied for a practical solution. As convolution-based methods, Fres-TF, Fres-IR, ASM, and RSC all need zero padding to avoid circular convolution errors during the calculation. Aiming to overcome this problem, we have proposed an adaptive-sampling ASM to automatically satisfy the sampling requirement to avoid circular convolution without zero padding [63]. Both ASM and Fres-TF model diffraction with a transfer function in the spatial frequency domain and are applicable for small propagation distances. Both RSC and Fres-IR model the diffraction with impulse response function in the space domain and are applicable for large propagation distances. The physical reason why they have different applicable distance ranges is given in Section 3.B, with Fig. 10. Besides, approaches to make these four methods applicable for a larger distance range have been analyzed and discussed in Section 4. Finally, we introduced how to use chirp-z transform, a kind of scaled FFT, to achieve scaling and zooming in the numerical diffraction calculation.

Furthermore, all these five methods are used to calculate Fresnel diffraction in free space. From the PSD shown in this paper, we can see that the required sampling SBP for the diffraction field is larger than the signal's SBP. Even though the signal's SBP does not change after diffraction, it spreads over a larger space, which requires a larger sampling SBP to describe the diffraction field. Thus, not all the diffraction field can be obtained by these methods when the sampling SBP is set to a number equivalent to the signal's SBP. In single FFT-based Fresnel transform, the spatial range of the calculated diffraction field is $\lambda z/\Delta_x$, while the whole diffraction field spreads over a range of $N\Delta_x + \lambda z/\Delta_x$, where $N\Delta_x$ is the spatial length of the input signal. Considering it can be only used when $z = N\Delta_x^2/\lambda$, the calculated spatial range is actually $N\Delta_x$. In the four convolution-based methods, the calculated spatial range is always $N\Delta_x$, the same as that of the input signal. Besides, when the propagation distance is relatively large, the bandwidth of the calculated diffraction field is also smaller than that of the input signal. To obtain the whole diffraction field and maintain the bandwidth, the only solution, currently, is to increase the sampling SBP, which would increase the computational complexity, as shown in Fig. 16(a). The signal's SBP is indicated by the blue parallelogram, while the required sampling SBP is indicated by the dashed-line rectangle, whose area is much larger than that of the parallelogram. A large amount of the sampling SBP is wasted, as indicated by the blank area.

Since we can control the spatial frequency range of the Fourier transform, as introduced in Section 4, it is possible to perform subdivision along the f_x axis, as shown in Fig. 16(b). That is to say, we do not calculate the whole diffraction field at once, but in three sections. First, we calculate the lower part, then the middle part, and, finally, the upper part. In this way, the required sampling SBP, the sum of three rectangular areas, can be greatly reduced. Going one step further, if more subdivisions are made, the required sampling SBP would trend to the signal's SBP, as shown in Fig. 16(c). Another potential way might be to model the diffraction in other geometric spaces, such as in Riemann space. In the Euclidean space, there is horizontal shearing between the PSDs before and after Fresnel diffraction, which increases the area of the circumscribed rectangle. The reason why we have to use a rectangle as the measure of SBP rather other shapes in the phase space is because we are using a Cartesian coordinate system in Euclidean space. Therefore, it is inevitable that we increase the sampling SBP to embrace the signal's SBP after diffraction, and some of it is wasted. It may be possible to make the sampling SBP and the signal's SBP always the same by using other geometric measures in a proper space; related work in Ref. [34] may be a good point. If so, the PSD should be in a different form in such a situation.

Overall, the task of numerical diffraction calculation is to efficiently and flexibly obtain the accurate diffraction field according to demand. Current methods can achieve one or two of these three functions. A generalized diffraction calculation method that can simultaneously achieve these three functions needs to be developed in the future. Also, new diffraction theory in different geometric spaces could be explored.

Funding. National Natural Science Foundation of China (62035003, 61875105); National Key Research and Development Program of China (2017YFF0106400).

Acknowledgment. The authors thank the anonymous reviewers.

Disclosures. The authors declare no conflicts of interest.

REFERENCES

1. M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (Elsevier, 2013).
2. J. W. Goodman, *Introduction to Fourier Optics* (W. H. Freeman, Macmillan Learning, 2017).
3. A. Sommerfeld, "Mathematische theorie der diffraction," Math. Ann. **47**, 317–374 (1896).

4. J. B. Keller, "Geometrical theory of diffraction," J. Opt. Soc. Am. **52**, 116–130 (1962).
5. A. Rubinowicz, "V the Miyamoto-Wolf diffraction wave," in *Progress in Optics* (Elsevier, 1965), pp. 199–240.
6. Y. Park, C. Depeursinge, and G. Popescu, "Quantitative phase imaging in biomedicine," Nat. Photonics **12**, 578–589 (2018).
7. D. G. Voelz, *Computational Fourier Optics: A MATLAB Tutorial* (SPIE, 2011).
8. Y. Z. Liu, F. A. South, Y. Xu, P. S. Carney, and S. A. Boppart, "Computational optical coherence tomography [Invited]," Biomed. Opt. Express **8**, 1549–1574 (2017).
9. W. Zhang, L. Cao, D. J. Brady, H. Zhang, J. Cang, H. Zhang, and G. Jin, "Twin-image-free holography: a compressive sensing approach," Phys. Rev. Lett. **121**, 093902 (2018).
10. W. Zhang, L. Cao, G. Jin, and D. Brady, "Full field-of-view digital lensfree holography for weak-scattering objects based on grating modulation," Appl. Opt. **57**, A164–A171 (2018).
11. W. Zhang, L. Cao, R. Li, H. Zhang, H. Zhang, Q. Jiang, and G. Jin, "Wavefront division digital holography," AIP Adv. **8**, 055304 (2018).
12. P. Marquet, B. Rappaz, P. J. Magistretti, E. Cuche, Y. Emery, T. Colomb, and C. Depeursinge, "Digital holographic microscopy: a noninvasive contrast imaging technique allowing quantitative visualization of living cells with subwavelength axial accuracy," Opt. Lett. **30**, 468–470 (2005).
13. H. Zhang, L. Cao, and G. Jin, "Three-dimensional computer-generated hologram with Fourier domain segmentation," Opt. Express **27**, 11689–11697 (2019).
14. Y. Zhao, L. Cao, H. Zhang, D. Kong, and G. Jin, "Accurate calculation of computer-generated holograms using angular-spectrum layer-oriented method," Opt. Express **23**, 25440–25449 (2015).
15. A. Jesacher, S. Bernet, and M. Ritsch-Marte, "Combined holographic optical trapping and optical image processing using a single diffractive pattern displayed on a spatial light modulator," Opt. Lett. **39**, 5337–5340 (2014).
16. M. J. Padgett, "Orbital angular momentum 25 years on [Invited]," Opt. Express **25**, 11265–11274 (2017).
17. S. S. Kou and C. J. R. Sheppard, "Image formation in holographic tomography: high-aperture imaging conditions," Appl. Opt. **48**, H168–H175 (2009).
18. Y. Sung, W. Choi, C. Fang-Yen, K. Badizadegan, R. R. Dasari, and M. S. Feld, "Optical diffraction tomography for high resolution live cell imaging," Opt. Express **17**, 266–277 (2009).
19. D. G. Voelz and M. C. Roggemann, "Digital simulation of scalar optical diffraction: revisiting chirp function sampling criteria and consequences," Appl. Opt. **48**, 6132–6142 (2009).
20. D. P. Kelly, "Numerical calculation of the Fresnel transform," J. Opt. Soc. Am. A **31**, 755–764 (2014).
21. K. Matsushima and T. Shimobaba, "Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields," Opt. Express **17**, 19662–19673 (2009).
22. Y.-H. Kim, C.-W. Byun, H. Oh, J. Lee, J.-E. Pi, G. Heon Kim, M.-L. Lee, H. Ryu, H.-Y. Chu, and C.-S. Hwang, "Non-uniform sampling and wide range angular spectrum method," J. Opt. **16**, 125710 (2014).
23. W. Zhang, H. Zhang, and G. Jin, "Band-extended angular spectrum method for accurate diffraction calculation in a wide propagation range," Opt. Lett. **45**, 1543–1546 (2020).
24. T. Shimobaba, K. Matsushima, T. Kakue, N. Masuda, and T. Ito, "Scaled angular spectrum method," Opt. Lett. **37**, 4128–4130 (2012).
25. T. Kozacki and K. Falaggis, "Angular spectrum-based wave-propagation method with compact space bandwidth for large propagation distances," Opt. Lett. **40**, 3420–3423 (2015).
26. F. Zhang, I. Yamaguchi, and L. P. Yaroslavsky, "Algorithm for reconstruction of digital holograms with adjustable magnification," Opt. Lett. **29**, 1668–1670 (2004).
27. R. P. Muffoletto, J. M. Tyler, and J. E. Tohline, "Shifted Fresnel diffraction for computational holography," Opt. Express **15**, 5631–5640 (2007).
28. J. F. Restrepo and J. Garcia-Sucerquia, "Magnified reconstruction of digitally recorded holograms by Fresnel–Bluestein transform," Appl. Opt. **49**, 6430–6435 (2010).
29. T. Shimobaba, T. Kakue, N. Okada, M. Oikawa, Y. Yamaguchi, and T. Ito, "Aliasing-reduced Fresnel diffraction with scale and shift operations," J. Opt. **15**, 075405 (2013).
30. C. S. Guo, Y. Y. Xie, and B. Sha, "Diffraction algorithm suitable for both near and far field with shifted destination window and oblique illumination," Opt. Lett. **39**, 2338–2341 (2014).
31. K. Matsushima, "Shifted angular spectrum method for off-axis numerical propagation," Opt. Express **18**, 18453–18463 (2010).
32. K. Matsushima, H. Schimmel, and F. Wyrowski, "Fast calculation method for optical diffraction on tilted planes by use of the angular spectrum of plane waves," J. Opt. Soc. Am. A **20**, 1755–1762 (2003).
33. L. Onural, "Exact analysis of the effects of sampling of the scalar diffraction field," J. Opt. Soc. Am. A **24**, 359–367 (2007).
34. H. M. Ozaktas, S. Ö. Arık, and T. Coşkun, "Fundamental structure of Fresnel diffraction: natural sampling grid and the fractional Fourier transform," Opt. Lett. **36**, 2524–2526 (2011).
35. H. M. Ozaktas, S. Ö. Arık, and T. Coşkun, "Fundamental structure of Fresnel diffraction: longitudinal uniformity with respect to fractional Fourier order," Opt. Lett. **37**, 103–105 (2012).
36. S. Ö. Arık and H. M. Ozaktas, "Optimal representation and processing of optical signals in quadratic-phase systems," Opt. Commun. **366**, 17–21 (2016).
37. B. M. Hennelly, J. J. Healy, and J. T. Sheridan, "Sampling in phase space," in *Phase-Space Optics Fundamentals and Applications*, M. E. Testorf, B. M. Hennelly, and J. Ojeda-Castaneda, eds. (McGraw-Hill, 2010).
38. A. Stern and B. Javidi, "Sampling in the light of Wigner distribution," J. Opt. Soc. Am. A **21**, 360–366 (2004).
39. A. Stern and B. Javidi, "Improved-resolution digital holography using the generalized sampling theorem for locally band-limited fields," J. Opt. Soc. Am. A **23**, 1227–1235 (2006).
40. L. Onural, "Sampling of the diffraction field," Appl. Opt. **39**, 5929–5935 (2000).
41. A. W. Lohmann, R. G. Dorsch, D. Mendlovic, Z. Zalevsky, and C. Ferreira, "Space–bandwidth product of optical signals and systems," J. Opt. Soc. Am. A **13**, 470–473 (1996).
42. B. M. Hennelly and J. T. Sheridan, "Generalizing, optimizing, and inventing numerical algorithms for the fractional Fourier, Fresnel, and linear canonical transforms," J. Opt. Soc. Am. A **22**, 917–927 (2005).
43. J. Serna, R. Martínez-Herrero, and P. M. Mejías, "Parametric characterization of general partially coherent beams propagating through ABCD optical systems," J. Opt. Soc. Am. A **8**, 1094–1098 (1991).
44. P. A. Bélanger, "Beam propagation and the ABCD ray matrices," Opt. Lett. **16**, 196–198 (1991).
45. S. A. Collins, "Lens-system diffraction integral written in terms of matrix optics," J. Opt. Soc. Am. **60**, 1168–1177 (1970).
46. T. Weihan, "Generality of optical ABCD theorem, its physical meaning and applications," Chin. J. Quantum Electron. **21**, 149–162 (2004).
47. C. J. Sheppard and K. G. Larkin, "Wigner function for highly convergent three-dimensional wave fields," Opt. Lett. **26**, 968–970 (2001).
48. C. J. Sheppard and K. G. Larkin, "Wigner function for nonparaxial wave fields," J. Opt. Soc. Am. A **18**, 2486–2490 (2001).
49. C. Sheppard and K. Larkin, "Wigner function and ambiguity function for non-paraxial three-dimensional wavefields," in *Optical Processing and Computing, 17-18 April 2001, Orlando, USA: A Tribute to Adolf Lohmann* (SPIE, 2002), pp. 37–56.
50. C. J. Sheppard, "Diffraction optics," in *Handbook of Biomedical Optics*, D. Boas, C. Pitris, and N. Ramanujam, eds. (CRC Press, 2011), pp. 11–31.
51. C. Sheppard and M. Hrynevych, "Diffraction by a circular aperture: a generalization of Fresnel diffraction theory," J. Opt. Soc. Am. A **9**, 274–281 (1992).
52. C. Sheppard, J. Lin, and S. Kou, "Rayleigh–Sommerfeld diffraction formula in k space," J. Opt. Soc. Am. A **30**, 1180–1183 (2013).
53. V. K. Ingle and J. G. Proakis, *Digital Signal Processing Using Matlab: A Problem Solving Companion* (Cengage Learning, 2016).
54. J. B. Almeida and V. Lakshminarayanan, "Wide angle near-field diffraction and Wigner distribution," Optik **114**, 333–335 (2003).
55. H. Zhang, Y. Zhao, L. Cao, and G. Jin, "Layered holographic stereogram based on inverse Fresnel diffraction," Appl. Opt. **55**, A154–A159 (2016).

56. T. Shimobaba and T. Ito, "Random phase-free computer-generated hologram," Opt. Express **23**, 9549–9554 (2015).
57. X. Yu, T. Xiahui, Q. Y. Xiong, P. Hao, and W. Wei, "Wide-window angular spectrum method for diffraction propagation in far and near field," Opt. Lett. **37**, 4943–4945 (2012).
58. B. M. Hennelly, D. P. Kelly, D. S. Monaghan, and N. Pandey, "Zoom algorithms for digital holography," in *Information Optics and Photonics* (Springer, 2010), pp. 187–204.
59. L. Rabiner, R. Schafer, and C. Rader, "The chirp z-transform algorithm," IEEE Trans. Audio Electroacoust. **17**, 86–92 (1969).
60. L. Bluestein, "A linear filtering approach to the computation of discrete Fourier transform," IEEE Trans. Audio Electroacoust. **18**, 451–455 (1970).
61. J.-Y. Lee and L. Greengard, "The type 3 nonuniform FFT and its applications," J. Comput. Phys. **206**, 1–5 (2005).
62. L. Greengard and J.-Y. Lee, "Accelerating the nonuniform fast Fourier transform," SIAM Rev. **46**, 443–454 (2004).
63. W. Zhang, H. Zhang, and G. Jin, "Adaptive-sampling angular spectrum method with full utilization of space-bandwidth product," Opt. Lett. **45**, 4416–4419 (2020).