VarRCWA: An Adaptive High-Order Rigorous Coupled Wave Analysis Method

Ziwei Zhu and Changxi Zheng*

ABSTRACT: Semianalytical methods, such as rigorous coupled wave analysis, have been pivotal in the numerical analysis of photonic structures. In comparison to other numerical methods, they have a much lower computational cost, especially for structures with constant cross-sectional shapes (such as metasurface units). However, when the cross-sectional shape varies even mildly (such as a taper), existing semianalytical methods suffer from high computational costs. We show that the existing methods can be viewed as a zeroth-order approximation with respect to the structure’s cross-sectional variation. We derive a high-order perturbative expansion with respect to the cross-sectional variation. Based on this expansion, we propose a new semianalytical method that is fast to compute even in the presence of large cross-sectional shape variation. Furthermore, we design an algorithm that automatically discretizes the structure in a way that achieves a user-specified accuracy level while at the same time reducing the computational cost.

KEYWORDS: numerical methods, nanophotonics, semianalytical methods, adaptive, coupling, GPU

INTRODUCTION

Numerical simulation is a fundamental tool for understanding photonic structures. Among many popular methods, semianalytical methods, such as rigorous coupled wave analysis (RCWA),¹ have been widely used for analyzing such devices as metasurfaces,² gratings,³ and waveguides.⁴ In comparison to other methods, such as finite-difference time-domain (FDTD) methods, semianalytical methods often have much lower computational costs. This advantage stems from how semianalytical methods discretize Maxwell’s equations. In contrast to other approaches (e.g., FDTD methods) that discretize the spatial domain fully (i.e., in all three dimensions),⁵ semianalytical methods discretize the spatial domain partially (e.g., in only x- and y-dimensions, but not z-dimension). This is possible because many photonic structures have a primary light propagation direction (referred to in this paper as the z-direction; see Figure 1). In some cases, along the light propagation direction, the structure’s cross-sectional shape stays unchanged (e.g., a metasurface unit). Therefore, we do not have to discretize the structure along the z-direction; instead, light propagation in the structure can be viewed as a superposition of individual propagating modes experiencing phase shifts. This is the fundamental view that enables semianalytical methods to reduce computational cost (see Solving Maxwell’s Equations).

However, this view becomes unsound for many photonic structures, wherein along the primary light propagation direction the structure’s cross section varies.⁶,⁷ A common example is photonic waveguides (such as a taper; see Figure 1a). To simulate these photonic structures using semianalytical methods, one has to further discretize the structure along the z-direction into a series of thin sections.⁸,⁹ In each section, the cross-sectional shape is assumed unchanged (see Figure 1b), and thereby a semianalytical method can be used to simulate that section. Yet, this approach requires a large number of sections, which in turn devastates the computational advantage of semianalytical methods. Apart from the computational cost, it is often unclear how many discrete sections are needed to achieve certain accuracy. In practice, one has to rely on trial and error to choose a proper resolution for sufficient accuracy. Oftentimes, to obtain satisfactory results, multiple runs of the simulation method (each with a different resolution) are needed.

In this work, we overcome these limitations. Our method requires no trial and error, thus much easier to use: provided a photonic structure and a user-specified accuracy level (i.e., a real number), our method automatically decides how to discretize the structure in z-direction, aiming to reduce the overall computational cost while achieving the desired accuracy.
accuracy. To obtain simulation results of user-specified accuracy, only one run is needed.

To this end, our core development is 2-fold: (1) We show that the conventional semianalytical methods (such as RCWA) are merely zeroth-order approximation with respect to the structure’s cross-sectional variation. Through a novel change of variable, we propose a high-order semianalytical method, which allows the structure’s cross section to vary over z-direction, without discretizing it into thin sections. (2) Leveraging this high-order method, we introduce an algorithm that automatically and adaptively discretizes the structure to achieve a user-specified accuracy level. For regions where the cross section varies rapidly in z-direction, our algorithm will slice the structure in fine resolution to ensure simulation accuracy; for regions with little cross-sectional variance, it will discretize them coarsely to save computational cost.

We use our method to analyze various photonic structures, and compare it with conventional semianalytical methods (such as RCWA) and the differential method.10 We show that our method, thanks to its higher-order approximation, indeed converges faster. As a result, to obtain the same level of accuracy, our method requires much less computational time and no resolution tuning at all.

## SOLVING MAXWELL’S EQUATIONS

We start by reviewing two widely used methods (namely, the differential method and RCWA) to solve Maxwell’s Equations in spatial and temporal Fourier space. We then present our method, which can be viewed as a higher-order generalization of RCWA. The differences of these methods are summarized in Table 1 and illustrated in Figure 1.

![Figure 1. Illustration of the three methods. (a) The differential method divides the structure into multiple sections (gray boxes) and approximates light propagation by a numerical integral (dashed pink arrow) of the electromagnetic fields (e and h, blue circles) in each section. Lastly, the integrals are combined using the Redheffer star product (•). (b) RCWA directly simulates using modal coefficients (a and b, green circles) instead of e and h. It computes the effective wavenumber \( \beta/\kappa_0 \) of each mode, assuming the waveguide’s cross-section stays fixed. It then approximates the scattering matrix by propagating modes directly. (c) Since many waveguides have varying cross sections, varRCWA (our method), as a high-order method, takes into account cross-mode coupling (orange and yellow arrows) caused by the cross-sectional variation (\( \delta A \) and \( \delta B \)) from a cross section at the reference position \( P_r \) and \( Q_r \).

### Table 1. Comparison of Three Simulation Methods

| method       | stability scaling | accuracy |
|--------------|-------------------|----------|
| differential | \( P_{z/2} \)    | +        |
| RCWA         | stable            | +        |
| varRCWA      | \( \delta P_z \)  | ++       |

*How the section length z and permittivity distribution \( P \) scale to divergence. +Error versus the number of sections. *It also depends on \( Q \) for simplicity, we only use \( P \). *A high-order relation may depend on the integration scheme, for example, \( (P_z)^4 \) for RK4. Detailed discussions in Results.

All these methods share the same starting point. They assume that there exists a primary light propagation direction (z-direction) in a photonic structure. Then, the electric and magnetic fields are only discretized on xy-plane and vary along z-direction. Under this setup, Maxwell’s Equations become

\[
\frac{\partial}{\partial z}[e] = j \frac{1}{\kappa_0} \begin{bmatrix} P \end{bmatrix} \begin{bmatrix} e \\ h \end{bmatrix}
\]

(1)

where the vectors \( e \) and \( h \) are discrete representations of the electric and magnetic fields (for example, in Fourier space) on an xy-plane at a z position. For simplicity of presentation, we assume that the material is isotropic and the permeability is uniform (\( \mu = 1 \)), resulting in vanishing diagonal blocks. The off-diagonal matrices \( P \) and \( Q \) encode the distributions of material permittivity on the xy-plane at the same z position.

**Differential Method.** One way of solving eq 1, which differs from the FDTD method, is the differential method. It integrates eq 1 by the fourth-order Runge–Kutta method (RK4).11 This integral connects the field \( e_r \) and \( h_r \) at the left end with the field \( e_k \) and \( h_k \) at the right end:

\[
\begin{bmatrix} e_k \\ h_k \end{bmatrix} = \frac{j}{\kappa_0} \int_0^z [P] \begin{bmatrix} e \\ h \end{bmatrix} dz' \approx \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} e_l \\ h_l \end{bmatrix}
\]

(2)

Yet, directly computing the \( T \) matrix using eq 2 is unstable. To overcome the instability, the \( T \) matrix is further converted into a modal scattering matrix12 \( S \). Through a change of variables \( a = W^{-1}e + V^{-1}h \) and \( b = W^{-1}e - V^{-1}h \), where \( W \) and \( V \) are eigenmodes of electric and magnetic fields (more discussion later). As a result, \( a \) and \( b \) are related through the modal scattering matrix \( S \):

\[
\begin{bmatrix} a_k \\ b_k \end{bmatrix} = S \begin{bmatrix} a_l \\ b_l \end{bmatrix}, \text{ where } S = \begin{bmatrix} T_{RR} & R_d \\ R_u & T_{RL} \end{bmatrix}
\]

(3)

Still, numerical computation of the \( S \) matrix is unstable unless the integration length in eq 2 is small. In practice, we have to slice the structure into many sections, each of which stays shorter than 0.1 \( \mu \). This often leads to a large number of sections and thus expensive computation.

**Rigorous Coupled Wave Analysis (RCWA).** A widely used method for numerically analyzing such structures as metasurfaces,\(^7\) gratings,\(^3\) and waveguides\(^4\) is the rigorous coupled wave analysis (RCWA),\(^1\) one of the most general semianalytical methods. When the structure’s cross-sectional shape is fixed along the z-direction, both \( P \) and \( Q \) in eq 1 are constant matrices, and eq 1 can be solved through an eigenvalue decomposition, that is, \( PQ = WAW^{-1} \) and \( V = QWA^{-1} \). In this case, we can directly construct the scattering matrix \( S \) in eq 3 without resorting to the \( T \) matrix in eq 2:
\[ T_{LR} = T_{RL} = e^{j/k_0 \Delta z} \quad \text{and} \quad R_L = R_R = 0 \quad (4) \]

where \( k_0 \) is the wavenumber in vacuum.

When the structure’s cross-sectional shape varies along \( z \)-direction, we must split the structure into a series of small sections so that every section can be approximated as having a fixed cross section (Figure 1b). Each section \( i \) is then analyzed through the aforementioned process, from which one can compute its scattering matrix \( S \). Finally, the entire structure’s scattering matrix is obtained by projecting all sections’ scattering matrices to match the bases of vectors \( a \) and \( b \) (see details in Supporting Information, Section 2) and then combining them using the Redheffer star product.\(^3\)

Unlike the FDTD method, both the differential method and RCWA discretize eq 1 in Fourier space, and thereby directly predict the full scattering matrix. Our method is in the same vein, described as follows.

**Our Method (varRCWA).** RCWA’s requirement of a photonic structure having a fixed cross-sectional shape can be viewed as a zeroth-order approximation of the structure’s material distribution along \( z \)-direction. A detailed derivation revealing this zeroth-order approximation is presented in Supporting Information, Section 1.4, where we further extend it into a higher-order approximation, one that takes into account cross-sectional shape change along \( z \)-direction.

Leaving the derivation details in the Supporting Information, Section 1, here we present the first-order expansion of the scattering matrix \( S \), although our derivation can produce the expansion of arbitrarily high order. The four sub-blocks of the \( S \) matrix are

\[
T_{LR} = e^{j/k_0 \Delta z} + \frac{j}{2k_0} \int_0^z e^{j/k_0 \Delta z} \delta A(z') e^{j/k_0 \Delta z'} dz' \\
R_R = -\frac{j}{2k_0} \int_0^z e^{j/k_0 \Delta z} \delta B(z') e^{j/k_0 \Delta z'} dz' \\
R_L = e^{j/k_0 \Delta z} + \frac{j}{2k_0} \int_0^z e^{j/k_0 \Delta z'} \delta A(z') e^{j/k_0 \Delta z'} dz' \\
T_{RL} = e^{j/k_0 \Delta z} + \frac{j}{2k_0} \int_0^z e^{j/k_0 \Delta z'} \delta B(z') e^{j/k_0 \Delta z'} dz' \quad (5)
\]

where we define transmission coupling matrix \( \delta A \) and reflection coupling matrix \( \delta B \) as follows:

\[
\delta A = W^{-1}(P - P)W + V^{-1}(Q - Q_0)W, \\
\delta B = W^{-1}(P - P)V - V^{-1}(Q - Q_0)W \quad (6)
\]

Here \( W, V, A, \) and \( V \) are obtained by solving an eigenvalue problem at a reference \( z \)-position (i.e., \( P, Q, = W^{-1} \)), the same as the conventional RCWA. Compare eq 5 with eq 4. The additional integral terms in eq 5 are to provide first-order correction of conventional RCWA. What the integrals account for is cross-mode coupling caused by the cross-sectional shape variation (see Figure 1c).

It is also worth noting that in eq 5 the integrands depend on \( \delta A \) and \( \delta B \), which, according to eq 6, indicate the difference of the cross-sectional shape from a reference shape (i.e., \( P, P, \) and \( Q, Q_0 \)). This contrasts starkly to the integral in the differential method eq 2, which integrates over \( P \) and \( Q \). When the cross section stays unchanged over \( z \)-direction, conventional RCWA should suffice. Indeed, in this case, \( \delta A = \delta B = 0 \), and eq 5 becomes identical to eq 4. If the cross section varies slowly over \( z \), the norms of \( \delta A \) and \( \delta B \) are small, and thus the first-order expansion converges even for a long section. This allows us to discretize the entire structure into fewer sections, and thereby reduces computational cost. If the cross section varies quickly, the norms of \( \delta A \) and \( \delta B \) may be large, and we have to use short section length to ensure convergence of our first-order expansion or adopt a higher-order expansion of the \( S \) matrix.

**Discussion on Other Methods.** Solution to Maxwell’s equations (eq 1) can also be expressed as a product integral that involves matrix exponentials:

\[
h_R = \prod_{i=0}^z \exp \left( \frac{j}{k_0} \begin{bmatrix} Q \end{bmatrix} \right) h_L \quad (7)
\]

Numerical evaluation of eq 7, however, is rather challenging. Prior works use a low-order Taylor expansion of the matrix exponentials to evaluate eq 7.\(^{16,17}\) But to use this expansion, section length must be excessively short (i.e., \( z_i - z_{i-1} \leq 0.1 \lambda \), where \( \lambda \) is the wavelength), and a large number of sections are needed. Another approach is to convert the product of matrix exponential into the exponential of matrix summation, similar to \( \prod \exp(x_i) = \exp(\sum x_i) \) for scalar values \( x_i \). However, this is not straightforward because matrix multiplication is non-commutative. As a result, Magnus expansion \(^{18-20}\) and Fer expansion\(^2\) have been used to correct the use of exponential of matrix summation. In this vein, RCWA can be viewed as an approximation of Magnus expansion, as shown in ref 22. Although a higher-order approximation of Magnus expansion was introduced in ref 22, it still requires each discrete section to be short, even with a small cross-sectional variation. These treatments all share the same stability issue as the differential method. Therefore, in our tests, we compare our method with the differential method and RCWA.

### NUMERICAL METHOD

**Numerical Integration.** The integrals in eq 5 can be numerically evaluated using a quadrature rule. In practice, we use the quadrature rule that samples three positions at the left end, right end, and the middle point of a section, and the integral is estimated by weighted summation of the integrand values at the sampled positions. This quadrature rule involves only matrix multiplications and additions, and thus can be computed at low cost. In practice, we implement it both on CPU and Graphics Processing Unit (GPU).

The major cost of evaluating wave propagation in a single section (using eq 5) comes from the eigen-decomposition that computes \( W, A, \) and \( V \) at the reference position. For a single structure section, the eigen-decomposition is also needed in conventional semi-analytical methods. Our perturbative expansion is more accurate at a trivial cost of the additional numerical integration. But when it comes to simulating an entire photonic structure, our expansion enables adaptive discretization of the structure along \( z \)-direction, thereby reducing the overall computational cost and outperforming the conventional methods. Next, we describe our algorithm for simulating an entire structure.

**Adaptive Discretization.** We now consider the simulation of an entire photonic structure, not just a single section. We propose an algorithm that adaptively discretizes the structure in \( z \)-direction and simulates wave propagation. Input to our algorithm is a photonic structure and a desired accuracy level (i.e., a real number). The goal here is to achieve the desired accuracy, while reducing the overall computational cost.
Our proposed algorithm is outlined in Algorithm 1. The key idea behind this algorithm is as follows: starting from the entire structure as a single section, it recursively subdivides a structure section into $M$ subsections (Line 6 in Algorithm 1). The subdivision occurs when an estimated simulation error of the current section is larger than the user-specified accuracy level $\alpha$ (Line 2 in Algorithm 1). Next, all subsections are simulated individually, and they may be subdivided further in a recursive way (Line 10). For each subsection, we compute its scattering matrix (Line 1) using eq 5. Lastly, the scattering matrices of individual subsections are combined through the Redheffer star product\(^{13}\) to form the scattering matrix of the parent section (Line 13).

We estimate the error by computing the discrepancy of asymptotic expansions in two consecutive orders (e.g., between zeroth order eq 4 and first order eq 5; see Supporting Information, Section 4), such an error estimation has been used in other asymptotic expansions.\(^{23,24}\) In practice, when a section is subdivided, we subdivide it into three subsections (i.e., $M = 3$), and use the midpoint of each section as the reference position (see Figure 1c). An alternative is to use binary subdivision (i.e., $M = 2$, endpoint reference position), which we will compare in our numerical experiments in Results.

Through recursive subdivision, adaptive discretization in the $z$-direction naturally emerges: regions with rapidly varying cross sections will be more subdivided, and thereby the algorithm automatically uses fine sections to ensure accuracy; meanwhile, in smoothly varying regions, it uses low resolution to save computation. An example is given in Figure 2.

### RESULTS

We conduct numerical experiments to validate the stability and accuracy of our method and compare it with the differential method and RCWA.

---

**Algorithm 1**: varRCWA($z_{\text{min}}$, $z_{\text{max}}$, $\alpha$)

**Input**: lower/upper limit $z_{\text{min}}$, $z_{\text{max}}$, error bound $\alpha$

**Output**: scattering matrix from $z_{\text{min}}$ to $z_{\text{max}}$

**Data**: simulated geometry $G$

1. Evaluate high order scattering matrix $S$ according to Eq. (5) by sampling on $G$.
2. if estimated error $< \alpha$ then
   - return $S$.
3. else
   - Subdivide this section into $M$ subsections $(z_0, z_1), (z_1, z_2), \ldots, (z_{M-1}, z_M)$ where $z_{\text{min}} = z_{0} < z_{1} < \cdots < z_{M} = z_{\text{max}}$.
   - Find a reference point for each subsection.
   - for $i = 1, \ldots, M$ do
     - $S_i(z_{i-1} \rightarrow z_i) \leftarrow \text{varRCWA}(z_{i-1}, z_{i}, \alpha)$.
   - Project the basis of $S_i$ to match $S_{i-1}$
   - (see Supporting Information Section 2).
4. return Redheffer star product of $S_1, \ldots, S_M$.

---

**Figure 2.** Adaptive subdivision process of a waveguide. At each subdivision step, the rapidly varying region (estimated error $\geq \alpha$, red color) is subdivided into $M = 3$ sections, while the smoothly varying sections (blue color) remain unchanged.

---

**Figure 3.** Stability and accuracy. We simulate a waveguide, visualized in (c), with no $z$-direction discretization (i.e., treating it as a single section). (a) We measure the resulting scattering matrix errors as length $L$ increases. (b) We also measure the scattering matrix errors as the cross-sectional shape changes. Its width on the right end is fixed at $2.6 \mu m$, while its left width varies linearly from $2.6$ to $3.7 \mu m$ as the parameter $p$ changes from 0 to 1. The length $L$ is fixed as $1 \mu m$. (d) In addition, we measure the scattering matrix errors as the number of sections increases.

---

ACS Photonics 2022, 9, 3310−3317

ACS Photonics

pubs.acs.org/journal/apchd5

https://doi.org/10.1021/acsphotonics.2c00662

© 2022 American Chemical Society

3313
scattering matrix $S^\ast$ using the conventional RCWA. Because RCWA converges as the resolution of discretization increases, this $S^\ast$ is then treated as a ground truth to measure a max norm error, $\| S - S^\ast \|_{\text{max}}$, where $S$ is the scattering matrix resulting from one of the three methods. In addition, we also benchmark our method against the commercial FDTD software to ensure its accuracy. The results are shown in Support Information, Section 5.

We then test the stability of the three methods against the waveguide length. We progressively increase the length of a waveguide $L$ from 0.05 to 1 $\mu m$, while keeping its left- and right-end cross section unchanged. We plot how the accuracy of these methods changes with respect to $L$ in Figure 3a, wherein we use a light blue dashed line to mark where the error equals to 1. As summarized in Table 1, for single section simulation, RCWA is always stable, regardless of the length of the waveguide. The differential method, however, can easily become unstable as the length increases. Our method in these tests remains stable as the length increases to 1 $\mu m$, especially when we use “midpoint” as the reference position. This is because our method scales as $\delta P_z$ instead of $P_z$.

Next, we test the stability of methods against the variance of each section. We measure the scattering matrix errors as the cross-section varies (controlled by $p$). As shown in Figure 3b, when we simulate it using a single section, the integration length is as long as 1 $\mu m$, and the differential method is always unstable. It becomes stable only when the waveguide length is smaller than 0.1 $\mu m$ (see Figure 3a). In contrast, RCWA and varRCWA are always stable as $p$ changes from 0 to 1. Note that our method has smaller errors when $p \in [0.1, 0.3]$, which implies it is a high-order method. When the variation is small (i.e., $p \to 0$), both RCWA and varRCWA have comparable accuracy. This is because in this case the waveguide is more or less straight, which can be approximated well by the conventional RCWA.

Further, we compare the accuracy of the three methods against $z$-direction resolution. Again consider the waveguide shown in Figure 3c (with $L = 1$ and $p = 1$). We evaluate the scattering matrix errors of the three methods while increasing the $z$-direction resolution. The results are reported in Figure 3d: our method converges faster; as the resolution becomes higher, our method is much more accurate. RCWA and the differential method have similar accuracy when the number of sections is large, although the differential method diverges for a small number of sections. The results also echo Table 1.

**Performance for Short Structures.** Next, we test the performance of our method for analyzing short photonic structures. The shapes of these structures are described in Figure 4a,d. For each structure, the ground-truth scattering matrix is computed using the conventional RCWA with a high resolution in $z$-direction ($N = 1024$). We then use it to measure the accuracy-cost curve for all the methods: for our method, we progressively reduce the error threshold (i.e., $\alpha$ in Algorithm 1), and for each error threshold, we measure the resulting scattering matrix error and the computational time cost. This allows us to plot a curve showing how the accuracy changes over the computation time. Similarly, for the differential method and conventional RCWA, we progressively increase the $z$-direction resolution, and measure the scattering matrix error and computation time. All the timings are measured on a workstation with 8 Intel Xeon(R) E5–1620 CPUs running at 3.60 GHz and an NVIDIA GeForce GTX 1080 GPU, and we report the results on both the CPU and GPU implementation.

The resulting accuracy-cost plots are shown in Figure 4. Given a fixed accuracy level ($10^{-4}$), our algorithm is the fastest on the CPU. Our algorithm with the “midpoint” reference position strategy is $2$–$3x$ faster than the differential method, and $3$–$8x$ faster than conventional RCWA. When implemented on GPUs, our algorithm is still much faster compared to conventional RCWA, but it has similar performance as the differential method, although the latter diverges when the resolution is too low. This is because the differential method only involves such operations as matrix multiplication and linear system solves, which can be easily parallelized on modern GPUs. But both RCWA and our method require eigenvalue decomposition, which is not fully GPU-friendly.

Moreover, the “midpoint” version of both RCWA and varRCWA perform better than the “end point” version, because typically at the midpoint the cross-sectional shape better represents a section’s overall shape than at its end points: the change with respect to the “midpoint” is often smaller than the change with respect to the “end point”. We also highlight that when the differential method or conventional RCWA is used for analyzing a particular photonic structure, there exists no guideline to determine the $z$-direction resolution for a certain accuracy level. For the differential method, the resolution is even harder to determine as it is more prone to the stability issue. In practice, one has to rely on multiple trials to choose a proper resolution, and thus spend more time than what is reported in Figure 4.

**Performance for Long Waveguides.** Our method is more advantageous when simulating long waveguides with length larger than 10 $\mu m$. In fact, many on-chip photonic...
devices are tens and hundreds of microns long. To simulate those long devices, a commercial software such as Lumerical on a quad-core CPU workstation takes up to several hours. In contrast, our method can finish in 10 min on the same workstation.

We demonstrate the results using two devices, one linear taper with a length of 10 μm, and one exponential taper with a length of 20 μm. For a given accuracy level ($10^{-4}$), as shown in Figure 5, our method is 2−3× faster on CPU and 3−4× faster on GPU than the conventional RCWA. The differential method converges very slow for these long structures. Therefore, our method is more than 10× faster on CPU and 3−8× faster on GPU than differential method.

**Performance for Metasurface.** Although our method is originally motivated to simulate on-chip nanophotonic devices, it can also speed up the simulation of metasurfaces, slanted gratings, and photonic crystals with varying cross sections.

Here we use our method to simulate a 3D metasurface with a curved cross-sectional shape along one direction. The metasurface shape and simulation results are shown in Figure 6. In this case, we simulate a metasurface unit and use a periodic boundary condition, and the speedup is similar to that of short waveguides. Our algorithm is 3−4× faster than the differential method, and 6−7× faster than RCWA on CPU. On GPU, our performance is similar to the differential method, 6−7× faster than RCWA.

**Real-World Applications.** Finally, we evaluate our method on two photonic devices reported in literature, namely, a short c/o band splitter and a long TE1−TE0 mode converter. Again, the ground-truth scattering matrices for both cases are obtained by the conventional RCWA with high resolution: 32768 discrete sections for the c/o band splitter and 8192 sections for the mode converter, since the former is more complex in shape. In practice, when the error is around $10^{-4}$, RCWA does not converge for Figure 7c. Therefore, a varRCWA simulation with error bound $5 \times 10^{-5}$ is chosen as the ground truth.

Because of the complex geometry and short length (3 μm) in Figure 7c, differential method is often more suitable since high-order method assumes the smoothness of the geometry. Surprisingly, on CPU, our method performs better than the conventional RCWA.
differential method, which is shown in Figure 7a. On GPU, because of the lack of parallelism of eigenvalue decomposition, our method (and RCWA) is slower for a large error (above $10^{-3}$), but for a smaller error ($10^{-4}$), our method is still better (more than 7× faster; Figure 7b). Compared to RCWA, our method is more than 10× faster on CPU and more than 7× faster on GPU, as shown in Figure 7d.

For the mode convertor shown in Figure 7e, our method is always faster because it has better scaling for long waveguides (18.616 μm). Again for the same accuracy ($10^{-4}$), our method is 4× faster than RCWA and 3× faster than the differential method on GPU.

### CONCLUSION

In summary, we present a high-order semianalytical method, which is best suited for simulating photonic structures whose cross-sectional shapes vary along the propagation direction. In comparison to the differential method, our method is more stable. In comparison to conventional semianalytical methods, our method is also faster and more accurate. Our method also allows the user to specify an accuracy level, and adaptively discretizes the structure along z-direction to achieve the desired accuracy.

In future, this work can be extended in various directions. Currently, when subdividing a section, we always divide it into three subsections uniformly. It is possible to subdivide nonuniformly, according to the local cross-sectional shape of the structure. Moreover, our implementation uses the first-order expansion that we derived, although we can easily extend it to use an even higher order expansion according to the derivation in Supporting Information, Section 1.4.

Lastly, although our high-order method works well for dielectric material, nonphysical field singularities can occur for metallic material—facing the same problem as the differential method and RCWA, especially for simulating metallic gratings. Nonphysical field singularities have been resolved for the differential method and for RCWA. Combining the fast Fourier factorization with our high-order method to handle metallic gratings is beyond the scope of this work but an interesting future direction to explore.

### ASSOCIATED CONTENT

#### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsphotonics.2c00662.

Detailed derivations and notes on the application of this technique (PDF)

### AUTHOR INFORMATION

#### Corresponding Author

Zhi-Wei Zhu — Department of Computer Science, Columbia University, New York, New York 10027, United States; orcid.org/0000-0001-5653-8094; Email: zwny@cs.columbia.edu

#### Author

Ziwei Zhu — Department of Computer Science, Columbia University, New York, New York 10027, United States; orcid.org/0000-0001-9228-1038; Email: cxz@cs.columbia.edu

Complete contact information is available at: https://pubs.acs.org/10.1021/acsphotonics.2c00662

### Funding

This work is supported by National Science Foundation (1816041 and 1910839).

### Notes

The authors declare no competing financial interest.

### REFERENCES

1. Moharam, M.; Gaylord, T. Rigorous coupled-wave analysis of planar-grating diffraction. JOSA 1981, 71, 811–818.

2. Divitt, S.; Zhu, W.; Zhang, C.; Lezec, H. J.; Agrawal, A. Ultrfast optical pulse shaping using dielectric metasurfaces. Science 2019, 364, 890–894.

3. Mohamad, H.; Essaidi, S.; Blaize, S.; Macias, D.; Benche, P.; Morand, A. Fast Fourier factorization for differential method and RCWA: a powerful tool for the modeling of non-lamellar metallic diffraction gratings. Optical and Quantum Electronics 2020, 52, 1–13.

4. Zhu, Z.; Dave, U. D.; Lipson, M.; Zheng, C. Ultra-broadband Nanophotonics via Adaptive Inverse Design. 2021 Conference on Lasers and Electro-Optics (CLEO), CLEO, 2021; pp 1–2.

5. Yee, K. Numerical solution of initial boundary value problems involving Maxwell’s equations in isotropic media. IEEE Transactions on antennas and propagation 1966, 14, 302–307.

6. Piggott, A. Y.; Lu, J.; Lagoudakis, K. G.; Petyukievicz, J.; Babinec, T. M.; Vučković, J. Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer. Nat. Photonics 2013, 59, 374–377.

7. Miller, S. A.; Chang, Y.-C.; Phare, C. T.; Shin, M. C.; Zadka, M.; Roberts, S. P.; Stern, B.; Ji, X.; Mohanty, A.; Gordillo, O. A. J.; et al. Large-scale optical phased array using a low-power multi-pass silicon photonic platform. Optica 2020, 7, 3–6.

8. Liu, V.; Fan, S. S4: A free electromagnetic solver for layered periodic structures. Comput. Phys. Commun. 2012, 183, 2233–2244.

9. Jing, X.; Jin, S.; Tian, Y.; Liang, P.; Dong, Q.; Wang, L. Analysis of the sinusoidal nanopatterning grating structure. Optics & Laser Technology 2013, 48, 160–166.

10. Popov, E.; Nefie, M. Maxwell equations in Fourier space: fast-converging formulation for diffraction by arbitrary shaped, periodic, anisotropic media. JOSA A 2001, 18, 2886–2894.

11. Nefie, M.; Vincent, P.; Petit, R. Sur la theorie du reseau conducteur et ses applications a l’optique. Nouv. Rev. Optique 1974, 5, 65–77.

12. Montiel, F.; Nefie, M.; Peyrot, P. Waveguide confinement of Cerenkov second-harmonic generation through a graded-index grating coupler: electromagnetic optimization. Journal of Modern Optics 1998, 45, 2169–2186.

13. Redheffer, R. Inequalities for a matrix Riccati equation. Journal of Mathematics and Mechanics 1959, 8 (3), 349–367.

14. Helton, J.; Stuckwisch, S. Numerical approximation of product integrals. Journal of Mathematical Analysis and Applications 1976, 56, 410–437.

15. Slavik, A. Product Integration, Its History and Applications; Matfyzpress: Prague, 2007.

16. Roberts, C. M.; Podolskiy, V. A. Rigorous diffraction interface theory. Appl. Phys. Lett. 2017, 110, 171108.

17. Li, J.; Shi, L. H.; Ma, Y.; Ran, Y.; Liu, Y.; Wang, J. Efficient and Stable Implementation of RCWA for Ultrathin Multilayer Gratings: T-matrix Approach without Solving Eigenvalues. IEEE Antennas and Wireless Propagation Letters 2020, 83–87.

18. Blanes, S.; Casas, F.; Oteo, J.; Ros, J. Magnus and Fer expansions for matrix differential equations: the convergence problem. Journal of Physics A: Mathematical and General 1998, 31, 259.

19. Iserles, A.; Norsett, S. P. On the solution of linear differential equations in Lie groups. Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences 1999, 357, 983–1019.

20. Blanes, S.; Casas, F.; Oteo, J.-A.; Ros, J. The Magnus expansion and some of its applications. Physics reports 2009, 470, 151–238.
(21) Takegoshi, K.; Miyazawa, N.; Sharma, K.; Madhu, P. Comparison among magnus/floquet/fer expansion schemes in solid-state nmr. *J. Chem. Phys.* 2015, 142, 134201.

(22) Chu, H. Finite difference approach to optical scattering of gratings. *Advanced Characterization Techniques for Optics, Semiconductors, and Nanotechnologies* 2003, 358–370.

(23) Chen, X.; Zheng, C.; Xu, W.; Zhou, K. An Asymptotic Numerical Method for Inverse Elastic Shape Design. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2014)* 2014, 33, 1–11.

(24) Cochelin, B. A path-following technique via an asymptotic-numerical method. *Computers & structures* 1994, 53, 1181–1192.

(25) Kwiecien, P.; Richter, I. Modeling of plasmonic nanostructures using efficient three dimensional aperiodic rigorous coupled wave analysis. *Laser Science* 2011, JWA41.

(26) NVIDIA, CUDA: Release 11.7. https://developer.nvidia.com/cuda-toolkit, Accessed: 2022−08−03.

(27) Gordillo, O. A. J.; Dave, U. D.; Lipson, M. Bridging between Si and few-mode fiber higher order modes. *CLEO: Science and Innovations* 2020, SM2O.

(28) Lumerical, GDS pattern extraction for inverse designed devices using contours method. https://optics.ansys.com/hc/en-us/articles/15000072852-GDS-pattern-extraction-for-inverse-designed-devices-using-contours-method, Accessed 2022−08−03.

(29) Chen, D.; Xiao, X.; Wang, L.; Yu, Y.; Liu, W.; Yang, Q. Low-loss and fabrication tolerant silicon mode-order converters based on novel compact tapers. *Opt. Express* 2015, 23, 11152−11159.

(30) Popov, E.; Nevière, M.; Gralak, B.; Tayeb, G. Staircase approximation validity for arbitrary-shaped gratings. *JOSA A* 2002, 19, 33–42.

(31) Popov, E.; Nevière, M. Grating theory: new equations in Fourier space leading to fast converging results for TM polarization. *JOSA A* 2000, 17, 1773–1784.