Supplemental Document

Motion-resolved, reference-free holographic imaging via spatiotemporally regularized inversion: supplement

YUNHUI GAO AND LIANGCAI CAO*

Department of Precision Instruments, Tsinghua University, Beijing 100084, China
*clc@tsinghua.edu.cn

Supplement DOI: https://doi.org/10.6084/m9.figshare.24658623

Parent Article DOI: https://doi.org/10.1364/OPTICA.506572

Motion-resolved, reference-free holographic imaging via spatiotemporally regularized inversion: supplemental document

1. ALGORITHM IMPLEMENTATIONS

In this section, we provide detailed instructions on implementing the proposed algorithmic framework for the specific near-field diffuser-modulated ptychographic imaging model described in the main text.

A. Forward Model

Mathematically, the forward imaging process can be divided into the following steps:

(1) Probe illumination. When the probe illuminates the sample, the exit field distribution can be mathematically interpreted as an element-wise multiplication between the sample transmission function and the probe field:

$$U_o = U_i \cdot P, \tag{S1}$$

where U_i and U_o denote the input and output complex field, respectively. P denotes the complex probe field distribution. This can be discretized and vectorized as a multiplication with a diagonal matrix $\mathrm{diag}(\mathbf{p})$. When using an extended and collimated plane wave for illumination, as in our case, the probe modulation is equivalent to an identity mapping (i.e., $\mathrm{diag}(\mathbf{p}) = \mathbf{I}$) and can thus be neglected.

(2) Lateral translation. During acquisition, the sample is laterally translated by the motorized stage. The translated sample field can be calculated with sub-pixel accuracy according to

$$U_o = \mathcal{T}(U_i) \overset{\text{def}}{=} \mathcal{F}^{-1}\left\{\mathcal{F}(U_i)\exp\left[-\mathrm{j}2\pi\left(f_x\frac{l_x}{L_x} + f_y\frac{l_y}{L_y}\right)\right]\right\}, \tag{S2}$$

where $\mathcal{F}$ represents the Fourier transform, with f being the spatial frequency coordinate. l and L are the displacement and the total dimension of the image. The subscripts x and y denote the corresponding spatial dimensions, respectively. j denotes the imaginary unit. The above process, represented by a linear operator $\mathcal{T}$, can be further discretized and vectorized as a matrix $\mathbf{T}$. In ptychography, the sample is translated to a different lateral position for each measurement, and therefore the translation operator corresponding to the kth measurement is denoted as $\mathbf{T}_k$. Because in practical implementation, the translation is calculated through fast Fourier transforms (FFTs), an image cropping operation, denoted as $\mathbf{C}_0$, should be applied to the shifted sample transmission function to avoid boundary artifact and save memory during computation [1].

(3) Free-space propagation from the sample to the diffuser. The diffraction of light in free space can be calculated using the angular spectrum method as [2]

$$U_o = \mathcal{Q}_1(U_i) \overset{\text{def}}{=} \mathcal{F}^{-1}\left\{\mathcal{F}(U_i)\exp\left[\mathrm{j}\frac{2\pi}{\lambda}d_1\sqrt{1-(\lambda f_x)^2-(\lambda f_y)^2}\right]\right\}, \tag{S3}$$

where λ denotes the illumination wavelength, and d_1 denotes the distance between the sample and the diffuser. Similarly, the propagation operator Q_1 can be discretized and vectorized into $\mathbf{Q}_1$. Because of the convolution effect of diffraction, the effective sample area usually has a larger dimension than the modulation area of the diffuser. As a result, an image cropping operation $\mathbf{C}_1$ is applied subsequently to model the limited size of the diffuser. The cropping size is determined by the convolution kernel size, which can be calculated according to the illumination wavelength, the propagation distance, and the sampling interval.

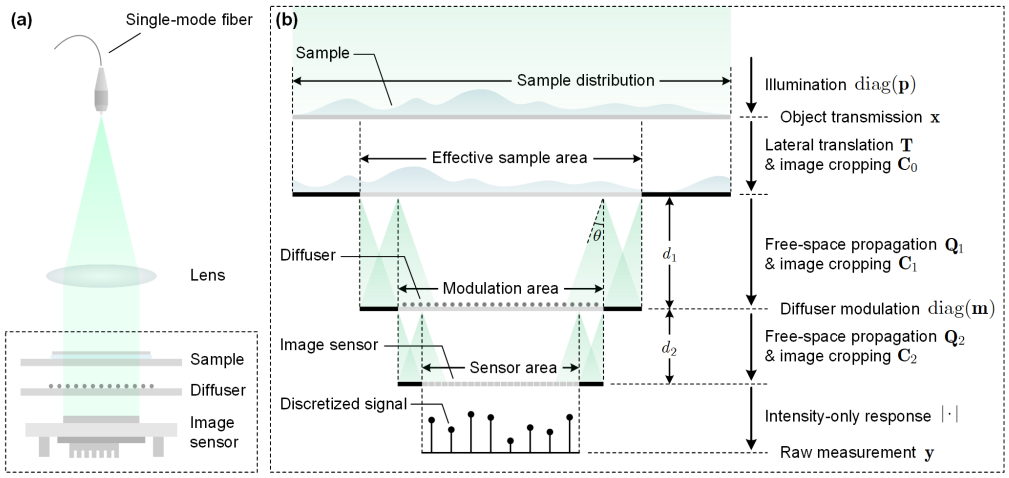

Fig. S1. Illustration of the forward model. (a) Diagram of the optical setup. (b) Enlarged view of (a), where the corresponding mathematical operations are highlighted.

(4) Modulation by the diffuser. The diffusing layer introduces an point-wise modulation to the incident wavefield:

$$U_o = U_i \cdot M, \tag{S4}$$

where M denote the transmission function of the diffuser. The modulation process can also be interpreted as the Hadamard product with the diffuser transmission $\mathbf{m}$, or equivalently as multiplication with a diagonal matrix $\mathrm{diag}(\mathbf{m})$.

(5) Free-space propagation from the diffuser to the sensor. Again, the diffraction is given by

$$U_o = \mathcal{Q}_2(U_i) \overset{\text{def}}{=} \mathcal{F}^{-1}\left\{\mathcal{F}(U_i)\exp\left[\mathrm{j}\frac{2\pi}{\lambda}d_2\sqrt{1-(\lambda f_x)^2-(\lambda f_y)^2}\right]\right\}, \tag{S5}$$

where d_2 denotes the distance between the diffuser and the imaging sensor. After discretization and vectorization, this process can be expressed as a propagation operator $\mathbf{Q}_2$ followed by an image cropping operator $\mathbf{C}_2$.

(6) Sensor response. The sensor pixel responds to the incident wavefield by converting the average power within the active area into a digital signal. This can be mathematically interpreted as calculating the modulus of the complex field:

$$Y = |U_i|. \tag{S6}$$

Based on the above analysis, the vectorized forward model is obtained by taking all the operators in sequence:

$$\mathbf{y}_k = |\mathbf{C}_2\mathbf{Q}_2\mathrm{diag}(\mathbf{m})\mathbf{C}_1\mathbf{Q}_1\mathbf{C}_0\mathbf{T}_k\mathrm{diag}(\mathbf{p})\mathbf{x}| = |\mathbf{A}_k\mathbf{x}|, \tag{S7}$$

where $\mathbf{A}_k = \mathbf{C}_2\mathbf{Q}_2\mathrm{diag}(\mathbf{m})\mathbf{C}_1\mathbf{Q}_1\mathbf{C}_0\mathbf{T}_k\mathrm{diag}(\mathbf{p})$ represents the overall linear measurement matrix in Eq. (1) of the main text. It should be emphasized that vectorization is only used to simplify the algorithm derivation and illustration, and is not required when implementing the algorithms in practice. A MATLAB implementation of the forward model can be found in [3].

B. Inverse Problem Formulation

As mentioned in the main text, phase retrieval is formulated as a regularized inverse problem by combining the forward model and the spatiotemporal priors:

$$\min_{\mathbf{x}} \underbrace{\frac{1}{2}\sum_{k=1}^{K}\||\mathbf{A}_k\mathbf{x}_k| - \mathbf{y}_k\|_2^2}_{F(\mathbf{x})} + \underbrace{\rho_s\|\mathbf{D}_{xy}\mathbf{x}\|_1 + \rho_t\|\mathbf{D}_t\mathbf{x}\|_1}_{R(\mathbf{x})}. \tag{S8}$$

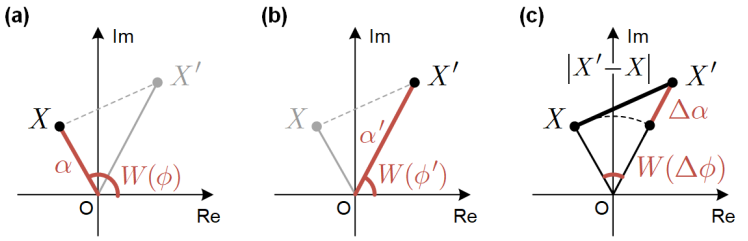

Fig. S2. Conceptual illustration of the complex total variation calculation. (a) and (b) show the values of two neighboring pixels, X and X', in the complex plane. (c) The absolute distance between the two complex numbers $|X' - X|$ is used to quantify the total variation. It can be seen that variations in both amplitude $\Delta\alpha = \alpha' - \alpha$ and wrapped phase $W(\Delta\phi) = W(\phi') - W(\phi)$ contribute to the result and are therefore regularized, where W denotes the phase wrapping function.

More specifically, the regularization term $R(\mathbf{x})$ is the spatiotemporal total variation function with different weights along the x, y and t dimensions:

$$\begin{aligned} R(\mathbf{x}) =& \rho_s \|\mathbf{D}_{xy}\mathbf{x}\|_1 + \rho_t \|\mathbf{D}_t\mathbf{x}\|_1 = \rho_s \|\mathbf{D}_x\mathbf{x}\|_1 + \rho_s \|\mathbf{D}_y\mathbf{x}\|_1 + \rho_t \|\mathbf{D}_t\mathbf{x}\|_1 \\ =& \rho_s \sum_{i=1}^{N_x-1}\sum_{j=1}^{N_y}\sum_{k=1}^{K} \left|X_{i+1,j,k} - X_{i,j,k}\right| + \rho_s \sum_{i=1}^{N_x}\sum_{j=1}^{N_y-1}\sum_{k=1}^{K} \left|X_{i,j+1,k} - X_{i,j,k}\right| + \rho_t \sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\sum_{k=1}^{K-1} \left|X_{i,j,k+1} - X_{i,j,k}\right|, \end{aligned} \tag{S9}$$

where $X_{i,j,k}$ denotes the element indexed by (i, j, k) in the spatiotemporal datacube $\mathbf{X} \in \mathbb{C}^{N_x \times N_y \times K}$. A noteworthy feature of the spatiotemporal total variation regularizer in Eq. (S9) is that it imposes sparsity regularization in the complex domain that involves both amplitude and phase variations, as is illustrated in Fig. S2. For thick samples, our method can directly retrieve the wrapped phases. In contrast, previously reported implementations often treat amplitude and phase separately, therefore requiring phase unwrapping during the iterative reconstruction, which may result in considerable time consumption and instability.

As introduced in the main text, Eq. (S8) can be solved via an accelerated proximal gradient algorithm, which is summarized in Algorithm S1. A detailed derivation and analysis of the algorithm can be found in Ref. [4].

Algorithm S1. Accelerated proximal gradient algorithm

Input: Initial guess $\mathbf{x}^{(0)}$, step size γ, and iteration number I.
Output: Estimate $\mathbf{x}^{(I)}$.

1: $\mathbf{u}^{(0)} = \mathbf{x}^{(0)}$
2: **for** $i = 1$ to I **do**
3: $\quad \mathbf{v}^{(i)} = \mathbf{u}^{(i-1)} - \gamma \nabla_{\mathbf{u}} F(\mathbf{u}^{(i-1)})$ ▷ Gradient update
4: $\quad \mathbf{x}^{(i)} = \mathrm{prox}_{\gamma R}(\mathbf{v}^{(i)})$ ▷ Proximal update via Algorithm S2
5: $\quad \mathbf{u}^{(i)} = \mathbf{x}^{(i)} + \frac{i}{i+3}(\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)})$

C. Step Size Selection for the Accelerated Proximal Gradient Algorithm

Let $\mathbf{y} = \left(\mathbf{y}_1^\mathsf{T}, \mathbf{y}_2^\mathsf{T}, \ldots, \mathbf{y}_K^\mathsf{T}\right)^\mathsf{T} \in \mathbb{R}^{M_x M_y K}$ and $\mathbf{x} = \left(\mathbf{x}_1^\mathsf{T}, \mathbf{x}_2^\mathsf{T}, \ldots, \mathbf{x}_K^\mathsf{T}\right)^\mathsf{T} \in \mathbb{C}^{N_x N_y K}$ denote the concatenated amplitude measurement and sample transmission function, respectively. The forward

model can be expressed as

$$\underbrace{\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_K\end{pmatrix}}_{\mathbf{y}} = \Bigg|\underbrace{\begin{pmatrix}\mathbf{A}_1 & & & \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ & & & \mathbf{A}_K\end{pmatrix}}_{\mathbf{A}}\underbrace{\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_K\end{pmatrix}}_{\mathbf{x}}\Bigg|, \tag{S10}$$

where $\mathbf{A} \in \mathbb{C}^{M_x M_y K \times N_x N_y K}$ is the measurement matrix. The data-fidelity function can then be simplified as

$$F(\mathbf{x}) = \frac{1}{2}\sum_{k=1}^{K} \| |\mathbf{A}_k \mathbf{x}_k| - \mathbf{y}_k \|_2^2 = \frac{1}{2}\| |\mathbf{A}\mathbf{x}| - \mathbf{y} \|_2^2 . \tag{S11}$$

According to Ref. [4], the proximal gradient algorithm converges to a stationary point if the step size γ satisfies

$$\gamma \leq \frac{2}{\rho(\mathbf{A}^{\mathsf{H}}\mathbf{A})}. \tag{S12}$$

Notice that $\mathbf{A}^{\mathsf{H}}\mathbf{A}$ is a block diagonal matrix with main-diagonal blocks $\mathbf{A}_1^{\mathsf{H}}\mathbf{A}_1, \mathbf{A}_2^{\mathsf{H}}\mathbf{A}_2, \ldots, \mathbf{A}_K^{\mathsf{H}}\mathbf{A}_K$. It satisfies

$$\rho(\mathbf{A}^{\mathsf{H}}\mathbf{A}) = \max_k \rho(\mathbf{A}_k^{\mathsf{H}}\mathbf{A}_k). \tag{S13}$$

Substituting Eq. (S13) into Eq. (S12), we arrive at

$$\gamma \leq \frac{2}{\max_k \rho(\mathbf{A}_k^{\mathsf{H}}\mathbf{A}_k)}. \tag{S14}$$

Therefore, an optimal step size of $\gamma = 2/\max_k \rho(\mathbf{A}_k^{\mathsf{H}}\mathbf{A}_k)$ can be used to ensure convergence, enabling highly efficient implementation and reconstruction.

D. The Proximal Solver

The proximal update in Eq. (5) of the main text coincides with a TV-regularized video denoising problem in the following form:

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|_2^2 + \gamma R(\mathbf{x}), \tag{S15}$$

where $\mathbf{v}$ and $\mathbf{x}$ can be interpreted as a noisy observation and the noiseless video to be recovered, respectively. The main difficulty of solving this convex optimization problem arises from the non-differentiability of the TV seminorm. We thus follow the idea from Ref. [5] and consider solving the dual problem instead.

D.1. Dual Formulation

The TV-regularized denoising subproblem can be equivalently expressed as a constrained optimization problem as follows:

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{u}} \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{v}\|_2^2 + H(\mathbf{u}) \\ &\text{s.t. } \mathbf{u} = \mathbf{D}\mathbf{x}, \end{aligned} \tag{S16}$$

where $\mathbf{D} = (\mathbf{D}_x^{\mathsf{T}}, \mathbf{D}_y^{\mathsf{T}}, \mathbf{D}_t^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{3N_xN_yK \times N_xN_yK}$ is the finite difference operator, and H is defined as

$$H(\mathbf{u}) = \rho_s\|\mathbf{u}_x\|_1 + \rho_s\|\mathbf{u}_y\|_1 + \rho_t\|\mathbf{u}_t\|_1, \tag{S17}$$

where $\mathbf{u} = (\mathbf{u}_x^{\mathsf{T}}, \mathbf{u}_y^{\mathsf{T}}, \mathbf{u}_t^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{C}^{3N_xN_yK}$ with $\mathbf{u}_x, \mathbf{u}_y, \mathbf{u}_t \in \mathbb{C}^{N_xN_yK}$. The Lagrangian is given by

$$L(\mathbf{x}, \mathbf{u}, \mathbf{w}) = \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{v}\|_2^2 + H(\mathbf{u}) + \mathrm{Re}\left(\langle \mathbf{w}, \mathbf{D}\mathbf{x} - \mathbf{u}\rangle\right), \tag{S18}$$

where $\mathbf{w} \in \mathbb{C}^{3N_xN_yK}$ is the dual variable, $\mathrm{Re}(\mathbf{x})$ denotes the real part of a complex vector $\mathbf{x}$, and $\langle \cdot, \cdot \rangle$ denotes the inner product. The Lagrange dual function, by definition, is

$$\begin{aligned}
\inf_{\mathbf{x},\mathbf{u}} L(\mathbf{x},\mathbf{u},\mathbf{w}) &= \inf_{\mathbf{x},\mathbf{u}} \left\{ \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{v}\|_2^2 + H(\mathbf{u}) + \mathrm{Re}\left(\langle \mathbf{w}, \mathbf{D}\mathbf{x}-\mathbf{u}\rangle\right) \right\} \\
&= \inf_{\mathbf{x}} \left\{ \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{v}\|_2^2 + \mathrm{Re}\left(\langle \mathbf{w}, \mathbf{D}\mathbf{x}\rangle\right) \right\} + \inf_{\mathbf{u}} \left\{ H(\mathbf{u}) - \mathrm{Re}\left(\langle \mathbf{w}, \mathbf{u}\rangle\right) \right\} \\
&= \inf_{\mathbf{x}} \left\{ \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{v}\|_2^2 + \mathrm{Re}\left(\langle \mathbf{D}^{\mathsf{T}}\mathbf{w}, \mathbf{x}\rangle\right) \right\} - H^*(\mathbf{w}) \\
&= \inf_{\mathbf{x}} \left\{ \frac{1}{2\gamma}\left\|\mathbf{x}-(\mathbf{v}-\gamma\mathbf{D}^{\mathsf{T}}\mathbf{w})\right\|_2^2 \right\} + \frac{1}{2\gamma}\|\mathbf{v}\|_2^2 - \frac{1}{2\gamma}\left\|\mathbf{v}-\gamma\mathbf{D}^{\mathsf{T}}\mathbf{w}\right\|_2^2 - I_S(\mathbf{w}), \quad \text{(S19)}
\end{aligned}$$

where H^* is the convex conjugate of H [6]. For H defined by Eq. (S17), in particular, H^* is the indicator function of a set S, which is denoted by I_S. The set S is a convex set which contains all $\mathbf{w} = (\mathbf{w}_x^{\mathsf{T}}, \mathbf{w}_y^{\mathsf{T}}, \mathbf{w}_t^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{C}^{3N_xN_yK}$ with $\mathbf{w}_x, \mathbf{w}_y, \mathbf{w}_t \in \mathbb{C}^{N_xN_yK}$ satisfying

$$\|\mathbf{w}_x\|_\infty \le \rho_s, \quad \text{(S20)}$$

$$\|\mathbf{w}_y\|_\infty \le \rho_s, \quad \text{(S21)}$$

$$\|\mathbf{w}_t\|_\infty \le \rho_t, \quad \text{(S22)}$$

where $\|\mathbf{w}\|_\infty = \max_i |w_i|$. The last equality in Eq. (S19) suggests that the primal optimal solution is obtained when $\mathbf{x} = \mathbf{v} - \gamma\mathbf{D}^{\mathsf{T}}\mathbf{w}$. The dual problem is to maximize the Lagrange dual function with respect to $\mathbf{w}$, which is equivalent to

$$\min_{\mathbf{w}\in S} \left\{ G(\mathbf{w}) \stackrel{\text{def}}{=} \left\|\mathbf{v}-\gamma\mathbf{D}^{\mathsf{T}}\mathbf{w}\right\|_2^2 \right\}. \quad \text{(S23)}$$

D.2. Solution to the Dual Problem

The problem of Eq. (S23) can be effectively solved using an accelerated gradient projection algorithm [7], where the dual variable is iteratively updated by a gradient step followed a projection operation $\mathcal{P}_S(\cdot)$ onto the set S. The denoising algorithm is presented in Algorithm S2. The Wirtinger gradient of the dual objective function $G(\mathbf{w})$ is given by

$$\nabla_{\mathbf{w}} G(\mathbf{w}) = -\gamma\mathbf{D}\left(\mathbf{v}-\gamma\mathbf{D}^{\mathsf{T}}\mathbf{w}\right). \quad \text{(S24)}$$

The step size can be set to $\eta = 1/(12\gamma^2)$ to ensure convergence, as is shown in the next subsection.

Algorithm S2. Accelerated gradient projection algorithm for the proximal update

Input: Observation $\mathbf{v}$, step size η, and iteration number I.
Output: Dual solution $\mathbf{w}^{(I)}$.

$\mathbf{w}^{(0)} = \mathbf{0}$, $\mathbf{z}^{(0)} = \mathbf{w}^{(0)}$
for $i = 1$ to I **do**
  $\mathbf{w}^{(i)} = \mathcal{P}_S(\mathbf{z}^{(i-1)} - \eta\nabla_{\mathbf{z}} G(\mathbf{z}^{(i-1)}))$
  $\mathbf{z}^{(i)} = \mathbf{w}^{(i)} + \frac{i}{i+3}(\mathbf{w}^{(i)} - \mathbf{w}^{(i-1)})$

D.3. Step Size Selection for the Accelerated Gradient Projection Algorithm

The accelerated gradient projection algorithm for solving Eq. (S23) is a special case of the general accelerated proximal gradient method (also known as FISTA) with the nonsmooth term being an indicator function. For convex objective functions with a Lipschitz continuous gradient, the algorithm has been shown to yield provable convergence behaviors with a step size of $\eta \le 1/L$, where L denotes the Lipschitz constant [8]. The following Lemma establishes an upper Lipschitz bound on $\nabla_{\mathbf{w}} G(\mathbf{w})$.

Lemma 1. *The Lipschitz constant of* $\nabla G(\mathbf{w})$ *is upper bounded by* $12\gamma^2$.

Proof. The first-order Wirtinger derivatives are given by

$$\frac{\partial G}{\partial \mathbf{w}} = -\gamma(\mathbf{v} - \gamma \mathbf{D}^{\mathsf{T}}\mathbf{w})^{\mathsf{H}}\mathbf{D}^{\mathsf{T}} \tag{S25}$$

$$\frac{\partial G}{\partial \bar{\mathbf{w}}} = -\gamma(\mathbf{v} - \gamma \mathbf{D}^{\mathsf{T}}\mathbf{w})^{\mathsf{T}}\mathbf{D}^{\mathsf{T}}. \tag{S26}$$

The second-order derivatives are calculated as follows:

$$\mathbf{H}_{\mathbf{ww}} = \frac{\partial}{\partial \mathbf{w}}\left(\frac{\partial G}{\partial \mathbf{w}}\right)^{\mathsf{H}} = \frac{\partial}{\partial \mathbf{w}}\left(-\gamma \mathbf{D}(\mathbf{v} - \gamma \mathbf{D}^{\mathsf{T}}\mathbf{w})\right) = \gamma^2 \mathbf{D}\mathbf{D}^{\mathsf{T}} \tag{S27}$$

$$\mathbf{H}_{\bar{\mathbf{w}}\mathbf{w}} = \frac{\partial}{\partial \bar{\mathbf{w}}}\left(\frac{\partial G}{\partial \mathbf{w}}\right)^{\mathsf{H}} = \mathbf{0} \tag{S28}$$

$$\mathbf{H}_{\mathbf{w}\bar{\mathbf{w}}} = \frac{\partial}{\partial \mathbf{w}}\left(\frac{\partial G}{\partial \bar{\mathbf{w}}}\right)^{\mathsf{H}} = \mathbf{0} \tag{S29}$$

$$\mathbf{H}_{\bar{\mathbf{w}}\bar{\mathbf{w}}} = \frac{\partial}{\partial \bar{\mathbf{w}}}\left(\frac{\partial G}{\partial \bar{\mathbf{w}}}\right)^{\mathsf{H}} = \frac{\partial}{\partial \bar{\mathbf{w}}}\left(-\gamma \mathbf{D}(\bar{\mathbf{v}} - \gamma \mathbf{D}^{\mathsf{T}}\bar{\mathbf{w}})\right) = \gamma^2 \mathbf{D}\mathbf{D}^{\mathsf{T}}. \tag{S30}$$

The second-order complex Hessian is given by

$$\nabla^2 G = \begin{pmatrix} \mathbf{H}_{\mathbf{ww}} & \mathbf{H}_{\bar{\mathbf{w}}\mathbf{w}} \\ \mathbf{H}_{\mathbf{w}\bar{\mathbf{w}}} & \mathbf{H}_{\bar{\mathbf{w}}\bar{\mathbf{w}}} \end{pmatrix} = \begin{pmatrix} \gamma^2 \mathbf{D}\mathbf{D}^{\mathsf{T}} & \mathbf{0} \\ \mathbf{0} & \gamma^2 \mathbf{D}\mathbf{D}^{\mathsf{T}} \end{pmatrix}. \tag{S31}$$

We now need to determine the upper bound on the eigenvalues of $\nabla^2 G$. Given any $\mathbf{x} \in \mathbb{C}^{N_x N_y K}$, we have

$$\begin{aligned}
\|\mathbf{D}\mathbf{x}\|_2^2 =& \sum_{i=1}^{N_x-1}\sum_{j=1}^{N_y}\sum_{k=1}^{K} |X_{i+1,j,k} - X_{i,j,k}|^2 + \sum_{i=1}^{N_x}\sum_{j=1}^{N_y-1}\sum_{k=1}^{K} |X_{i,j+1,k} - X_{i,j,k}|^2 \\
&+ \sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\sum_{k=1}^{K-1} |X_{i,j,k+1} - X_{i,j,k}|^2 \\
\leq& \sum_{i=1}^{N_x-1}\sum_{j=1}^{N_y}\sum_{k=1}^{K} (|X_{i+1,j,k}| + |X_{i,j,k}|)^2 + \sum_{i=1}^{N_x}\sum_{j=1}^{N_y-1}\sum_{k=1}^{K} (|X_{i,j+1,k}| + |X_{i,j,k}|)^2 \\
&+ \sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\sum_{k=1}^{K-1} (|X_{i,j,k+1}| + |X_{i,j,k}|)^2 \\
\leq& 2\sum_{i=1}^{N_x-1}\sum_{j=1}^{N_y}\sum_{k=1}^{K} (|X_{i+1,j,k}|^2 + |X_{i,j,k}|^2) + 2\sum_{i=1}^{N_x}\sum_{j=1}^{N_y-1}\sum_{k=1}^{K} (|X_{i,j+1,k}|^2 + |X_{i,j,k}|^2) \\
&+ 2\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\sum_{k=1}^{K-1} (|X_{i,j,k+1}|^2 + |X_{i,j,k}|^2) \\
\leq& 12\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\sum_{k=1}^{K} |X_{i,j,k}|^2 \\
=& 12\|\mathbf{x}\|_2^2,
\end{aligned} \tag{S32}$$

where the last inequality follows from the fact that for all $1 \leq i \leq N_x$, $1 \leq j \leq N_y$, and $1 \leq k \leq K$, the term $X_{i,j,k}$ appears at most twelve times in the summation. Therefore, we have

$$\|\mathbf{D}\mathbf{x}\|_2 \leq 2\sqrt{3}\|\mathbf{x}\|_2, \quad \forall \mathbf{x} \in \mathbb{C}^{N_x N_y K}, \tag{S33}$$

which implies that the eigenvalues of $\mathbf{D}$ are less or equal than $2\sqrt{3}$. Consequently, the eigenvalues of $\mathbf{D}\mathbf{D}^{\mathsf{T}}$ is upper bounded by 12. We thus conclude that

$$\nabla^2 G \preceq 12\gamma^2 \mathbf{I}, \tag{S34}$$

where $\mathbf{I}$ denotes the identity matrix. □

Lemma 1 ensures convergence of the fast gradient projection algorithm using a fixed step size of $\eta = 1/(12\gamma^2)$ [8].

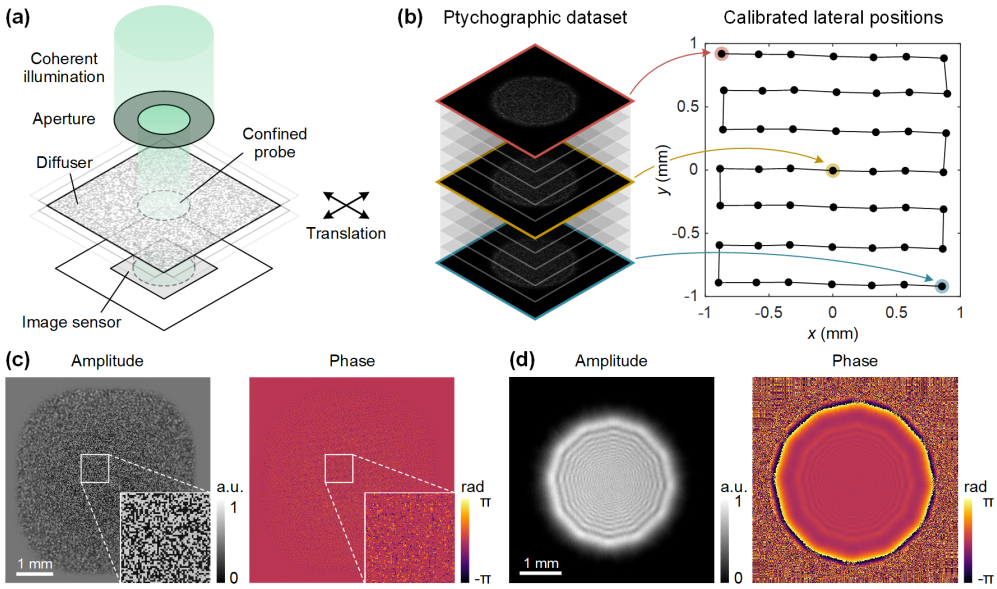

Fig. S3. Diffuser calibration based on near-field ptychography. (a) Experimental setup. The aperture is closed down to generate a confined probe for illumination. (b) The captured raw ptychographic dataset and the translation positions recovered from the modified ePIE algorithm. (c) The retrieved diffuser profile. (d) The retrieved probe profile.

2. EXPERIMENTAL DETAILS

A. System calibration

To accurately model the imaging process, system parameters including propagation distances, diffuser profile, and lateral translation positions need to be experimentally calibrated or estimated during post-processing. Since neither the sample nor the diffuser undergoes axial movement during measurement, the sample-to-sensor and diffuser-to-sensor distances can be calibrated beforehand using autofocusing algorithms from an inline hologram of the sample and the diffuser, respectively.

The amplitude and phase transmission of the diffuser are calibrated by conventional near-field ptychography, as illustrated in Fig. S3. The sample is removed, and the diffuser itself is treated as the object to be imaged. Meanwhile, the aperture of the iris diaphragm is reduced to provide a confined probe for illumination. During calibration, the diffuser is translated to a grid of 7×7 lateral positions in the x-y plane with an interval of approximately 0.3 mm by the 2D motorized stage. A modified extended ptychographic iterative engine (ePIE) with positional estimation is used to jointly recover the diffuser profile, the probe field distribution, and the translation positions from the captured ptychographic dataset [9]. Figure S3(c) shows the retrieved diffuser amplitude and phase. Once calibrated, the diffuser remains stationary throughout the experiment.

During measurement, the sample is continuously scanned at roughly the same speed. But due to inevitable mechanical variations, the positional shift between adjacent frames still needs refinement. To do this, we first obtain a coarse estimation of the sample field using the conventional single-shot method with only spatial-domain regularization. Although the retrieved field is still corrupted by artifacts, a qualitative image of the sample can be resolved. As a result, several stationary control points from the sample can be selected for registration [10].

B. Sample preparation

As a proof-of-concept demonstration, we performed holographic imaging of live paramecia, which are widely found in nature and can be easily obtained from a biological supply company. In the experiment, a drop of paramecium culture containing tens of live paramecia is sandwiched between two coverslips. The droplet is relatively small, allowing us to neglect the axial motion of the sample during the measurement. If this were not the case, frame-by-frame autofocusing would need to be performed to ensure an accurate axial location of the sample.

3. SUPPLEMENTAL RESULTS

A. Simulation

A.1. Comparison of Different Measurement Numbers

To quantitatively evaluate the effect of different numbers of measurements on the reconstruction quality, we tested the algorithm on a stationary object simulated from the Shepp-Logan phantom. Figure S4 shows the results obtained from $K = 2, \ldots, 20$ measurements and under varying signal-to-noise ratios (SNRs). The quality of the reconstruction improves as the number of measurements increases. To achieve similar reconstruction quality, more measurements are required under noisier conditions. $K = 10$ was chosen for most of the experiments as a balance between reconstruction quality and computation time.

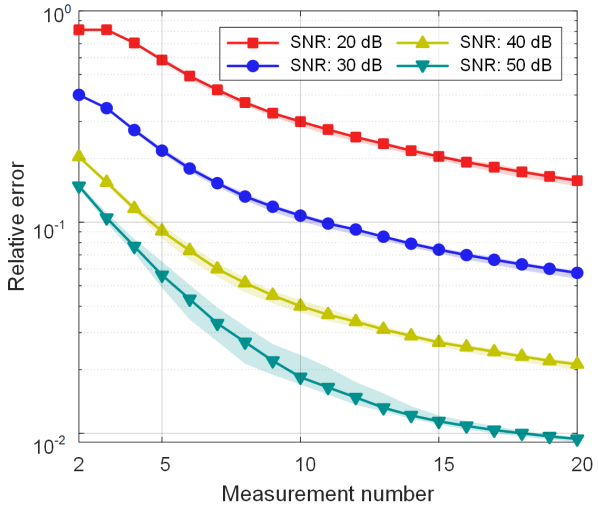

Fig. S4. Relative errors under different measurement numbers and SNRs. The solid lines and the shaded regions indicate the median and the range of the results obtained from ten repeated simulations, respectively.

A.2. Comparison of Different Diffuser Feature Sizes

Figure S5 shows the results obtained under identical simulation settings, but with varying diffuser feature sizes and noise levels. The sample translation speed was set to 4 pixels per frame for all simulations. The results suggest that the reconstruction quality is inversely related to the diffuser feature size. This can be partly explained by the fact that diffusers with finer features provide more uniform modulation across different spatial frequency components, and are more effective in achieving measurement diversity and acquiring redundant information. In the optical experiment, the diffuser has a feature size of 16 μm, corresponding to 5$\sim$6 pixels. This is due to the tradeoff between fabrication and calibration accuracy and modulation efficiency.

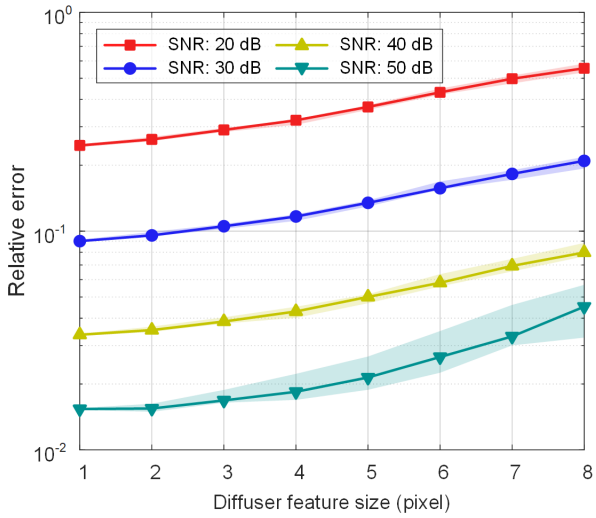

Fig. S5. Relative errors under different diffuser feature sizes and SNRs. The solid lines and the shaded regions indicate the median and the range of the results obtained from ten repeated simulations, respectively.

A.3. Comparison of Different Diffuser Types

Figure S6 presents the results obtained using six different types of diffusers. The diffusers provide amplitude-only, phase-only or complex modulation to the wavefront, and their profiles are either binary ($0/1$ for amplitude and $0/\pi$ for phase) or uniformly distributed (between $[0, 1]$ for amplitude and $[0, 2\pi]$ for phase). The results indicate that phase-only diffusers are the most effective due to their superior efficiency and ability to offer fully random modulation in the complex plane. For amplitude-only and complex diffusers, binary profiles are preferred due to their higher contrast. Considering that amplitude diffusers are simpler and less costly to fabricate, we used an amplitude-only diffuser with a binary pattern in the experiments.

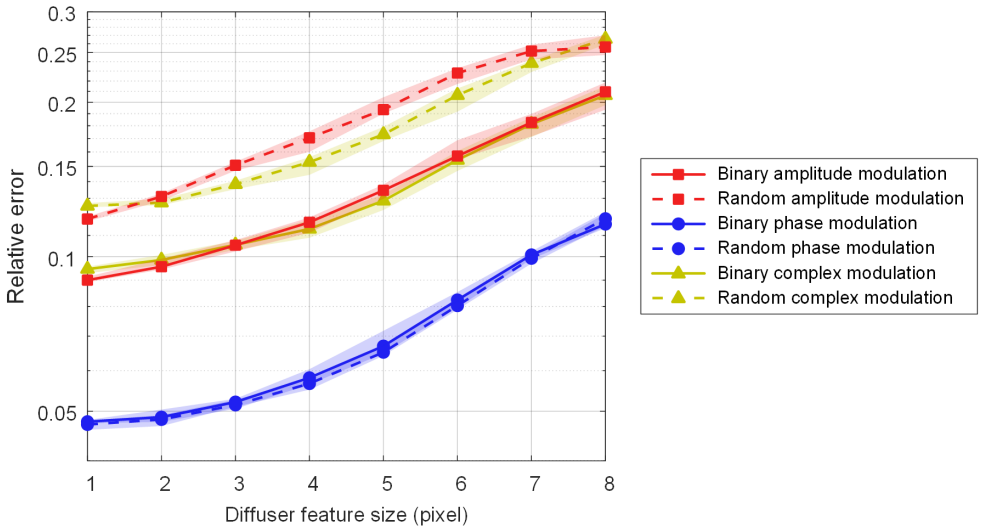

Fig. S6. Relative errors under different diffuser feature sizes and modulation types. The solid lines and the shaded regions indicate the median and the range of the results obtained from ten repeated simulations, respectively.

A.4. Holographic Reconstruction of Dynamic Samples with Varying Speeds

To evaluate the optimal choice of the regularization parameters, we tested the algorithm on simulated samples with different translation and rotation speeds. Specifically, we use the same sample based on the Shepp-Logan phantom but with different moving speeds, as shown in Figs. S7. In Fig. S7(a), the sample is completely motionless, and the reconstruction error decreases monotonically as ρ_t increases. In principle, for motionless samples, STRIVER can be reduced to a conventional multi-frame phase retrieval algorithm as $\rho_t \to \infty$. From Figs. S7(b), (c) and (d), it can be seen that for a larger sample movement, a smaller ρ_t is preferred. The relative error of the optimal case increases as the sample moves faster. It is expected that STRIVER based on the spatiotemporal total variation regularizer is suitable for moderate sample movements. The performance can be further improved by exploiting advanced video priors that can capture long-term spatiotemporal features.

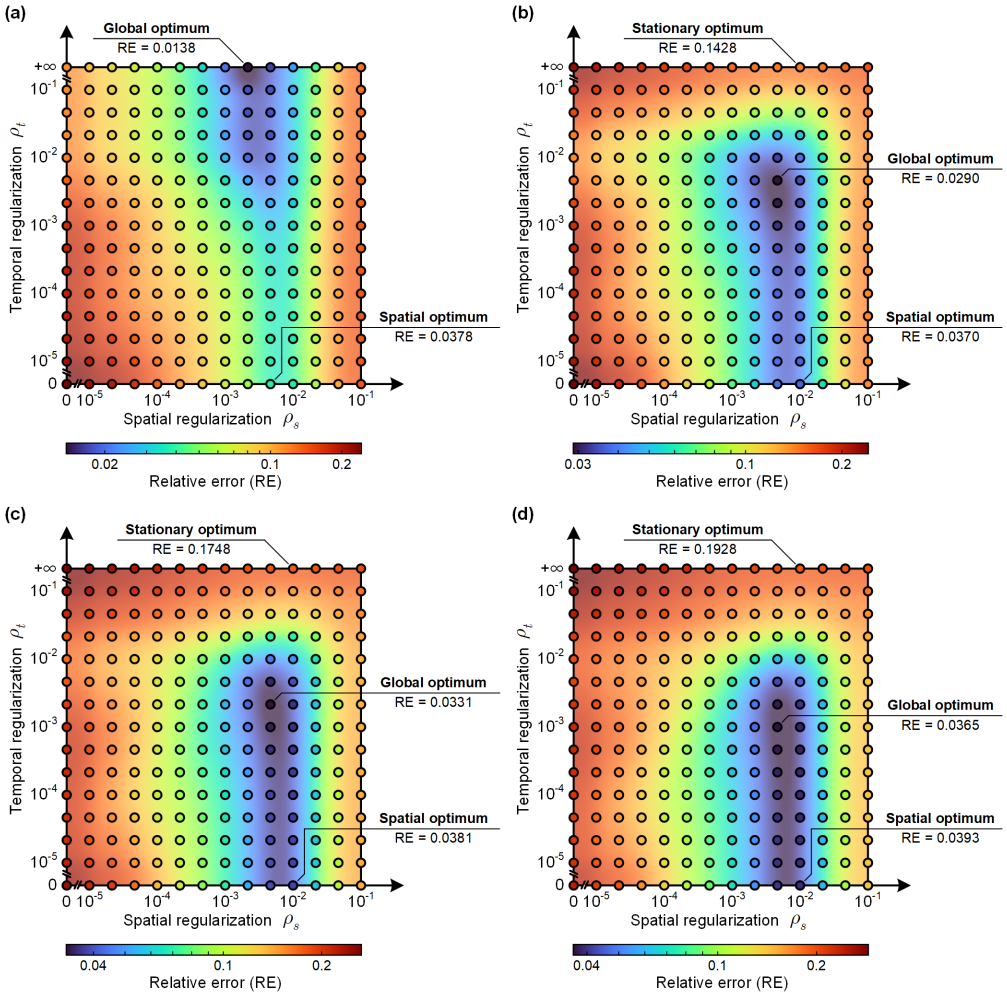

Fig. S7. Regularization parameter evaluation under different sample speeds: (a) stationary sample, (b) 1 pixel/frame translation and 1°/frame rotation, (c) 2 pixel/frame translation and 2°/frame rotation, and (d) 3 pixel/frame translation and 3°/frame rotation.

A.5. Holographic Reconstruction under Different Noise Levels

As indicated by Eq. (2) in the main text, the regularization parameters ρ_s and ρ_t serve as a balance between model consistency and prior knowledge. As a result, their choices are also dependent on the measurement noise level. To quantify this dependency, we performed STRIVER reconstruction with identical settings to those described in the main text, but under varying noise levels. Additive white Gaussian noises were added to the recorded intensity images with an SNR of 30 dB, 40 dB, and 50 dB, as shown in Figs. S8(a), (b) and (c), respectively. The noiseless case is also studied and the results are shown in Fig. S8(d). With reduced noises, smaller ρ_s and ρ_t can be used to obtain a high reconstruction quality.

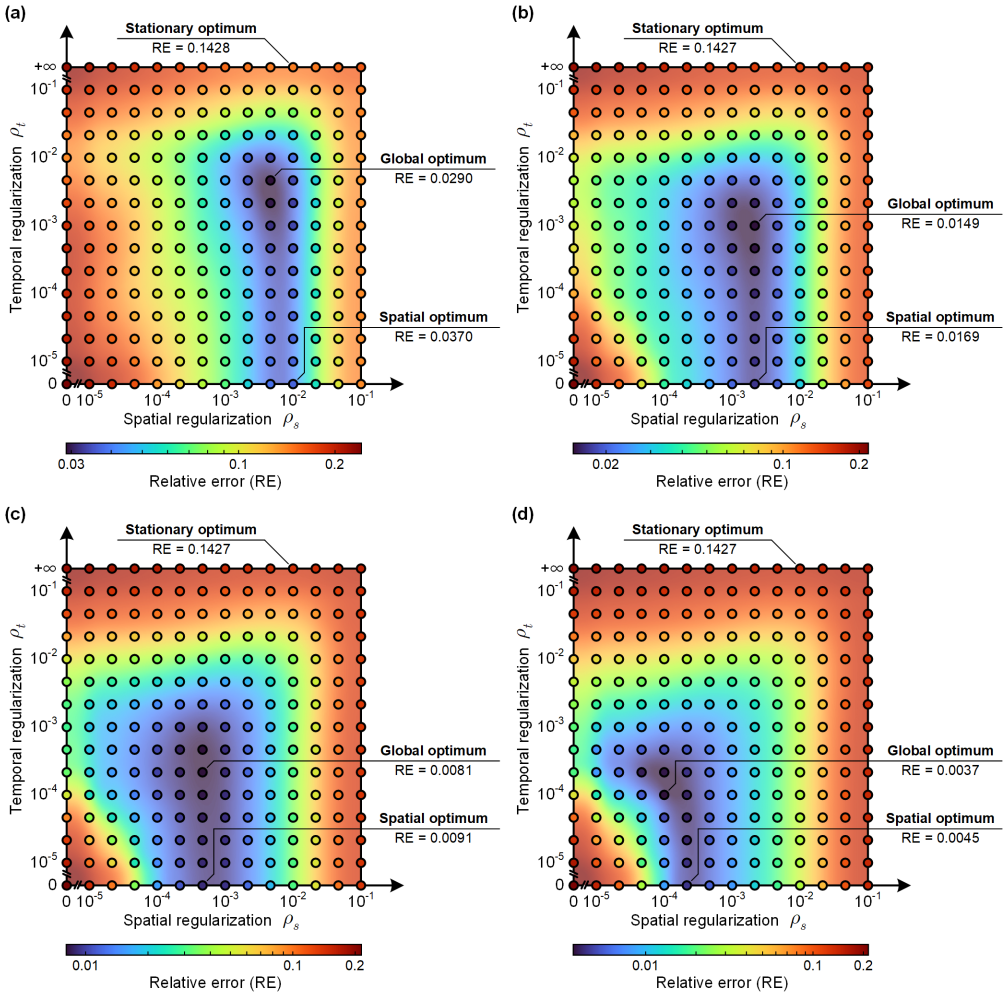

Fig. S8. Regularization parameter evaluation under different noise levels: (a) SNR = 30 dB, (b) SNR = 40 dB, (c) SNR = 50 dB, and (d) noiseless case.

A.6. Holographic Reconstruction of Different Dynamic Samples

To quantitatively study the generalizability of STRIVER, we synthesized different dynamic complex fields from publicly available video datasets [11, 12]. Figures S9, S10 and S11 show the reconstruction of *salesman*, *traffic* and *coastguard* video with identical simulation settings to those described in the main text. It can be seen from the results that the optimal choice of the regularization parameters is not very sensitive to the content of the scene. The optimal ρ_t is slightly larger for the *salesman* sample, because the sample exhibits less movements than the other two. In all three cases, STRIVER outperforms conventional methods according to both quantitative and visual comparisons, indicating that the complex spatiotemporal total variation regularizer used in this work is generally applicable to various dynamic scenes.

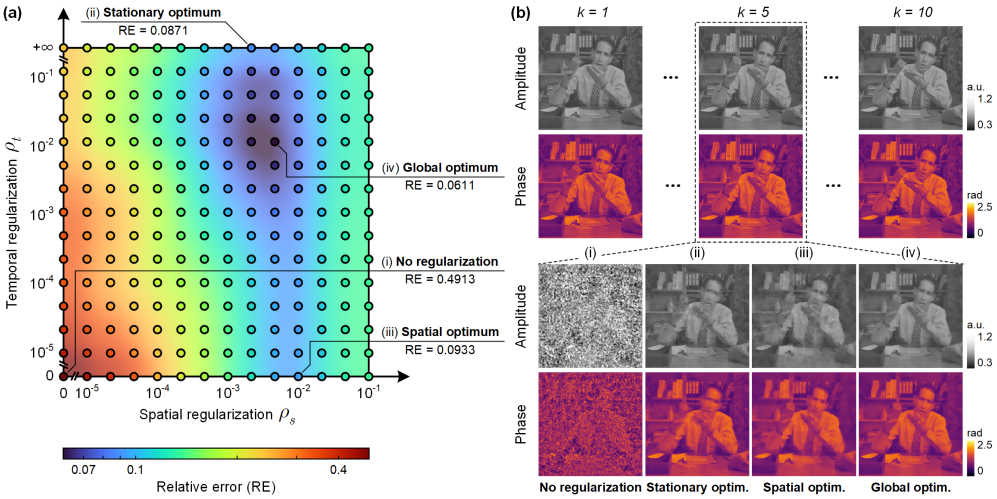

Fig. S9. (a) Regularization parameter evaluation for the *salesman* video. (b) Visualized results of representative cases.

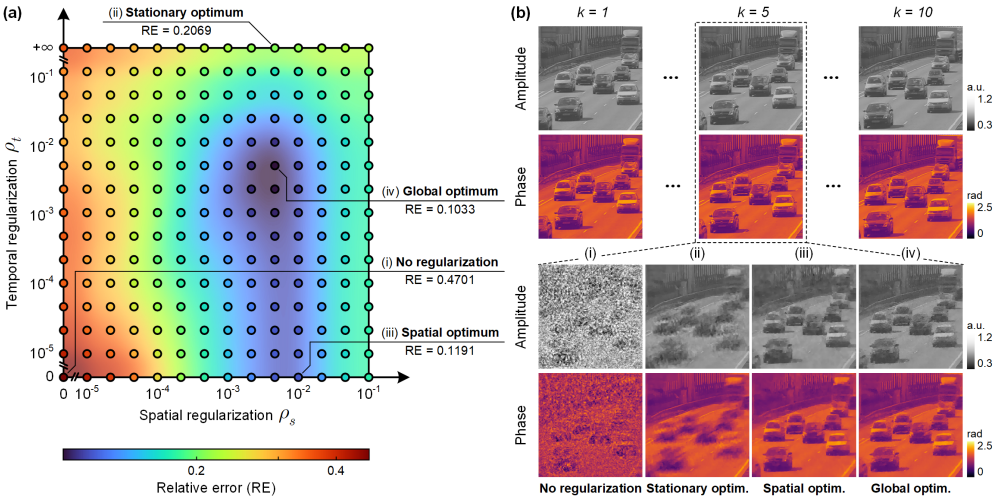

Fig. S10. (a) Regularization parameter evaluation for the *traffic* video. (b) Visualized results of representative cases.

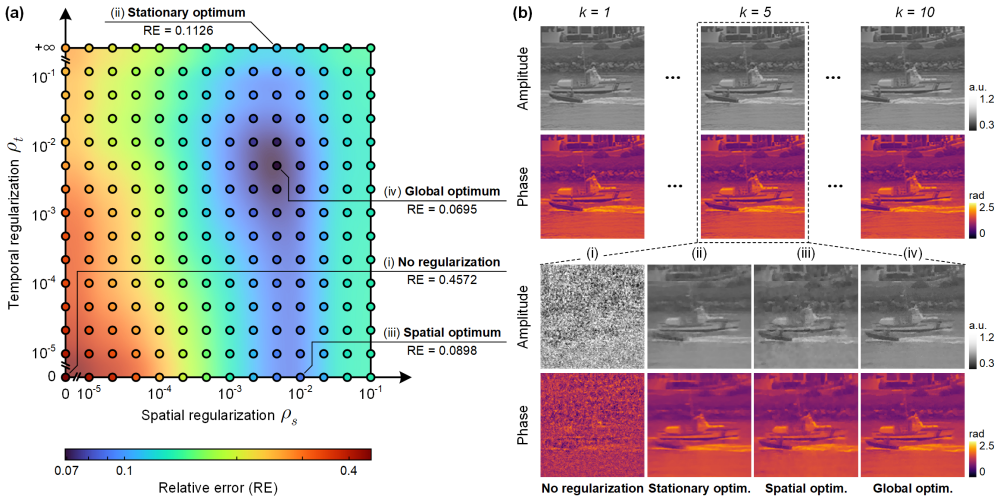

Fig. S11. (a) Regularization parameter evaluation for the *coastguard* video. (b) Visualized results of representative cases.

A.7. Runtime Evaluation

The numerical operations involved in the reconstruction algorithm support highly parallel computation. Therefore, an accelerated version using a graphical processing unit (GPU) card is also implemented in MATLAB [3], which significantly reduces the computation time. The runtimes for 200 iterations using CPU and GPU were quantitatively evaluated on a laptop computer with an Intel Core i7-12700H (2.30 GHz) CPU and an Nvidia GeForce RTX 3060 GPU. The results are shown in Fig. S12.

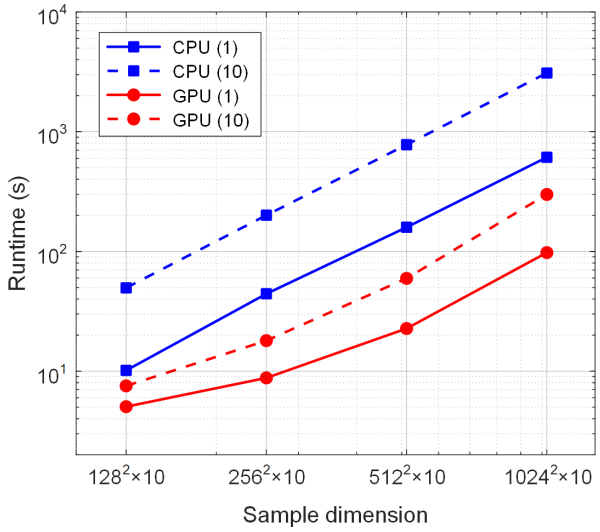

Fig. S12. Runtimes (200 iterations) using CPU and GPU for different sample dimensions. The number in the parenthesis denotes the subiteration number for the proximal update.

B. Experiment

To further evaluate compare STRIVER with the conventional method, we implemented the latter with different diversity measurements, namely $K = 1, 2, 3, 5$, and 10. The results are shown in Fig. S13.

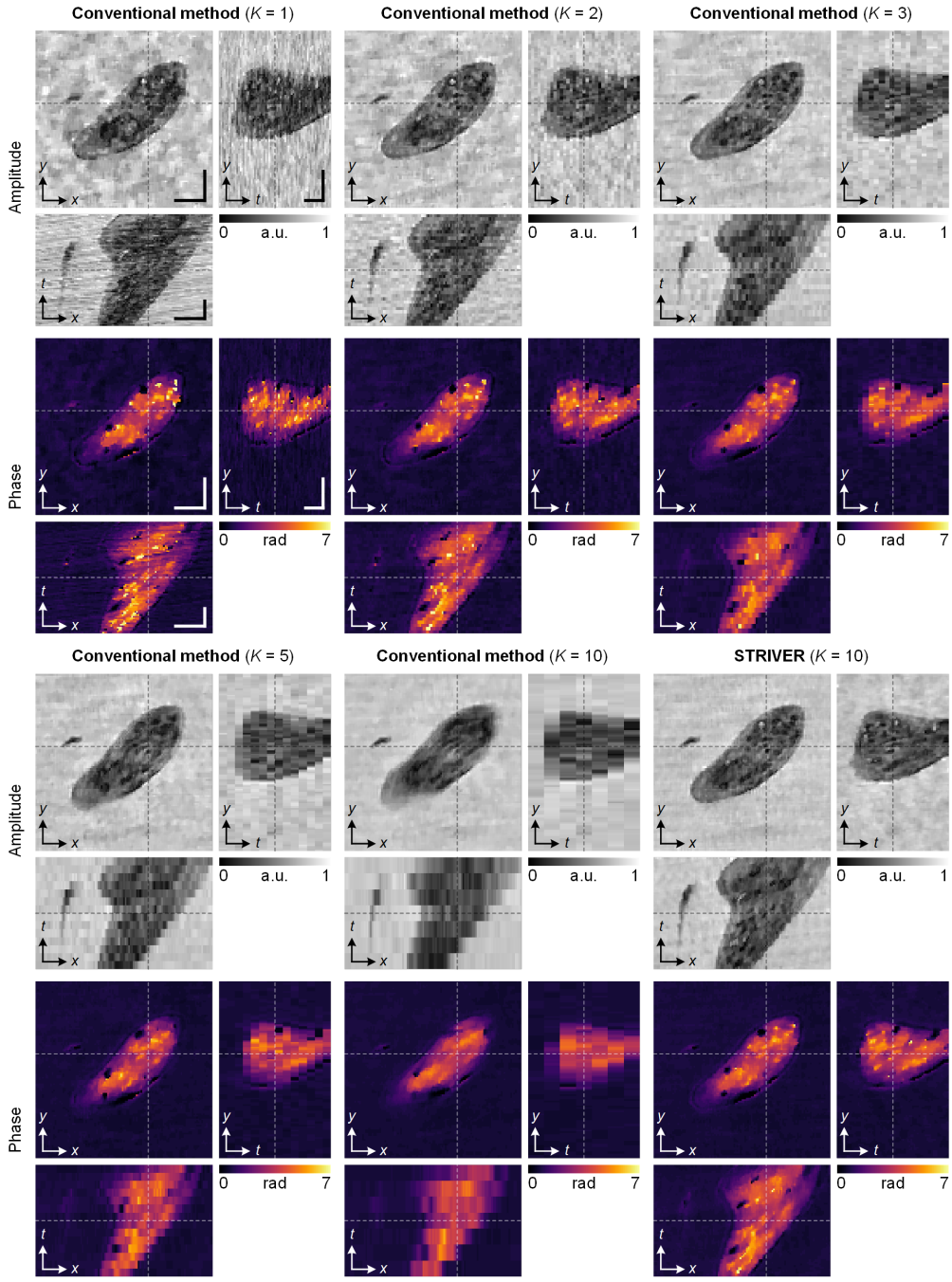

Fig. S13. Comparison of different reconstruction algorithms on a live paramecium sample. 2D amplitude and unwrapped phase slices from the 3D spatiotemporal datacube are shown for the results obtained by the conventional method with different number of diversity images ($K = 1, 2, 3, 5$, and 10) and STRIVER ($K = 10$). Scale bars are 50 μm (x, y) and 0.1 s (t).

REFERENCES

1. A. Matakos, S. Ramani, and J. A. Fessler, "Accelerated edge-preserving image restoration without boundary artifacts," IEEE Transactions on Image Process. **22**, 2019–2029 (2013).
2. J. W. Goodman, *Introduction to Fourier Optics* (Roberts and Company publishers, 2005).
3. https://github.com/THUHoloLab/STRIVER.
4. Y. Gao and L. Cao, "Iterative projection meets sparsity regularization: towards practical single-shot quantitative phase imaging with in-line holography," Light. Adv. Manuf. **4**, 1–17 (2023).
5. A. Chambolle, "An algorithm for total variation minimization and applications," J. Math. Imaging Vis. **20**, 89–97 (2004).
6. S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization* (Cambridge university press, 2004).
7. A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," IEEE Transactions on Image Process. **18**, 2419–2434 (2009).
8. A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," SIAM J. on Imaging Sci. **2**, 183–202 (2009).
9. F. Zhang, I. Peterson, J. Vila-Comamala, A. Diaz, F. Berenguer, R. Bean, B. Chen, A. Menzel, I. K. Robinson, and J. M. Rodenburg, "Translation position determination in ptychographic coherent diffraction imaging," Opt. Express **21**, 13592–13606 (2013).
10. M. Guizar-Sicairos, S. T. Thurman, and J. R. Fienup, "Efficient subpixel image registration algorithms," Opt. Lett. **33**, 156–158 (2008).
11. https://media.xiph.org/video/derf/.
12. J. Yang, X. Yuan, X. Liao, P. Llull, D. J. Brady, G. Sapiro, and L. Carin, "Video compressive sensing using gaussian mixture models," IEEE Transactions on Image Process. **23**, 4863–4878 (2014).