“In-chip” microstructures and photonic devices fabricated by nonlinear laser lithography deep inside silicon

SUPPLEMENTARY INFORMATION

Onur Tokel[1], Ahmet Turnalı[2], Ghaith Makey[1], Parviz Elahi[1], Tahir Çolakoglu[3], Emre Ergeçen[4], Özgün Yavuz[2], René Hübner[5], Mona Zolfaghari Borra[3,6], Ihor Pavlov[1], Alpan Bek[3,6,7], Raşit Turan[3,6,7], Denizhan Koray Kesim[2], Serhat Tozburun[8], Serim Ilday[1] and Fatih Ömer Ilday [1,2,9†]

[1]Department of Physics, Bilkent University, Ankara, 06800, Turkey

[2]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, 06800, Turkey

[3]The Center for Solar Energy Research and Applications, Middle East Technical University, Ankara, 06800, Turkey

[4]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, USA

[5]Helmholtz-Zentrum Dresden - Rossendorf, Institute of Ion Beam Physics and Materials Research, Bautzner Landstraße 400, 01328 Dresden, Germany

[6]Micro and Nanotechnology Graduate Program, Middle East Technical University, Ankara, 06800, Turkey

[7]Department of Physics, Middle East Technical University, Ankara, 06800, Turkey

[8]Harvard Medical School, Boston, MA, 02115, USA

[9]UNAM - National Nanotechnology Research Center and Institute of Materials Science and Nanotechnology, Bilkent University

† To whom correspondence should be addressed: ilday@bilkent.edu.tr

Table of Contents

1. Comparison to Previous Work

Most of the previous work on subsurface modification of silicon focused on using ultrafast lasers. The first attempt appears to be by Nejadmalayeri, et al., who have reported creation of optical waveguides[1]. However, the waveguides were confined to the immediate vicinity of the surface and could not be created at arbitrary depths inside silicon. Additional efforts did not succeed in creating truly subsurface modifications without damaging surface either due to absorption or strong plasma-shielding effects[2,3]. In these studies, either the beam damages chip surface, or laser intensity is clamped by plasma effects, precluding subsurface modifications[2,3]. Even the use of very high pulse energies up to 90 μJ has failed to induce subsurface modifications[4]. These limitations, which appear to stem from the complex nonlinear effects arising in Si are analyzed in refs[4,5]. Thus, these studies were limited to fabrication on interfaces, such as on the interface of silicon and silica[1], nanopatterns[6] or backsurface damage[7].

Our early efforts in subsurface processing of Si with lasers were made with nanosecond pulses and were first reported in[8,9]. Similar approaches were subsequently reported by several other groups. While two-photon absorption (TPA) was suggested as the mechanism leading to subsurface modification[10], this is difficult to understand, given the TPA coefficient of silicon: Considering a TPA of 0.7 cm/GW[11], the laser beam would have to propagate approximately 5 mm, a distance that greatly exceeds the silicon sample's thickness to impart merely %30 of its energy. Conversely, the beam would transfer only approximately 3 nJ of energy localized on the observed features, corresponding to an energy density that is several orders of magnitude below the estimated modification threshold in Si ($\sim$7x10^3 J/cm^3)[10]. Furthermore, in-chip devices or 3D sculpturing through chemical etching have not been reported to date. Recently, claims of laser writing of waveguides have been reported[12,13]. However, this manuscript did not provide evidence of guidance of light, only side image of scattered light from a linear structure. Based on our results showing that the nanosecond laser-induced regions result in depressed optical indices, it is not entirely clear if the reported structure, that requires a higher index, was an actual waveguide.

2. Nonlinear Feedback in Silicon

Silicon (Si) based functional electronic and photonic devices are manufactured with mainly planar architectures on wafer surfaces. Current non-planar chip architectures are limited to vertically stacked dies, silicon-via geometries or wire-bonds around chips[14–16]. Recent optical lithography and laser-writing methods similarly do not exploit the bulk of silicon[17,18]. In order to build in the third dimension, we exploit nonlinear feedback mechanisms that naturally arise from the interaction of Si and infrared laser pulses. This leads to the formation of laser-driven self-organised structures directly and at any selected position deep inside Si, without damaging the surface.

We regard the approach outlined here as an extension to the third dimension for nonlinear laser lithography (NLL), which we developed recently for surface nanostructuring[19]. This NLL technique is an enabling technology, where the third dimension in Si can be exploited for the first time for functional devices, while keeping the wafer surface unaltered. In what follows, we detail the physical mechanisms, *i.e.*, carrier generation, thermal nonlinearity, pulse propagation and nonlinear feedback mechanisms, responsible in creating the building blocks of complex structures.

2.1 Carrier generation and recombination mechanisms in silicon

Pulse propagation and thermal profile in Si are both affected by the carrier dynamics. Therefore, we first analyse the sources of free electrons and their recombination mechanisms in Si to first arrive at the carrier profile, and later the temperature profile at the relevant time scale. We later show how carrier and temperature profiles relate to the main nonlinear effects in pulse propagation, *i.e.*, free carrier induced diffraction and thermal focusing, in the feedback mechanism for the formation of subsurface structures.

Silicon is transparent to photons with wavelengths higher than 1.12 μm. However, for high intensities, two-photon-absorption (TPA) can excite an electron to the conduction band. Unlike linear absorption, the absorption coefficient for TPA is proportional to intensity, and can result in an intensity profile strongly deviating from the Beer-Lambert Law. Further, the generated carriers in Si will both diffuse and recombine, resulting in a quasi-steady state carrier profile. Thus, we analyze the carrier generation, recombination and

diffusion mechanisms in Si.

The dominant carrier generation mechanism in the experiment is the TPA process. The free carrier generation rate due to TPA is given as $\xi = \beta I(t)^2/2hf$, where β is the TPA coefficient, $I(t)$ is the instantaneous intensity and f is the photon frequency. For every two photons absorbed, an electron-hole pair is created. The equal electron and hole densities are denoted by N. The dominant recombination mechanism is the Auger recombination process for carrier densities N>10^{18} - 10^{19} cm^{-3}. In this regime, other recombination processes such as radiative and Shockley-Read-Hall (SRH) processes can be neglected[20]. A simple estimate shows that a typical laser pulse in our experiments can easily reach to these carrier densities. For a single pulse, the final carrier density without recombination can be estimated as:

$$N_{\text{avg}} = \int_{-\infty}^{\infty} \frac{\beta I(t)^2}{2hf} dt = \int_{-\infty}^{\infty} \frac{\beta I_0^2 \exp\left(-2t^2/\tau_{\text{pulse}}^2\right)}{2hf} dt = \frac{\sqrt{\pi}\beta I_0^2}{2\sqrt{2}hf} \tau_{\text{pulse}}, \tag{1}$$

where β is the TPA coefficient, τ_{pulse} is the pulse duration and I_0 is the peak intensity intensity. For our experimental parameters ($\tau_{\text{pulse}} = 3$ ns, $P_{\text{avg}} = 2$ W, $w_0 = 3$ μm and $I_0 = 250$ W/μm^2), we obtain a carrier density in the order of 10^{21} cm^{-3}. However, before reaching this value, Auger recombination will limit the carrier density to a lower value.

With the preceding considerations, the equation governing the evolution of carrier density is given as:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{\beta I^2}{2hf}\left(1 - \frac{N}{N_{\text{at}}}\right) - \gamma_3 N^3, \tag{2}$$

where γ_3 is the Auger recombination constant[21], β is the TPA coefficient, N_{at} is the atomic density in the crystal, I is laser intensity and hf gives the photon energy. Here, the first term on the right is due to TPA, while the second term is due to Auger recombination. From the latter, we observe that the carrier recombination time is $\tau_{\text{rec}} = 1/\gamma_3 N^2$, which becomes comparable to the pulse duration for high carrier densities ($N \approx 10^{19}$ cm^{-3}). Thus, the relevant time scale for the carrier rate equation is determined by the average carrier density (since the recombination time is a function of carrier density) and the pulse length (several nanoseconds). To get a better feeling for the carrier dynamics in the presence of recombination,

Equation 2 can be solved numerically (Fig. S1). The carrier density is in the order of $10^{19} - 10^{20}$ cm^{-3} range, consistent with previous assumptions. Further, in this carrier density range, the recombination time and pulse duration are comparable, thus the density profile closely follows the pulse profile. We will simply use pulse duration in the following estimations for diffusion.

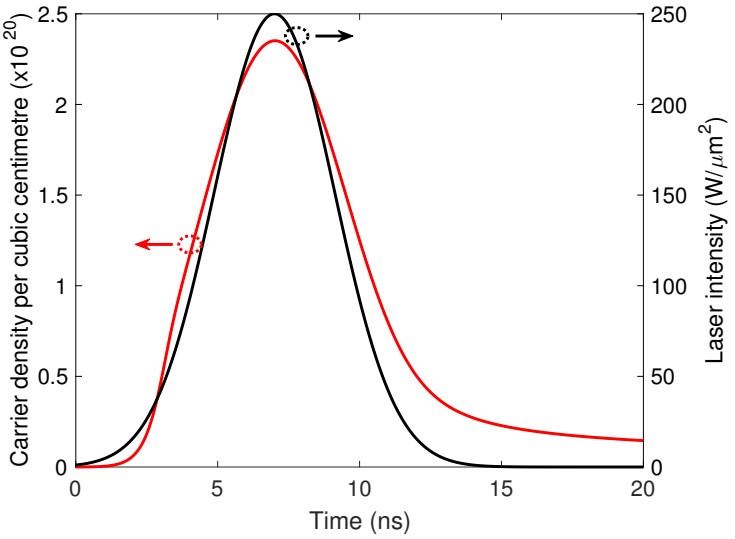

Figure S1: Carrier density and laser intensity during a single pulse. The values are obtained from the numerical solution of Eq. 2 for typical experimental parameters. The black curve indicates the intensity profile, and the red curve indicates the carrier density.

The diffusion term is ignored In Eq. 2. However, if the carriers can escape the excitation volume before recombining or before the pulse ends, carrier diffusion should be accounted for. Fortunately, a simplification arises here, for the carrier diffusion can be neglected if the diffusion length in the relevant time scale (either recombination time or pulse duration) is less than the characteristic distance (beam diameter at the focal plane). Therefore, if $L_{\text{diff}} = \sqrt{6Dt_{\text{pulse}}} \ll S$, where L_{diff} is the diffusion length for carrier diffusion, D is the corresponding diffusion constant, t_{pulse} is the pulse duration and S is the smallest diameter of the beam in Si, carrier diffusion can be ignored. In this section, we show that the carrier diffusion can be neglected in this work, and in the next section we will show that heat diffusion can similarly be ignored.

The carrier diffusion constant in Si is, $D_{\text{carrier}} = 2\text{k}_\text{B}T_\text{e}\ \mu_\text{e}\mu_\text{h}/(e(\mu_\text{e} + \mu_\text{h}))$, where T_e is the electron

temperature, and μ_e and μ_h are the electron and hole mobilities respectively. The diffusion length at T = 300 K is found as, $L_{diff} = \sqrt{6 D_{carrier} \tau_{pulse}} = 1.8$ μm, which is already smaller than the beam width at focus (3 μm). Moreover, μ_h and μ_e terms strongly depend on temperature, thus the increased temperature will significantly reduce the diffusion length. Further, since the recombination time decreases with increasing carrier density, a significant portion of carriers will recombine before they diffuse out of the beam. Therefore, carrier diffusion can safely be neglected in Eq. 2.

A similar simplification arises for the contribution of avalanche ionisation to carrier generation. During avalanche ionisation (impact ionisation), an energetic carrier may interact with valence band electrons to create an electron-hole pair, while losing some of its energy. This effect can be introduced to the carrier rate equation simply by modifying the source term as:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = (\frac{\beta I^2}{2hf} + \delta N)(1 - \frac{N}{N_{at}}) - \gamma_3 N^3. \tag{3}$$

The avalanche ionisation coefficient, δ, is a function of electron temperature (T_e), and can be written as $\delta(T_e) = 3.6 \times 10^{10} \exp(-1.5 E_g / k_B T_e)$, where k_B is the Boltzmann constant[22]. Impact ionisation is physically the inverse of Auger recombination mechanism and may be expected to remove the heat effects of Auger recombination. However, we show that Auger recombination is dominant in this work. The maximum avalanche ionisation rate is equal to 2.3×10^5 s^{-1} at the Si melting temperature, T = 1600 K. For $N = 7.8 \times 10^{17}$ cm^{-3}, the Auger recombination rate is already equal to this rate. In our case, the carrier densities can be much higher, rendering avalanche ionisation rates a few orders of magnitude smaller than the Auger recombination rate. Therefore, we ignore avalanche ionisation.

We next analyse the thermal nonlinearities and temperature profile in Si. The detailed assumptions about electron temperature, lattice temperature, heat diffusion and how the temperature profile is related to the carrier density are discussed in the next section.

2.2 Nonlinear heating mechanisms and temperature profile in silicon

In order to find the thermal profile in silicon during heating by laser beams, we first assess the thermal diffusion and heat generation mechanisms in Si. The temperature evolves according to the heat equation,

$$\rho C_{\mathrm{p}}(T)\frac{\mathrm{d}T}{\mathrm{d}t} - \nabla.(\kappa\nabla T) = Q, \tag{4}$$

where ρ is the Si density (2.33 g/cm^3), $C_{\mathrm{p}}(T)$ is the specific heat[23], κ is the thermal conductivity (1.6 W/cm K)[24], T is lattice temperature and Q is the heat generation rate. A simplification arises if one notices that the heat diffusion length is $L_{\mathrm{diff}} = \sqrt{6D_{\mathrm{heat}}\tau_{\mathrm{pulse}}} \ll S$, where S is the laser spot size and $D_{\mathrm{heat}} = \frac{\kappa}{\rho C_{\mathrm{p}}}$. Thus the second term on the left hand side of Eq. 16 can be ignored, and the heat generation translates to ΔT.

We note that, in the analysis of heat sources, we exploit the single temperature model. During the laser-Si interaction, laser energy is first transferred to electrons, and these carriers, in turn go on to thermalize through electron-electron collisions. After thermalization, carriers with temperature T_{e} will behave as an external heat bath for the lattice. Carrier and lattice temperatures will then equilibrate on a time scale defined by the electron-phonon coupling time. In Si, energetic electrons relax in less than 500 fs through carrier - LO phonon channel[22]. Since the pulse duration in this work (few ns) is much larger than this value, one can assume the electron and lattice temperatures are always in equilibrium ($T_{\mathrm{e}} = T$).

With the preceding considerations, the heat generation mechanisms and their contributions in Si are:

1. Two-photon absorption based heating: If the total energy of the two absorbed photons is larger than the band gap of silicon, the residual energy ($2hf$ - E_{gap}) will heat the sample. The heating rate due to two-photon absorption (TPA) can be written as:

$$Q_{\mathrm{TPA}} = \beta I^2(1 - \frac{N}{N_{\mathrm{at}}})(1 - \frac{E_{\mathrm{gap}}}{2E_{\mathrm{photon}}}). \tag{5}$$

2. Heating due to Auger recombination: At high carrier densities (N > 10^{18} cm^3), the dominant recombination mechanism is due to the three-body interaction, Auger recombination. This effect occurs

when an electron recombines with a hole, and an excess energy of E_{gap} is transferred to another free carrier at the rate of recombination. The generated heat is given as[25]:

$$Q_{\mathrm{rec}} = E_{\mathrm{gap}} \gamma_3 N^3, \tag{6}$$

where γ_3 is the total Auger recombination constant. The recombination constant, γ_3, takes eeh and ehh processes into account, where the excess energy is transferred to electrons or holes, respectively.

3. Free carrier absorption (FCA) based heating: The electrons generated due to interband absorption processes contribute to conduction as free carriers. These free carriers can further absorb photons through intraband absorption (FCA). Thus, interband absorption processes seed electrons for intraband FCA, which can lead to an inhomogeneous and nonlinear absorption profile, resulting in heating. Note that due to Kramers-Kronig relations, FCA should also result in a decrease in the real part of the refractive index, namely the free carrier index (FCI) change.

The changes in FCA and FCI ($\Delta\alpha_{\mathrm{FCA}}$ and Δn_{FCI}) can be estimated by the Drude model, which is also useful for building an intuition for the temperature dependence of these free carrier effects present in our system. $\Delta\alpha_{\mathrm{FCA}}$ and Δn_{FCI} can be evaluated with Drude model as[26]:

$$\Delta\alpha_{\mathrm{FCA}} = \frac{e^3 \lambda^2 \Delta N}{4\pi^2 c^3 \epsilon_0 n} \left(\frac{1}{{m^*_{\mathrm{ce}}}^2 \mu_{\mathrm{e}}} + \frac{1}{{m^*_{\mathrm{ch}}}^2 \mu_{\mathrm{h}}} \right) \tag{7}$$

$$\Delta n_{\mathrm{FCI}} = -\frac{e^2 \lambda^2 \Delta N}{8\pi^2 c^2 \epsilon_0 n} \left(\frac{1}{m^*_{\mathrm{ce}}} + \frac{1}{m^*_{\mathrm{ch}}} \right) \tag{8}$$

It is seen that both free carrier effects are expected to be linear in carrier density. The temperature dependence of these effects is due to the temperature dependence of electron/hole mobilities (μ_{e}, μ_{h}) and effective carrier masses (m^*_{ce}, m^*_{ch}). While carrier mobilities are temperature dependent in Si, the effective masses are determined by the band structure and have a weak temperature dependence. Thus, while $\Delta\alpha_{\mathrm{FCA}}$ is expected to depend on the temperature, Δn_{FCI} is not expected to change drastically with temperature. In fact, from Eq. 8, Δn_{FCI} is expected to change only 15% between 300 K and 1600 K. Therefore, we neglect

the temperature dependence of FCI.

In the literature, $\Delta\alpha_{\mathrm{FCA}}$ and Δn_{FCI} are usually used with their empirical values as[27]:

$$\Delta\alpha_{\mathrm{FCA}} = \Delta\alpha_{\mathrm{h}} + \Delta\alpha_{\mathrm{e}} = 0.51 \times 10^{-20}\lambda^2 TN + 1.01 \times 10^{-20}\lambda^2 TN, \tag{9}$$

$$\Delta n_{\mathrm{FCI}} = -[8.8 \times 10^{-22}N + 8.5 \times 10^{-18}N^{0.8}], \tag{10}$$

where λ is the photon wavelength and T is the equilibrium temperature. The FCA based heating rate is then given as $Q_{\mathrm{FCA}} = \Delta\alpha_{FCA} I$.

Thus, the total heat generation rate is given as, $Q = Q_{\mathrm{TPA}} + Q_{\mathrm{rec}} + Q_{\mathrm{FCA}}$. This equation needs to be solved simultaneously with Eq. 16 to find the temperature profile. We note that, in addition to the nonlinearity of light absorption, we have nonlinearities in the heat generation mechanisms through the carrier density. Next, we demonstrate how these nonlinearities are exploited in feedback mechanisms to create the necessary heat localisation to enable subsurface modification in Si.

2.3 Propagation of light in silicon

The propagation and interaction of counter-propagating beams in silicon can be modelled with approaches that solve Maxwell equations such as Finite Difference Time Domain (FDTD) or Finite Element Method (FEM). However, coupled nonlinear heat- and carrier-generation mechanisms in silicon and the feedback between the two counter-propagating laser beams render the problem impractical for these methods. Further, since typical Si wafer thickness is three orders of magnitude higher than the laser wavelength, FDTD/FEM techniques will require extensive storage and computing power. In order to probe further into the pulse propagation dynamics, we will use the nonlinear paraxial equation (NPE), which captures the essential dynamics of the light propagation and evolution. The NPE is given by[28]:

$$\frac{\partial A}{\partial z} = \frac{i}{2k}\nabla_T^2 A + \frac{ikA}{n_0}(\Delta n_{\mathrm{total}} + i\Delta k_{\mathrm{total}}), \tag{11}$$

where n_0 is the wavelength dependent refractive index that can be found from Sellmeier relation, A is the

electric field distribution, k is the wave vector in Si, and Δn_{total} and Δk_{total} are the total change in real and imaginary parts of the refractive index, respectively. The total change in the real part of the refractive index of Si is given as:

$$\Delta n_{\text{total}} = \Delta n_{\text{Kerr}} + \Delta n_{\text{FCI}} + \Delta n_{\text{Thermal}}, \tag{12}$$

where Δn_{Kerr} is due to Kerr nonlinearity, Δn_{FCI} is due to free-carrier induced changes and $\Delta n_{\text{Thermal}}$ is due to temperature induced refractive index change. These components are given as:

1. Δn_{Kerr}: This term is due to instantaneous Kerr response of Si and can be written as $\Delta n_{\text{Kerr}} = n_2 I$. The Kerr coefficient for Si is given as[11], $n_2 = 5 \times 10^{-14}$ cm^2/W at 1.5 μm.

2. Δn_{FCI}: This term represents the refractive index change due to free electron-hole pairs. The refractive index change per cm^{-3} change in electron-hole pairs is given as $\Delta n_{\text{FCI}} = -[8.8 \times 10^{-22} N + 8.5 \times 10^{-18} N^{0.8}]$ from Eq. 10.

3. $\Delta n_{\text{Thermal}}$: This term is due to the refractive index change of Si with temperature. The relation between the temperature change and the refractive index change is given as[29,30]: $\Delta n_{\text{Thermal}} = 1.86 \times 10^{-4} \Delta T$.

The total change in the imaginary part of the refractive index of silicon can be decomposed as follows:

$$\Delta k_{\text{total}} = \Delta k_{\text{TPA}} + \Delta k_{\text{FCA}}, \tag{13}$$

where the terms are given as:

1. Δk_{TPA}: This term represents the losses due to two-photon absorption. For two-photon absorption, the attenuation coefficient (cm^{-1}) is given as: $\alpha_{\text{TPA}} = \beta I$. The change in the imaginary part of refractive index can be calculated from the absorption coefficient as follows: $\Delta k_{\text{TPA}} = \frac{\alpha_{\text{TPA}} \lambda_0}{4\pi} = \frac{\beta I \lambda_0}{4\pi}$.

2. Δk_{FCA}: The generated electron-hole pairs increase the absorption through free-carrier absorption. The relation between the absorption coefficient (cm^{-1}) and the free electron and hole carrier densities is given as $\Delta \alpha_{\text{FCA}} = \Delta \alpha_{\text{h}} + \Delta \alpha_{\text{e}} = 0.51 \times 10^{-20} \lambda^2 T N + 1.01 \times 10^{-20} \lambda^2 T N$. Note that for high carrier densities, free electrons and holes have the same densities. The change in the imaginary part of refractive index can be calculated from the absorption coefficient as follows: $\Delta k_{\text{FCA}} = \alpha_{\text{FCA}} \lambda_0 / 4\pi$.

2.4 Light propagation under nonlinear feedback conditions

Thermal lensing due to $\Delta n_{\text{thermal}}$ and FCI diffraction due to Δn_{FCI} are the main diffractive effects responsible for the formation of subsurface structures in Si. For a single laser beam at low intensities or weak focusing conditions, thermal lensing is not substantial, FCI diffraction precludes self-focusing, and thus no subsurface modification is observed. However, the competition between these two opposing diffractive effects can be won by thermal lensing when two counter-propagating laser beams nonlinearly couple involving a self-focusing feedback process. In the experiment, the laser beam reflects from the Si-air interface, to produce the counter-propagating beams that are nonlinearly coupled, the beam self-focuses and subsurface modification is observed. Moreover, the process restarts with every laser pulse, but with a difference, *i.e.*, modified areas by the previous pulse reconfigures the next pulse's propagation. In close analogy to the moving-focus model in filamentation[31], this results in high-aspect ratio structures, elongating with every pulse. Here, we present a model for subsurface modification with a single laser pulse under nonlinear feedback conditions, mainly due to $\Delta n_{\text{thermal}}$ and Δn_{FCI}.

Single pulse model implementation

The strong interaction between intensity, carrier and temperature distributions (Eqs. 3, 16, 9, 15) and also the nonlocal feedback between counter-propagating beams render the system highly nonlinear. Thus, the temperature and carrier distributions strongly depend on the previous history of the system, and a method that is capable of handling both high optical nonlinearities and nonlocal feedback is necessary. For this purpose, we used the split step Fourier method, which is commonly used to solve nonlinear differential equations numerically[32]. The method propagates linear and nonlinear terms separately, in small steps. In each step, the linear term is solved exactly in the frequency domain, whereas the nonlinear evolution is evaluated in the spatial domain[32].

A propagating laser pulse will excite free carriers in Si to produce a nonuniform refractive index profile due to FCI. In parallel, heating through interband- (TPA) and intraband absorption (FCA) will modify the refractive index. These two opposing diffractive effects will continually reconfigure the beam propagation, complicating the numerical implementation. However, a simplification arises, if one notes that the pulse

duration is much larger than the round trip time of light in the wafer. For instance, the round trip time in a 500 μm thick Si wafer is $\approx$ 10 ps, and during this time the refractive index, temperature and carrier distributions distribution of a nanosecond pulse can be considered quasi-stationary. Thus, a nanosecond pulse can be propagated with Eq. 15, as a consecutive set of discrete 10-ps temporal slices (Fig. S2).

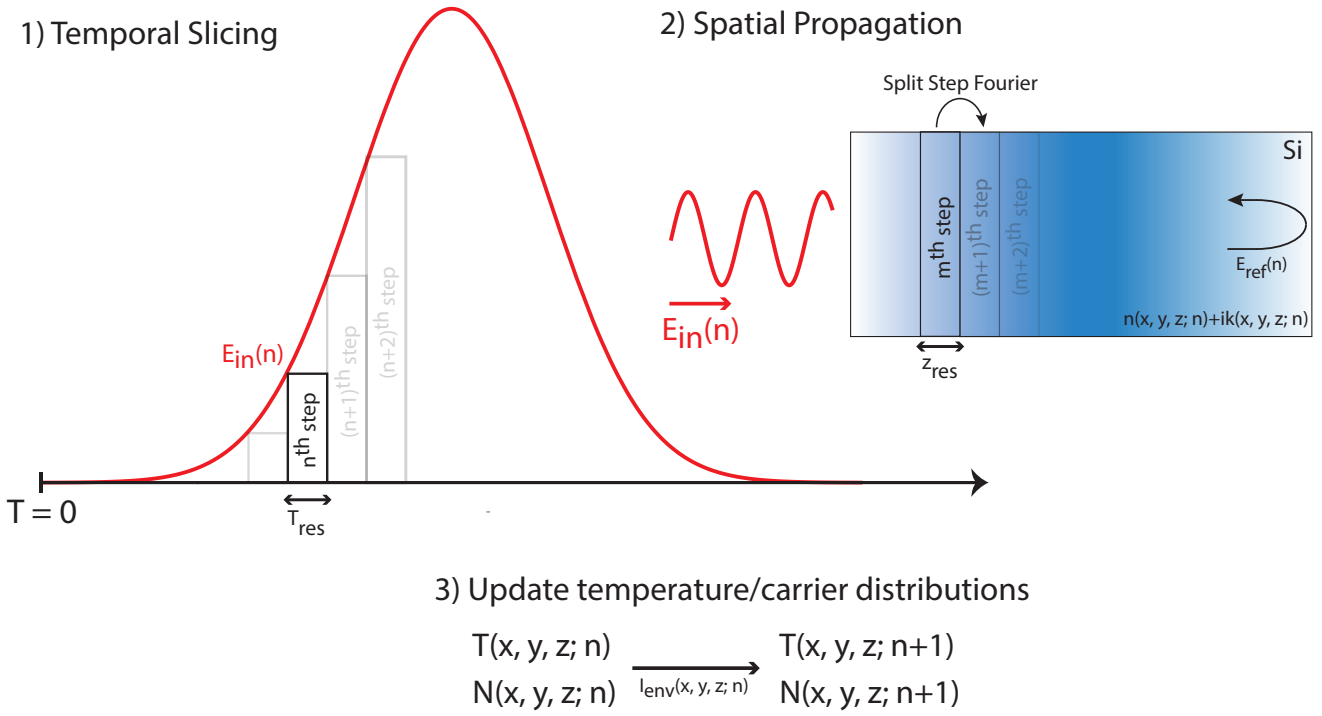

Figure S2: Conceptual diagram for the solution of pulse propagation in silicon. For each temporal slice, the intensity distribution (I_{env}) is obtained using split step Fourier method using the refractive index, temperature and carrier distributions from the previous iteration. These distributions are updated at the end of the propagation for each temporal slice.

For each temporal slice, the intensity distribution (I_{env}) in Si is obtained using the split step Fourier method (Eq. 15, Fig. S2), with refractive index ($n_{\text{total}}, k_{\text{total}}$), temperature ($T$) and free carrier density (N) values used as input from the previous iteration. First, the field is linearly propagated by half of the spatial resolution ($\Delta z/2$), then the nonlinear term is propagated for a length of Δz, and the calculation is completed by propagating the field linearly for another $\Delta z/2$. Thus, the total linear and nonlinear propagation length is equal to one spatial step size, Δz. In this implementation of the split-step method, the error is third order in Δz[32]. At the end of the iteration, $I_{\text{env}}(\mathbf{r})$ is used to update temperature (T), carrier density (N), and

refractive index ($n_{\text{total}}, k_{\text{total}}$) distributions through Equations 3,16,12,13 for the next temporal slice. Note that, in the case of two counter-propagating pulses, an interference pattern will form, with an oscillatory period of $\lambda_{\text{Si}}/2 = 225$ nm. The length scale of these fluctuations is much smaller than both the beam spot size, and the carrier/heat diffusion lengths, and the effect smears out. Therefore, the intensity envelope (I_{env}) is used in the simulations:

$$I \, \alpha \, |E_{\text{forward}} + E_{\text{backward}}|^2 \leq (|E_{\text{forward}}| + |E_{\text{backward}}|)^2 = I_{\text{env}}. \tag{14}$$

Single-pulse simulation results

In the experiments, when a low power beam is weakly focused (with no feedback from a second beam) subsurface modification is not observed. However, if the same beam is accompanied by a counter-propagating, weakly focused *dressing* beam, subsurface modification is observed. In order to shed light on the role of the second beam (dressing beam) in creating subsurface structures, two scenarios were simulated: In the first case, a single pulse is weakly focused in Si, without a dressing beam. In the second case, a counter-propagating dressing beam is added to monitor the nonlinear coupling between the beams.

In the experiments corresponding to the second scenario, the laser is focused after the wafer, such that the reflected portion due to Fresnel reflection at the air-Si interface ($R = 0.3$) interacts with the forward-propagating component in Si. The forward propagating dressing beam carries more energy (14 μJ) in comparison to the reflected beam (4.2 μJ), however since the dressing beam is focused outside the sample, it has has a lower peak intensity in comparison to the reflected beam in Si.

The simulation results are given in Fig. S3. In the first scenario, the increase in intensity and temperature remains relatively limited (Figs. S3a, S3c). This corresponds to the case of $\alpha = \gamma = 0$ in the Toy Model (see Supplementary section 4). The beam first diffracts due to FCI in the first few ns, and after a short temporal lag, focusing starts due to thermal nonlinearity (Figs. S3c). Thermal lensing lags behind, since the main heating mechanism is FCA, which first requires seed electrons from TPA. This effect demonstrates the competition between diffraction due to FCI and thermal lensing, which present

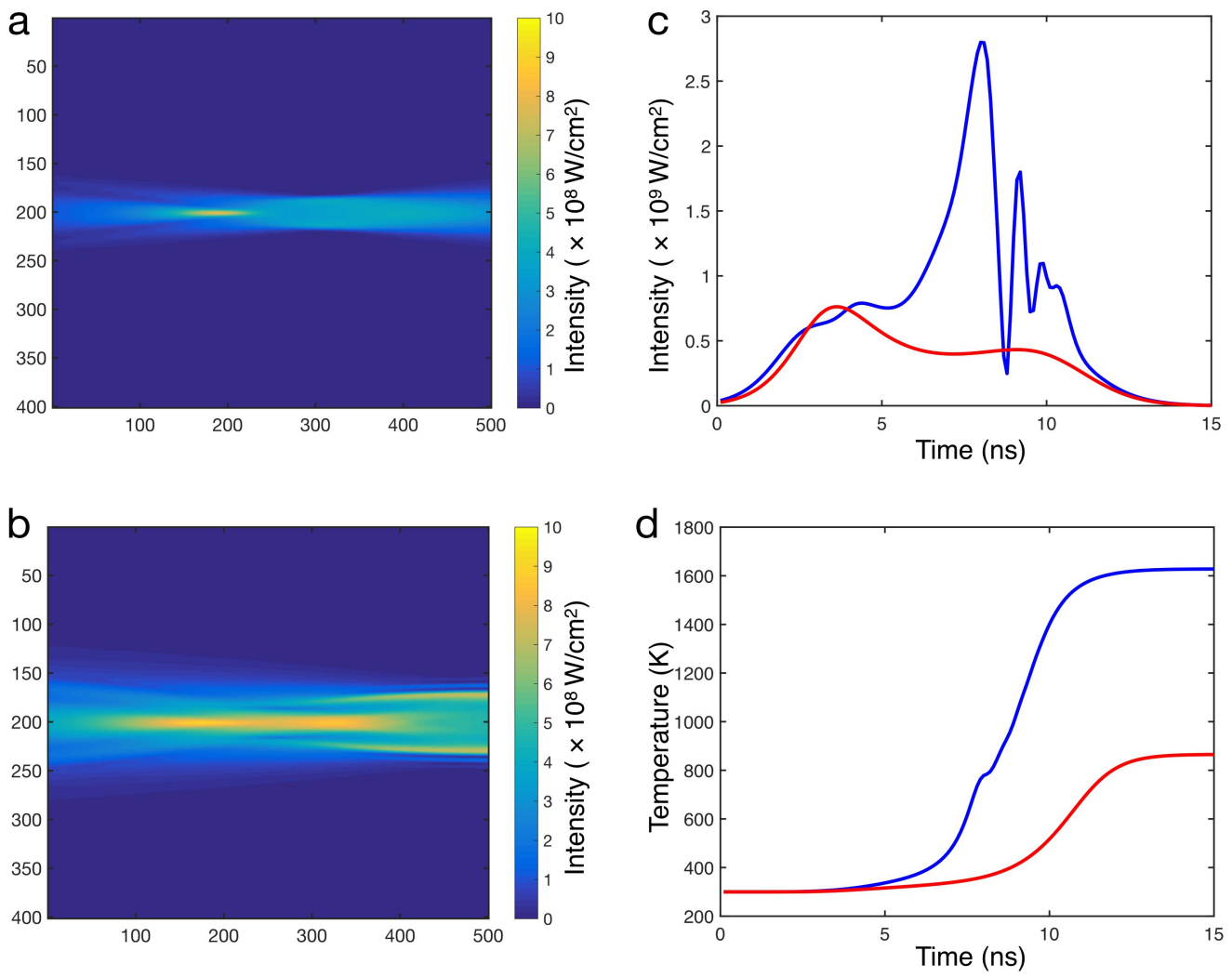

Figure S3: Simulations for single pulse propagation in silicon. **a**, Intensity distribution for the undressed beam case. **b**, Intensity distribution of the dressed beam case. **c**, Intensity evolution on a location where modification occurs. The blue curve is for the dressed beam case, and the red curve for the undressed case. The double peak is due to the lag of thermal lensing, which presents its effect a few ns later than FCI. This is an indication of the competition between the two diffractive effects. **d**, Comparison of temperature evolution on the location where modification occurs. The blue curve is for the dressed beam, and the red curve is for the undressed beam. The pulse used in the simulation is temporally and spatially Gaussian, with a temporal width of 5 ns and a spatial width of 3 μm.

their influence at different timescales (Fig. S3c.)

In the second scenario (Fig. S3b), the dressing beam is turned on, with the rest of the simulation conditions same as in the first scenario. In this case, the two beams nonlinearly couple, and the intensity

and temperature reach to higher values in comparison to the first scenario (Figs. S3b, S3d). This situation corresponds to the condition of $\alpha > 0,\ \gamma > 0$ in the Toy Model. In Fig. 2b of the Manuscript, the total self-induced refractive index for the single beam case and dressed beam case are shown. For the latter, it is seen that thermal lensing is stronger than diffraction due to FCI effects. Such high intensity and temperature values are characteristic for self-focusing and beam collapse[33], through which the optical wave produces material changes in Si (see Supplementary Information section 3).

This behaviour is possible due to self-focusing feedback and realised by the cooperation of two coupled beams. The temperature gradient generated by the reflected beam behaves like a lens to cause a redistribution on the dressing beam. The reconfigured dressed beam in turn creates a different environment for the focused beam to propagate. This continuous nonlocal, nonlinear feedback between the beams help initiate the subsurface modification, which then leads to rod elongation with consecutive pulses.

3. Material Characterisation

We modify the bulk of silicon with infrared laser pulses. The wafer is studied with infrared (IR) transmission microscopy, and a representative cross-sectional image is shown in Fig. S4. It is seen that each pulse can be used to create a voxel, with dimensions of $\approx 1\mu\text{m} \times 1\mu\text{m} \times 1\mu\text{m}$. In order to characterize the crystal structure of modified volume, it is more practical to modify a large volume of 0.2 cm^3 which is processed with the multiplexed approach described in the manuscript. The laser-processed volume inside Si is analysed using transmission electron microscopy (TEM), which is exposed first by diamond pen scribing then by simply fracturing the Si chip. We intentionally avoid exposing the processed area using chemical or mechanical pretreatment in order to prevent further modification of the microstructure by these techniques.

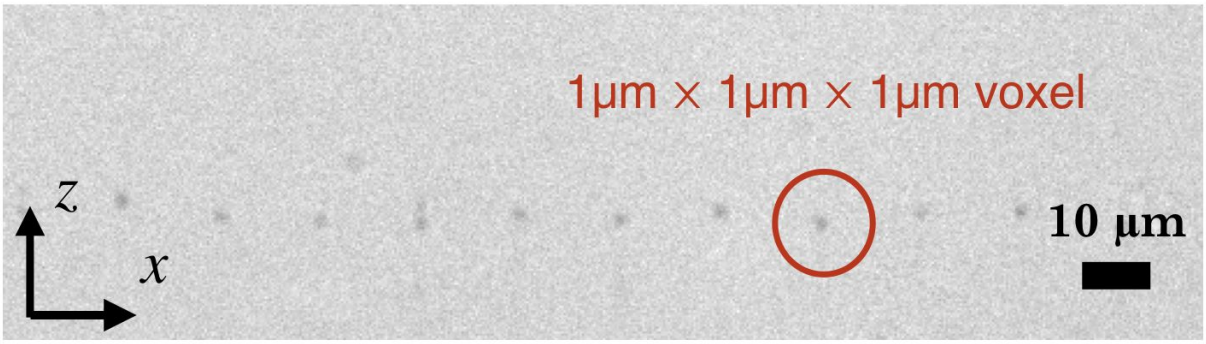

Figure S4: IR microscope image from single-pulse modified volume. The x-z cross-section is imaged with IR microscopy. Voxels of $\approx$ 1-μm feature size are created with $\approx$ 8μJ laser pulses. The laser propagates along the z axis as in the manuscript concention.

A protective carbon layer is deposited to the subsurface processed volume before the TEM lamella preparation using first electron beam assisted then Ga focused ion beam (FIB) assisted precursor decomposition. Then, a TEM lamella of processed area with 2 μm height and 6 μm width is prepared by the *in-situ* lift-out technique using 30 keV Ga FIB with adapted currents in a Zeiss Crossbeam NVision 40 system. Transfer of the lamella to a 3-post copper lift-out grid (Omniprobe) is performed using a Kleindiek micromanipulator. Low-energy Ga ions (5 keV) are used in the final thinning stage for minimal sidewall damage to the lamella. Finally, the specimen is mounted in a double-tilt analytical holder and placed in a Model 1020 Plasma Cleaner (Fischione) for 10 s in order to remove any organic contamination. TEM investigations are performed using an image C_s-corrected Titan 80-300 microscope (FEI) operated at an

accelerating voltage of 300 kV.

A bright-field scanning TEM image of the prepared lamella is shown in Fig. S5, where defect-free, single-crystalline Si region is observed along with multiple defective regions. On the left-hand-side of the image, strong bend contours[34] originating from the defective regions can be identified from varying diffraction contrasts.

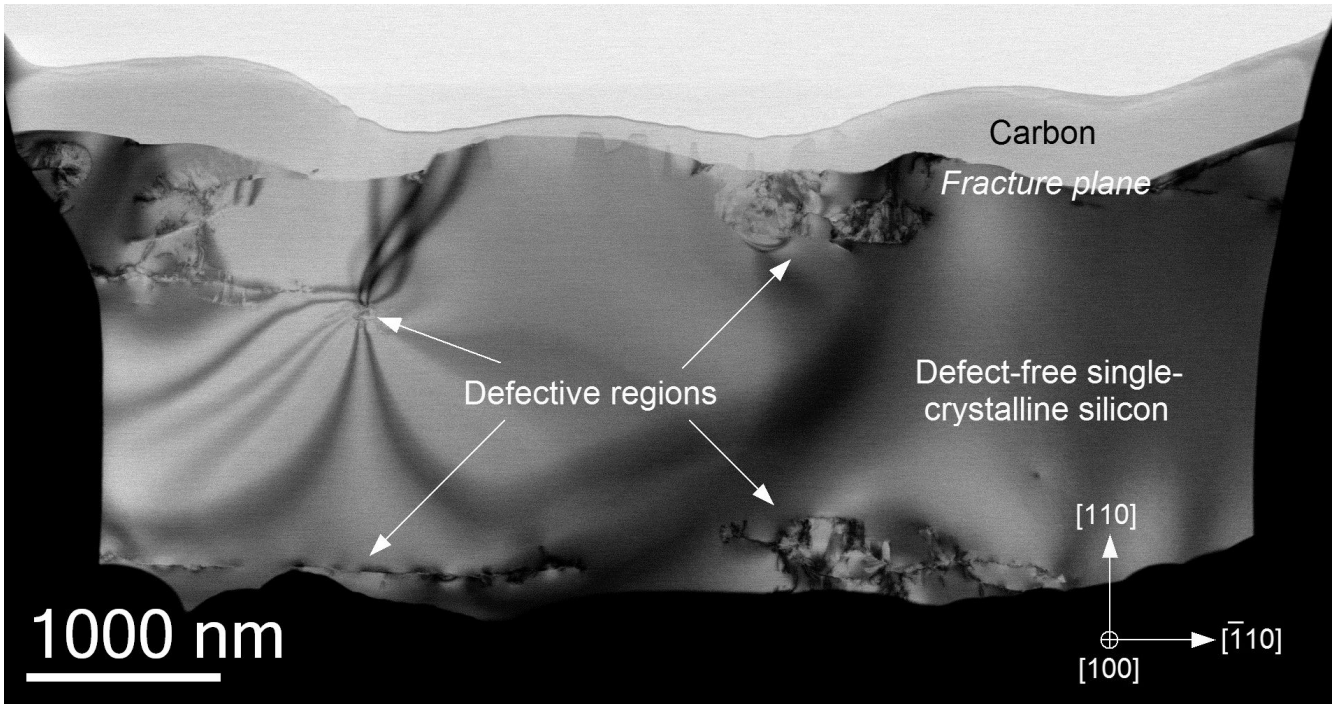

Figure S5: TEM image from the *x-y* plane of the lamella. Quasi bright-field scanning TEM image of the laser-processed area (6 μm in width, 2 μm in height) recorded with a high-angle annular dark-field detector and an adjusted camera length of 3 m showing local defective regions with strong bending contours.

In order to better understand the structure of the defective regions, a magnified TEM image (Fig. S6) along with selected area electron diffraction (SAED) pattern (inset) is taken. The broad rings seen in the SAED pattern strongly suggest the presence of an amorphous Si (a-Si) formation as a result of laser-material interaction. However, the a-Si formation is local, confined to a few tens of nanometers, which is surrounded by single-crystalline Si, evidenced from the dark spots (diamond cubic structure, space group Fd$\bar{3}$m). It should be noted that this pattern is slightly deviating from that of the original Si chip as a result of the lattice bending due to the defective regions.

The diffraction pattern analyses are repeated for another defective region (Fig. S4). This time single-

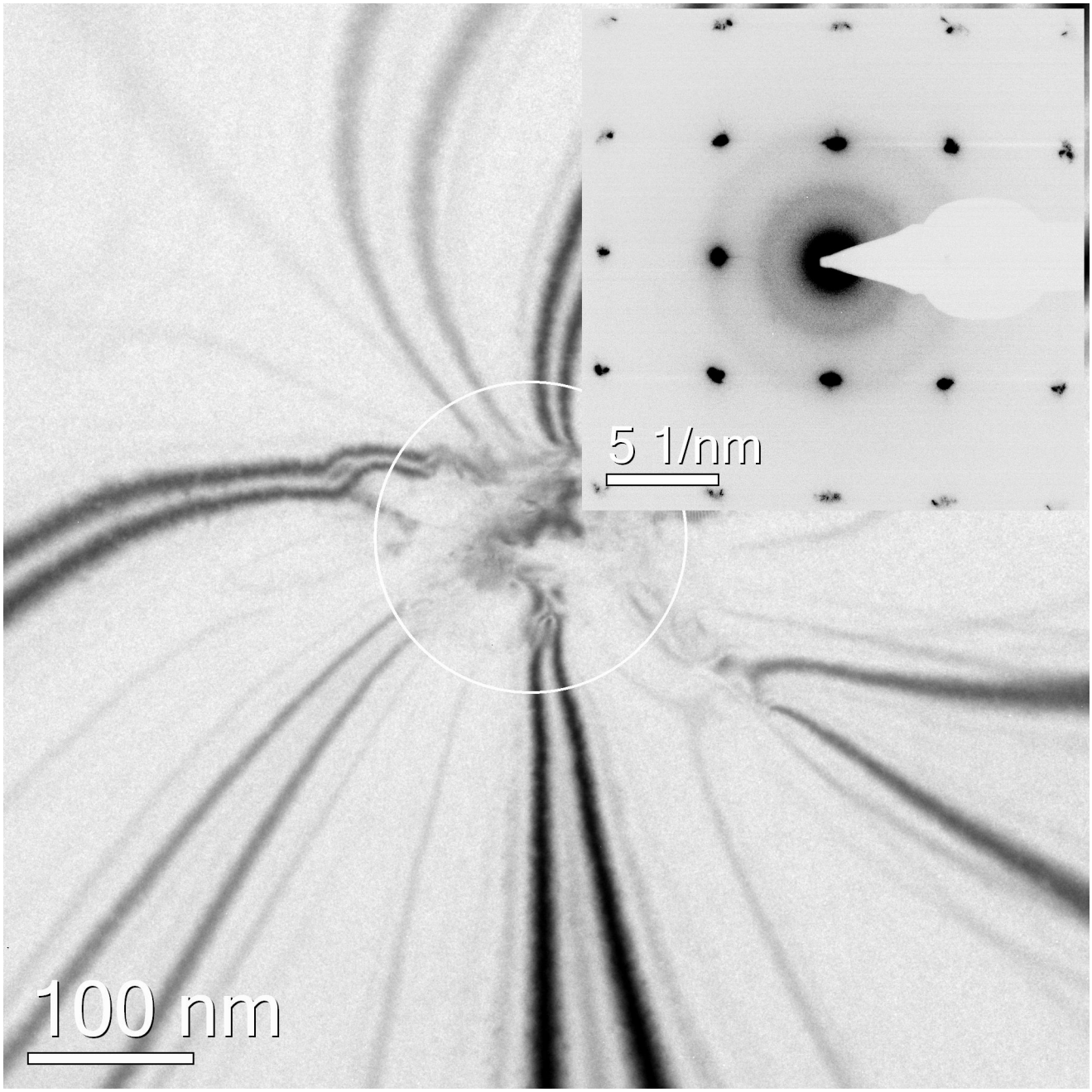

Figure S6: A magnified bright-field TEM image of the defective area and the SAED pattern (inset) taken from the area marked with a white circle.

crystalline Si pattern is accompanied by a few additional Bragg reflections, possibly originating from high-pressure Si structures such as Si-III (body-centered cubic structure with 8 atoms per unit cell, space group Ia3) and/or Si-XII (rhombohedral structure with 8 atoms per unit cell, space group R3)[35].

These observations are not entirely unexpected since laser pulses can deposit localized high-pressures

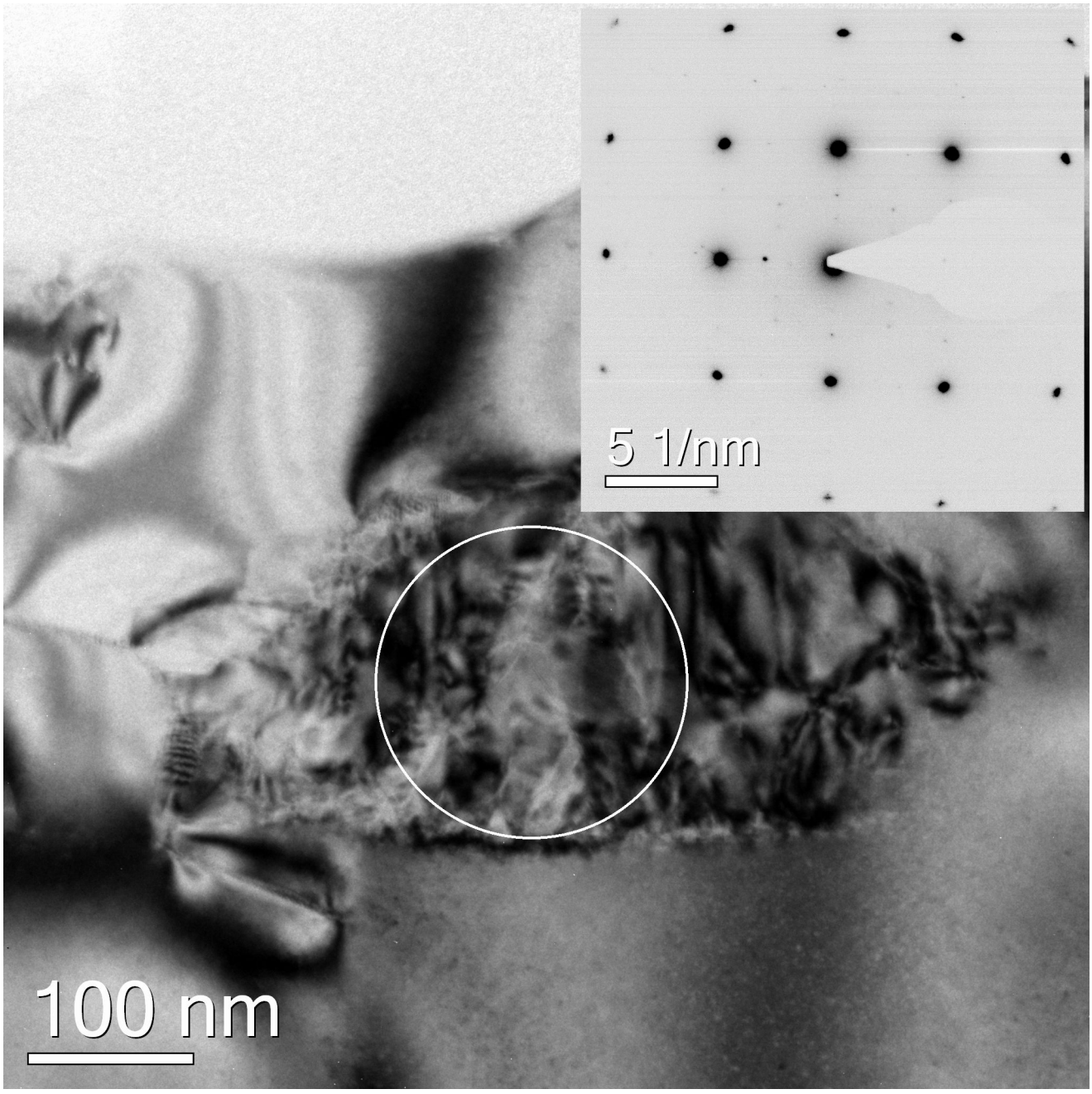

Figure S7: A magnified bright-field TEM image of the defective area and the SAED pattern (inset) taken from the area marked with a white circle.

into the bulk and this could cause formation of Si-III and/or Si-XII phases. The deposited, highly-localized energy can also melt a localized area; and upon rapid cooling this melt can be solidified in the form of a-Si. However, most of the laser-modified volume remains as single-crystal.

4. Description of the Toy Model

Truly 3D silicon (Si) based devices employing embedded or in-chip functionalities, created without altering wafer surfaces have not been realised to date. A promising direction to overcome current limitations in fabricating controlled structures in Si is to study the rich nonlinear dynamics of interacting optical beams, with implications in multiple fields. For instance, thermally coupled wavepackets have been shown to emulate nonlinear gravitational effects[28]. In an analogous way, we exploit the nonlinear dynamics of thermally coupled laser beams to enable the creation of subsurface functional structures in Si.

In order to explain the formation of structures and their elongation in Si, we developed a toy model. The model is a simplistic one, with many details removed so it can be used for qualitative understanding. It captures the salient features in the formation of structures, and makes predictions that are in good agreement with experiments. Specifically it has three claims, (i) structures do not form when there is no feedback from a second, counter-propagating beam, (ii) the structures form and get elongated by each pulse, along the optical axis, when there is a counter-propagating laser beam which initiates nonlinear feedback, (iii) their elongation saturates for high pulse numbers, self-regulating and stopping before any surface damage occurs.

We first focus on the single laser pulse case, then on the two pulse case, and finally extend the model to any number of pulses, quite similar to inference with mathematical induction method. Pulse propagation is essentially governed by the nonlinear paraxial equation (NPE) and heat equation. The NPE is given as,

$$\frac{\partial A}{\partial z} = \frac{i}{2k}\nabla_T^2 A + \frac{ikA}{n_0}\Delta n, \tag{15}$$

where n_0 is the refractive index, A is the electric field distribution, k is the wave vector in Si, and Δn is the self-induced refractive index. The last term is responsible for the feedback of self-induced opposing diffractive effects due to thermal nonlinearity and free carrrier induced (FCI) refractive index change. The temperature evolves according to the heat equation,

$$\rho C_{\mathrm{p}}(T)\frac{\mathrm{d}T}{\mathrm{d}t} - \nabla.(\kappa\nabla T) = Q, \tag{16}$$

where ρ is Si density, $C_p(T)$ is the specific heat, κ is the thermal conductivity, T is lattice temperature and Q is the heat generation rate term.

This equation system is mathematically equivalent to the Newton-Schrödinger system[28], inherently nonlinear and describes a wave providing feedback back onto itself. During beam propagation, a pulse induces two opposite refractive index changes, due to thermal nonlinearity and Free Carrier Induced (FCI) effects. These two self-induced diffractive effects effectively operate as concatenated converging and diverging lenses. If thermal lensing is stronger than diffraction due to FCI effects, the optical wave self-focuses in Si, eventually collapses and produces material changes. This constitutes the first feedback mechanism, which is responsible for the formation of the building blocks of more complex structures. We later relate the competition between these diffractive effects to the nonlinear coupling between two counter-propagating beams.

When additional pulses are incident, the second feedback mechanism is activated: Each pulse locally modifies silicon; this modification (with a permanent refractive index change), in turn, shifts the focal point for the next pulse similar to the moving focus model of self-focusing. We assume that the self-induced lensing effects during pulse propagation can be described with two concatenated thin lenses, one with a positive focal length, f_{therm}, and another with a negative focal length, f_{FCI}. Once the first pulse focuses and creates a local modified volume, due to different optical properties in this volume, the focal position of the next pulse is displaced with respect to the previous pulse. This focal translation is found from the lens equation, $\frac{1}{l_2} = \frac{1}{f_t} + \frac{1}{l_1}$, where l_1 is the focal position of the first pulse measured from the lens located at $l_1/2$ before the exits surface, l_2 is the focal position of the second pulse, and f_t is the total induced focal length per pulse, given by $f_t = \frac{f_{\text{therm}}}{(1+\eta)}$. Here, $\eta = \frac{f_{\text{therm}}}{f_{\text{FCI}}}$ is a measure of the competition between the self-induced lenses. For thermal nonlinearity to overcome FCI effects and start subsurface modification, we have $-1 < \eta < 0$, otherwise $\eta < -1$.

The position of the n^{th} modification (n>1) located at l_n can be obtained with simple reiteration from,

$\frac{1}{l_n} = \frac{n-1}{f_t} + \frac{1}{l_1}$. After the n^{th} pulse, the total length of subsurface structure, $\delta l_n = l_1 - l_n$ is given by,

$$\delta l_n = \frac{l_1^2(n-1)}{f_t + l_1(n-1)}. \tag{17}$$

We will show that f_t can be written as a combination of two thin lenses, with focal lengths of f_{therm} and f_{FCI}, and then calculate the structure length for a given number of laser pulses from the preceding equation. We will then compare this model's prediction with experiments corresponding to different pulse numbers. We start by forming the forward and backward propagating Gaussian beams with,

$$I_{f(b)}(z,r) = I^0_{f(b)} \frac{w_0^2}{w^2_{f(b)}} e^{\frac{-2r^2}{w^2_{f(b)}}}, \tag{18}$$

where subscripts are for forward (f) and backward (b) propagating beams, w_0 is beam waist, r is radial distance, and the beams propagate along the z axis. The beam radii $w_{f(b)}(z)$ for the forward and backward propagating beams are given as, $w_{f(b)}(z) = w_0\sqrt{1 + \left(\frac{z-L\mp l}{z_R}\right)^2}$, where L is the wafer thickness, z_R is Rayleigh length, and l is the distance of focal point relative to the laser exit surface.

In order to calculate the self-induced refractive index changes due to these beams, we need to estimate the carrier density and temperature profiles. The former is needed for estimating FCI diffraction, while the latter is required to find thermal focusing. The total change in carrier density, δN_{tot}, after the first pulse is assumed to be,

$$\delta N_{tot} = \delta N_1 + \delta N_2 + \delta N_3, \tag{19}$$

where δN_1, δN_2, and δN_3 are the changes in carrier density due to forward propagating beam, backward propagating beam, and their coupling, respectively. For notational simplicity, we will use the index $i = 1, 2$ for the effects due to forward and backward propagating beams, and $i = 3$ for their coupling in the remainder of this exposition, and include the effects of beam coupling semi-phenomenologically. The change in carrier density due to the beams defined by Eq. 18 becomes,

$$\delta N_i(z,r) \approx \frac{\beta I_i^2(z)\delta t}{2E} e^{\frac{-4r^2}{w_i^2(z)}}, \tag{20}$$

where δt is pulse width, E is photon energy, β is two-photon-absorption coefficient and I denotes intensity. Thus, the total refractive index profile due to induced carriers is given by,

$$\delta n_{\mathrm{FCI}}(z,r) \approx -A\delta N_{\mathrm{tot}}(z,r) = -\sum_{i=1}^{3} \frac{A\beta I_{\mathrm{i}}^2(z)\delta t}{2E} e^{\frac{-4r^2}{w_{\mathrm{i}}^2(z)}}, \tag{21}$$

where $A = 8.8 \times 10^{-22}$ cm^3 is constant[36].

In parallel, the refractive index changes due to heating, in proportion to the thermo-optic coefficient of Si. If the same spatial profile for temperature and intensity is assumed, the temperature change induced by the pulse is given by,

$$\delta T_{\mathrm{i}}(z,r) \approx \frac{\beta I_{\mathrm{i}}^2(z)\delta t}{\rho c} e^{\frac{-4r^2}{w_{\mathrm{i}}^2(z)}}, \tag{22}$$

where ρ is density and c is specific heat capacity. Thus, the refractive index change due to heating is,

$$\delta n_{\mathrm{therm}}(z,r) \approx \sum_{i=1}^{3} \frac{\beta I_{\mathrm{i}}^2(z)\delta t}{\rho c} \frac{\mathrm{d}n}{\mathrm{d}T} e^{\frac{-4r^2}{w_{\mathrm{i}}^2(z)}}. \tag{23}$$

We now resort to paraxial ray approximation, expand the terms around the optical axis and keep the first two terms for $\delta n_{\mathrm{FCI}}(z,r)$ and $\delta n_{\mathrm{therm}}(z,r)$ which provides,

$$\delta n_{\mathrm{FCI}}(z,r) \approx -\sum_{i=1}^{3} g_{\mathrm{i}}(z) \left(1 - \frac{4r^2}{w_{\mathrm{i}}(z)^2}\right) \tag{24}$$

and

$$\delta n_{\mathrm{therm}}(z,r) \approx \sum_{i=1}^{3} h_{\mathrm{i}}(z) \left(1 - \frac{4r^2}{w_{\mathrm{i}}(z)^2}\right), \tag{25}$$

where $g_{\mathrm{i}}(z) = A\frac{\beta I_{\mathrm{i}}^2(z)\delta t}{2E}$ and $h_{\mathrm{i}}(z) = \frac{\beta I_{\mathrm{i}}^2(z)\delta t}{\rho c}\frac{\mathrm{d}n}{\mathrm{d}T}$ give the changes in refractive indices at $r = 0$. To simplify our model, we average over the propagation direction z, and the general form of these refractive index profiles becomes,

$$\delta\bar{n}(r) = \delta\bar{n}_0(1 - \frac{4r^2}{\bar{w}^2}), \tag{26}$$

with $\bar{n}$, $\bar{w}$ indicating average values over z axis. We note that this is the familiar form of refractive index profile in graded index lenses.

Next, we introduce matrix optics formalism, which we will use to characterise ray paths in inhomogeneous media. Assume that a paraxial beam is propagating in the z direction, in a medium with an r-dependent refractive index profile. If after propagating $\mathrm{d}z$, the incidence angle is θ at an interface, then from Snell's law we have,

$$n(r)\cos(\theta) = n(r+\mathrm{d}r)\cos(\theta+\mathrm{d}\theta) \tag{27}$$

If we use Taylor expansion, by keeping the first two terms for $\cos(\theta+\mathrm{d}\theta)$ we have,

$$n(r)\cos(\theta) = \left[n(r) + \frac{\partial n}{\partial r}\mathrm{d}r\right][\cos(\theta)\cos(\mathrm{d}\theta) - \sin(\theta)\sin(\mathrm{d}\theta)]. \tag{28}$$

We divide by $\cos(\theta)$, and by using $\tan(\theta) = \frac{\mathrm{d}r}{\mathrm{d}z}$, the paraxial ray equation is found:

$$\frac{\partial n}{\partial r} = n(r)\frac{\mathrm{d}^2 r}{\mathrm{d}z^2}. \tag{29}$$

Finally, by substituting Eq. 26 into Eq. 29, we find the propagation equation for our case,

$$\frac{\mathrm{d}^2 r}{\mathrm{d}z^2} + z_0^2 r = 0, \tag{30}$$

where $z_0^2 = \frac{8}{\bar{w}^2}$. This equation provides the transmission matrix of the medium[37]:

$$T = \begin{bmatrix} \cos(zz_0) & \frac{1}{z_0}\sin(zz_0) \\ -z_0\sin(zz_0) & \cos(zz_0) \end{bmatrix} \tag{31}$$

It can be shown[37,38] that the T matrix is equivalent to a lens with a focal length of f :

$$f = \frac{1}{\delta\bar{n}_0 z_0}. \tag{32}$$

Here, we assume the change of refractive index for forward and backward is the same and introduce two coefficients, α and γ, which represent the strength of coupling terms in Eqs. 24 and 25, as $\bar{g}_3 = \alpha\bar{g}_1 = \alpha\bar{g}_2$ and $\bar{h}_3 = \gamma\bar{h}_1 = \gamma\bar{h}_2$. Thus, the induced focal lengths due to FCI and thermal effects can be calculated as,

$$f_{\mathrm{FCI}} \approx -\frac{1}{\delta\bar{n}_{0,\mathrm{FCI}} z_0(2+\alpha)} = \frac{2E}{A\beta I^2 \delta t z_0(2+\alpha)}, \tag{33}$$

$$f_{\mathrm{therm}} \approx \frac{1}{\delta\bar{n}_{0,\mathrm{therm}} z_0(2+\gamma)} = \frac{\rho c}{\beta I^2 \delta t \frac{dn}{dT} z_0(2+\gamma)}. \tag{34}$$

4.1 Claim 1

The model suggests that the subsurface structures do not form when there is no feedback from the second beam (Claim 1). To see this, we remove the coupling simply by turning one of the beams off, and use $\alpha = \gamma = 0$. We plug in typical experimental parameters in Eqs. 33 and 34, e.g, $\lambda = 1.55$ μm, $\bar{w} = 30$ μm, $\delta t = 5$ ns, and $E_p = 10$ μJ, and the self-induced refractive indices and corresponding focal lengths are found as, $\delta\bar{n}_{0,\mathrm{FCI}} = -8.3\times10^{-5}$ and $\delta\bar{n}_{0,\mathrm{therm}} = 1.4\times10^{-6}$, and $f_{\mathrm{therm}} = 7.4\times10^{6}\mu$m and $f_{FCI} = -2.5\times10^{5}\mu$m. The net effect is a diverging lens ($|f_{FCI}| < f_{\mathrm{therm}}$ and $\eta < -1$), which can preclude subsurface modification.

We test this claim experimentally by coating half of the surface of a double-side polished Si wafer with a thin anti-reflection coating layer (Fig. S8a) and evaluate the bulk of the wafer after subsurface writing procedure. As expected, the part of the wafer where there is no anti-reflection coating is observed to carry subsurface structures (Fig. S8b). In contrast, we did not observe any discernable change in the bulk of Si located directly below the coated areas (Fig. S8b).

4.2 Claims 2 and 3

The model also predicts that structures form and elongate, pulse to pulse, along the optical axis (Claim 2), and also self-regulate with increasing pulse numbers, their elongation saturating and stopping before any surface damage (Claim 3). In the model, for the case of two counter-propagating beams, the nonlinearity due to coupling is turned on, *i.e.*, nonzero α and γ, and we can have $-1 < \eta < 0$ with the net effect of a converging lens. This is demonstrated experimentally and is shown in Fig. S9, where structure lengths for different number of pulses are compared with the toy model prediction from Eq. 17. In order to fit the data, we used $f_{\mathrm{t}} = 26$ mm and assume for simplicity the same intensity and beam size for both forward

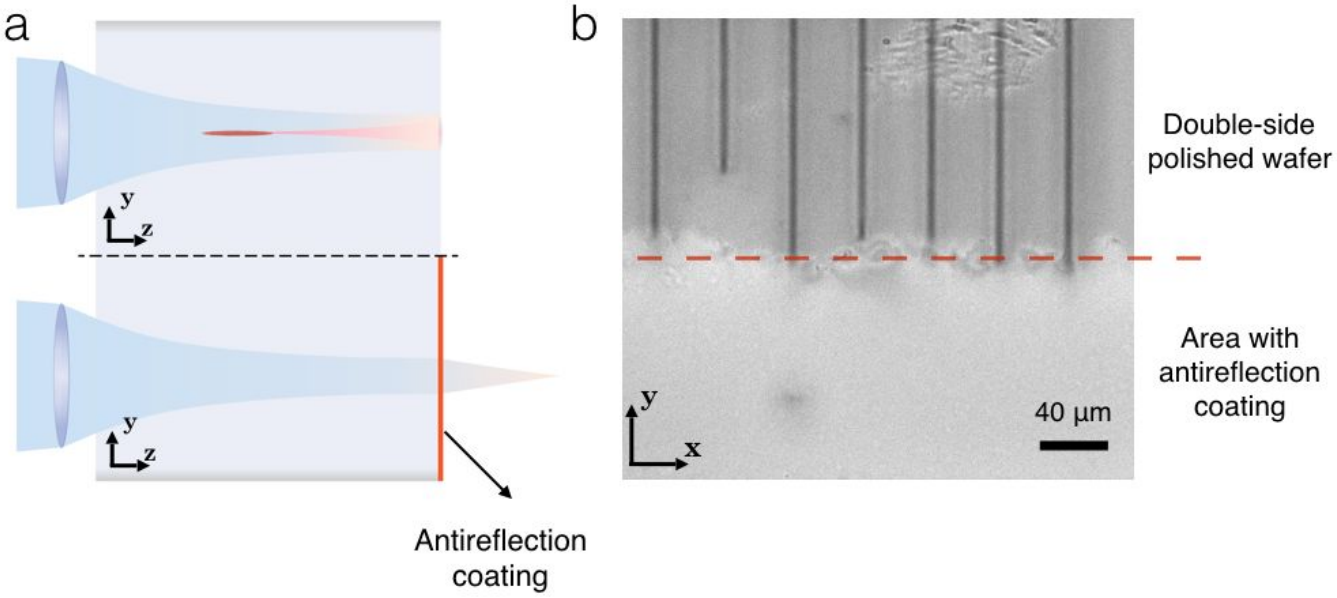

Figure S8: Experimental test of the first claim of the toy model. a, Schematic showing the experiment for probing the role of coupling between two counter-propagating beams. The wafer is coated with a thin anti-reflective coat (200 nm thick Si_3N_4) on one half, and it is left untouched for the other. **b**, Infrared transmission microscope image showing both halves of the processed wafer. The subsurface structures immediately stop when the laser moves into the areas with anti-reflection coating. The laser propagates along the z axis and is set to operate at 150 kHz with 14 μJ pulses. The sample is scanned at 0.2 mm/s along y axis relative to the laser.

and backward beams. The corresponding values of α and γ can be estimated from,

$$\gamma = \frac{f^0_{\text{therm}}}{f_{\text{t}}} - \alpha \frac{f^0_{\text{therm}}}{f^0_{\text{FCI}}}, \tag{35}$$

where f^0_{therm} and f^0_{FCI} are thermal and FCI induced focal lengths calculated from Eqs. 33, 34 when one of the beams is turned off.

We now have $-1 < \eta < 0$ as expected, with the resulting self-induced focal lengths in the range of $f_{\text{therm}} = 9.5 \times 10^3 \mu$m, $f_{\text{FCI}} = -1.5 \times 10^4 \mu$m. In addition, we have $\gamma > \alpha$, indicating that the competition between thermal and FCI effects is won by the former, enabling subsurface modification in Si.

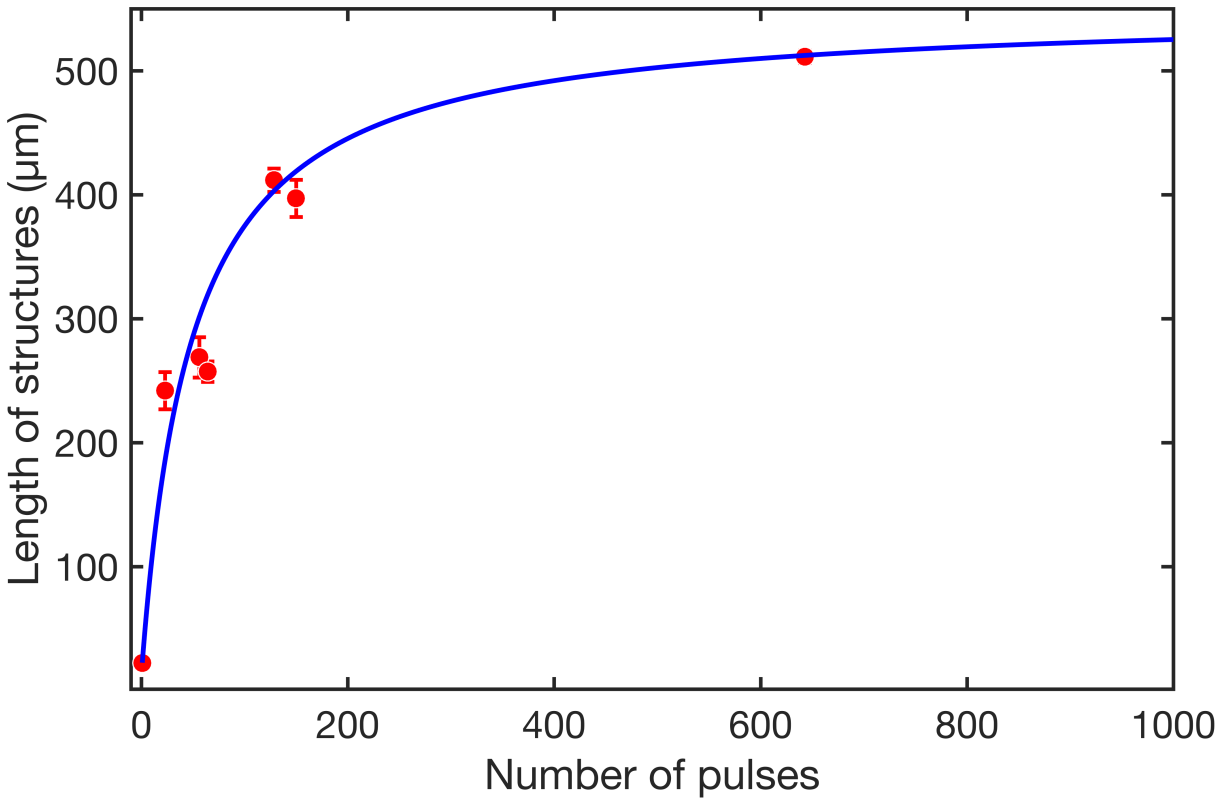

Figure S9: Experimental test of the second and third claims of the toy model. Experimental result (red circles) and the toy model prediction (solid line). In the toy model the total self-induced focal length for single pulse is $f_t = 26$ mm, resulting in $-1 < \eta < 0$ as expected. The error bars correspond to standard deviations from averaging lengths of 20 - 30 subsurface lines.

This level of control enables us to create subsurface structures with different aspect ratios as shown in Fig. S10. For instance, we demonstrate alternating short and long subsurface structures created with 20 and 200 pulses (Fig. S10a). As predicted by the toy model, the structures elongate with each laser pulse along the optical axis. For higher number of laser pulses, the structures self-regulate and stop before reaching the wafer surface. The structures also extend along the y axis, where their extent is not fundamentally limited. These multi-axis extended structures are notable, considering the difficulties in creating continuous subsurface microstructures[39].

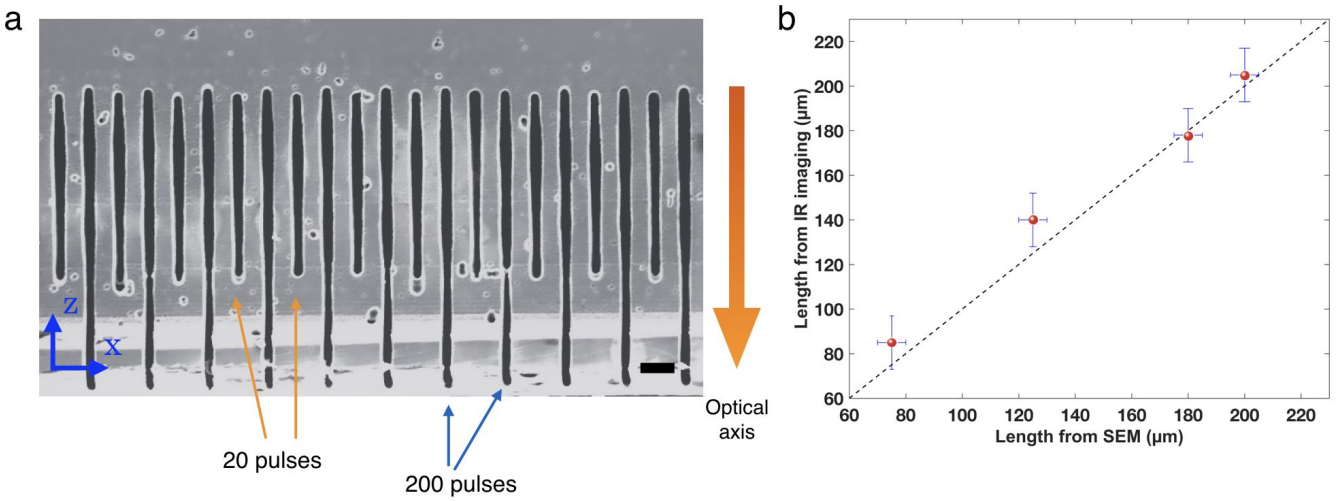

Figure S10: Controlled formation of structures with different number of laser pulses. **a**, Scanning Electron Microscope (SEM) images of high-aspect-ratio subsurface structures with different elongations. The structures form along the laser propagation direction (z axis), and are created either with 20 or 200 laser pulses. The scale bar is 40 μm. **b**, Comparison of structure length measurements from SEM images with *in-situ* measurements from infrared transmission microscopy.

The buried structures in Si are also diagnosed *in-situ*. Infrared (IR) transmission microscopy is used, such that relative positions of subsurface structures can be located and their aspect ratios can be determined. Fig. S10b compares the measurements from IR microscopy with SEM measurements, and both methods agree nicely. The fact that different material properties are used for these measurements (optical for IR microscopy, and electrical conductivity for SEM) indicate that the length measurements are very reliable.

5. Subsurface laser writing in doped Si

5.1 Complex structures in doped Si

Complex subsurface structures can be created both in undoped and doped Si wafers using structures of controllable length placed at any desired location by translating the beam position. Here, we give a complete set of experiments for creating these structures in n- and p-doped Si samples (1 Ω.cm). We will show increasing complexity all the way to 3D helix structures buried in Si. We start by evaluating wall-like 3D structures. These are evaluated with IR microscopy, and the structures in n- and p-doped Si are given in Fig. S11a and Fig. S11b, respectively. Sample cross sections have been evaluated with SEM, and the corresponding images are given in Fig. S11c and Fig. S11d. In all figures, z-axis corresponds to the laser propagation direction.

We studied any possible directional effects of scanning direction and polarisation in doped Si samples (Fig. S12). The experiment is repeated with laser polarisation rotated by 90 degrees. In both cases, vertical orientation is written first. We observed that the writing order can be changed, *i.e.*, horizontal lines can be written first (with either polarisation direction), without any discernible change in the pattern. Thus, the high-aspect-ratio structures can be created irrespective of scanning direction and polarisation, without suffering from directional heat accumulation effects. Such patterns can be used to create 3D architectures when combined with pattern transfer techniques, as shown in the creation of micro-pillar arrays; as well as purely optical objects entirely buried inside the wafer.

The already large parameter space is further expanded by the observation that one can create the structures in any scanning direction, or write arbitrary patterns in 3D. For instance, we show representative raster scans in Fig. S13, creating buried patterns at different orientations with respect to the crystal axes. We then show low-dimentional "dots" (Fig. S14a) to wall-like structures (Fig. S14b), to circles (Fig. S14c), 3D spirals (Fig. S14d) and finally a subsurface helix (Fig. S14e). These capabilities can be directly used to create functional elements inside Si, as demonstrated with Fresnel Zone Plates, and for information storage applications as will be shown in the next section, among others.

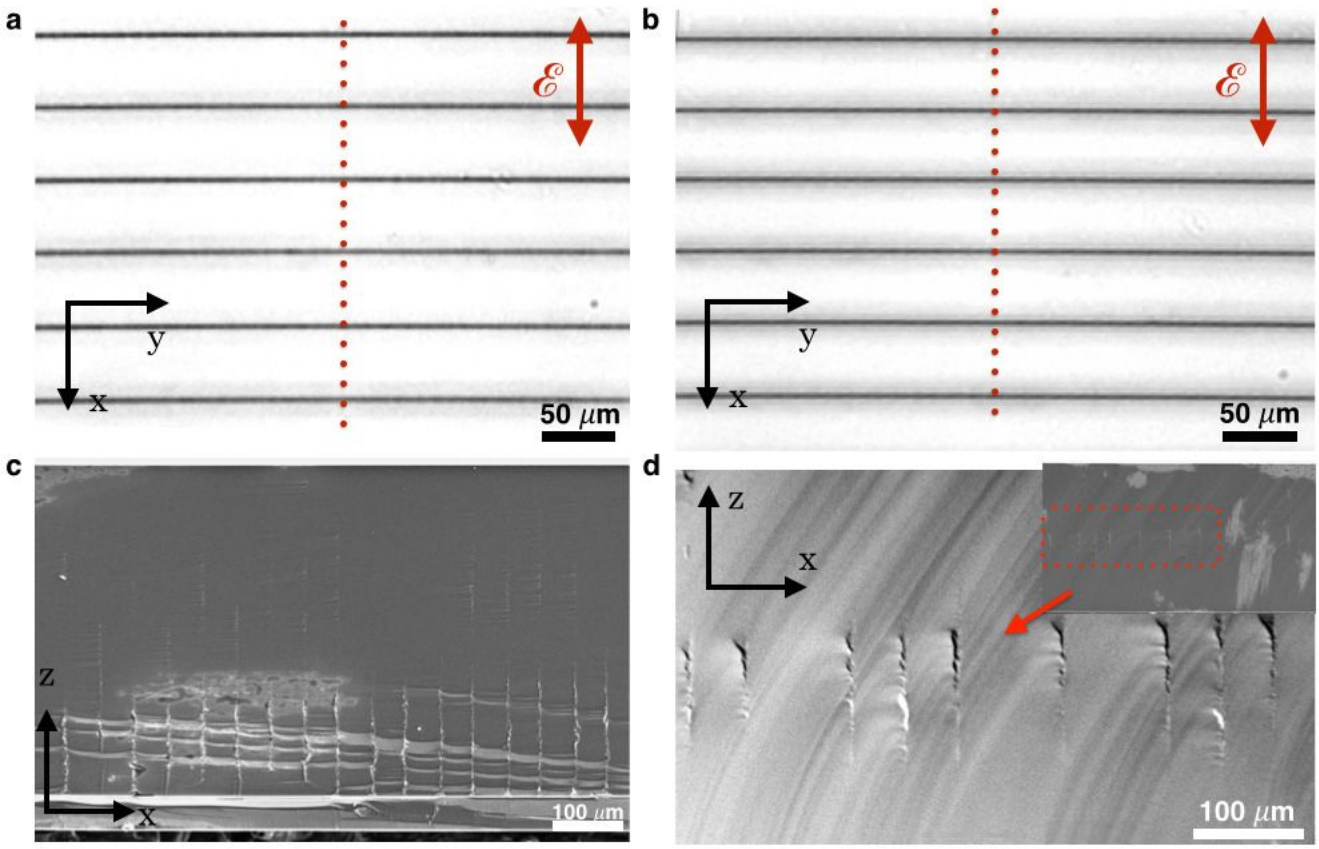

Figure S11: Laser scribing in n-doped and p-doped silicon. **a**, Subsurface 3D wall-like structures written in n-doped Si. The dotted red line shows the cross-sectional plane studied with the SEM. **b**, Subsurface structures created in p-doped Si. The dotted red line shows the cross-sectional plane studied with the SEM. **c**, SEM image of the cross-sectional plane indicated in (a). The laser entrance surface of the wafer is seen at the top of the image, the exit surface is seen at the bottom. **d**, Representative SEM image of the cross-sectional plane indicated in (b). The inset shows the full wafer. A zoomed image of the patterns is also shown. 20 μJ pulses were used.

We evaluated the interwall distance due to its relevance in creating in-chip photonic crystals. In Fig. S15, IR microscope images of wall-like structures with interwall separation ranging from 14 μm to 2 μm are given. When imaged with SEM at the cross-sectional plane (x-z plane), these samples demonstrate the uniform, high-aspect-ratio structures (Fig. S15f, Fig. S15g). We have created down to 2 μm separated subsurface walls with high control.

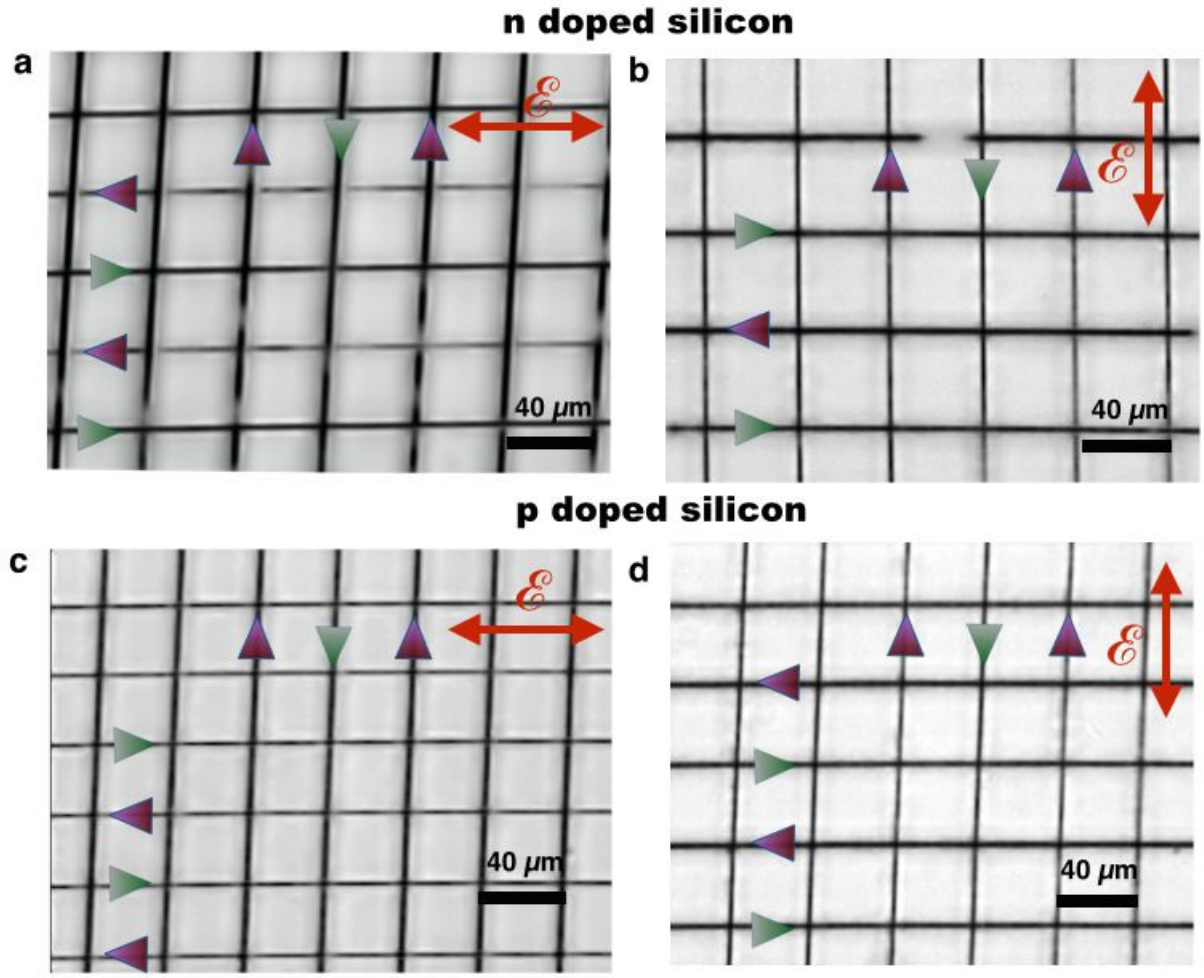

Figure S12: IR microscope images of mesh patterns in doped Si (1 Ω.cm) with alternating scanning directions. The subsurface wall-like structures are written parallel or perpendicular to crystal axes. The arrows indicate the scanning direction. Subsurface structures in n-doped Si with **a**, horizontally polarised light, **b**, vertically polarised light. Subsurface structures in p-doped Si created with **c**, horizontally polarised light, **d**, vertically polarised light. The pulses used were in the range of 20 μJ .

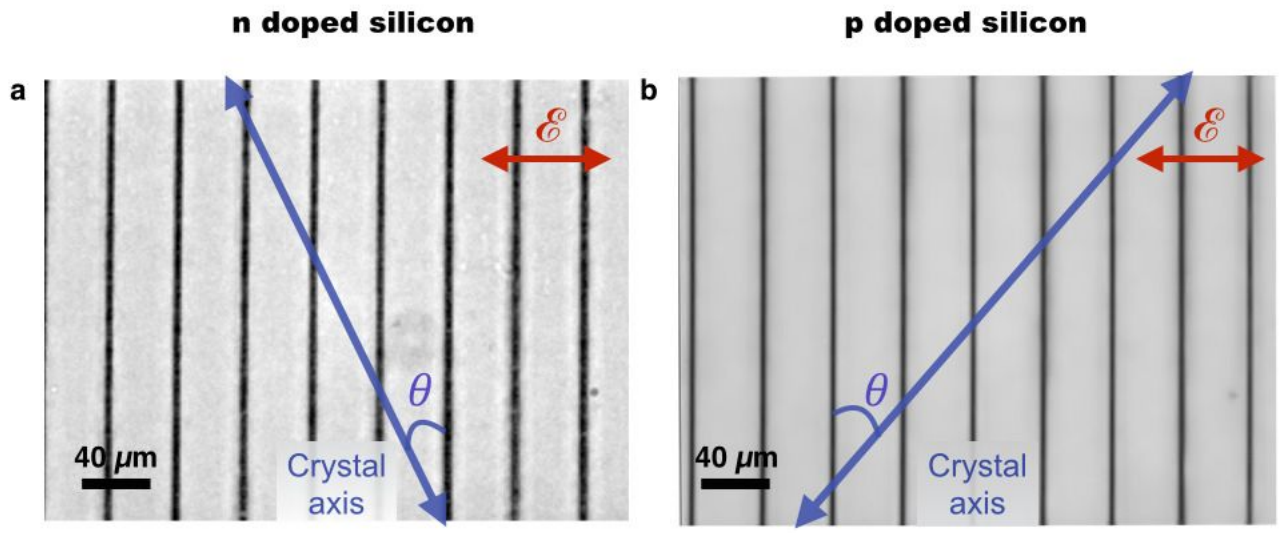

Figure S13: Subsurface patterns at arbitrary angles with respect to crystal axis in doped Si (1 Ω.cm). The blue lines indicate the crystal axis. **a**, The subsurface lines were written at $\theta = 27$ degrees with respect to the crystal axis in n-doped Si. **b**, The subsurface lines were scribed at $\theta = 49$ degrees with respect to the crystal axis in p-doped Si. For both samples, horizontally polarised laser pulses (20 μJ) were used, and the sample was scanned with alternating scanning directions.

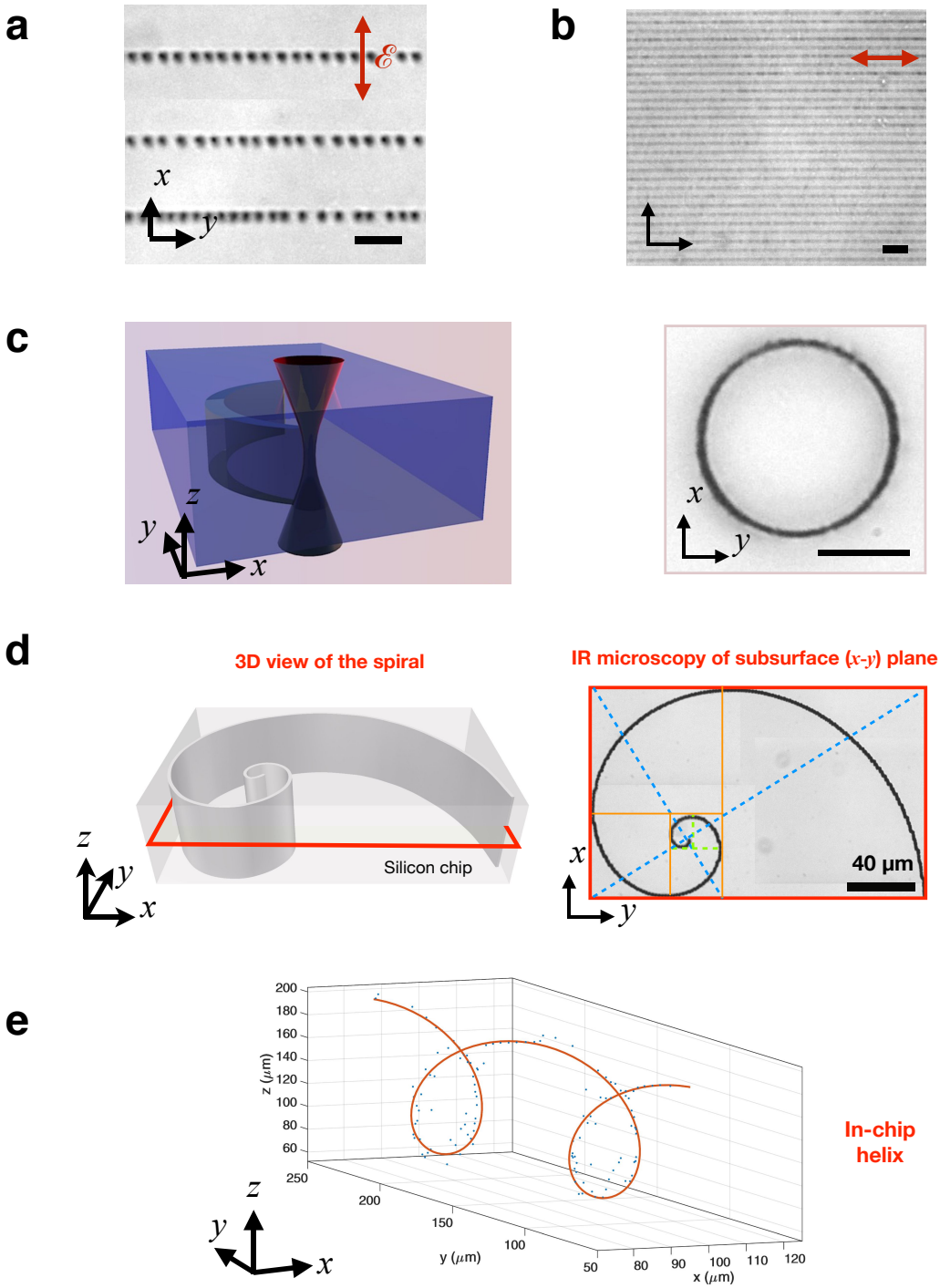

Figure S14: Increasingly complex structures buried in Si (1 Ω.cm). **a**, Discrete in-chip structures (n-type Si). **b**, Wall-like, uniform structures with small seperation (p-type Si). **c**, Schematic describing the creation of curved 3D structures (left). The dressing beam is not shown for clarity. IR microscope image of circular subsurface structure (right). Scale bars are 30 μm. **d**, Illustration of a buried Golden Spiral (left), and its composite IR image showing $x-y$ cross-section. **e**, A subsurface 3D helix structure is written with single laser pulses, and then reconstructured from 64 IR trasnmission mode images. Blue dots show experimental data, and the red curve indicates expected track of helix. At different depths dot-by-dot modifications require optimisation due to intensity dependence on depth and spherical aberration. The rod-like structures are not effected as much, as they are based on multi-beam, multi-pulse interaction.

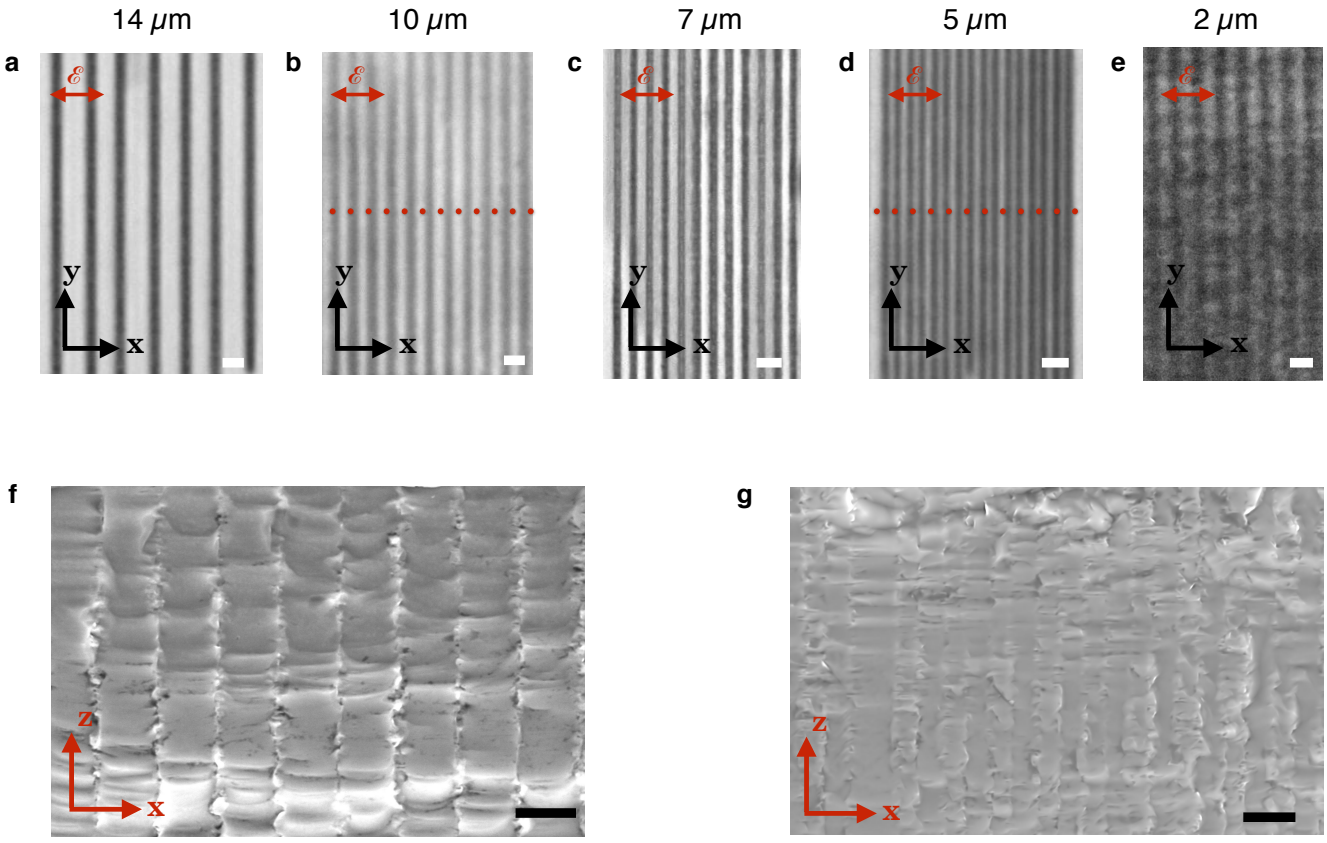

Figure S15: Control on interwall distance. Subsurface patterns produced in n-doped Si (1 Ω.cm). The interwall separations are, **a**, 14 μm, **b**, 10 μm, **c**, 7 μm, **d**, 5 μm, **e**, 2μm. **f**, SEM image of cross section from the sample in (*b*) along red dotted line. **g**, SEM image of cross section from the sample in (*d*) along red dotted line. All patterns were created by raster scans with alternating scanning directions. Horizontally polarised, ≈20-μJ pulses were used, and the samples were scanned at 0.2 mm/sec. All scale bars correspond to 10 μm, except in (*e*) where it is 2 μm. The laser propagates along the *z* axis.

5.2 3D information storage and erasure in Si

3D data storage in Si requires multi-level writing capability. We have demonstrated 25 levels along z axis as proof of concept, buried in a 1-mm thick wafer (7 levels are shown in Movie 3). The number of levels can easily be increased, with 1-μm feature sizes in all axes.

In order to create binary data storage elements in 3D, single laser pulses were used to create an array of "dots" in Si. These dots have have $\approx$ 1 μm features, and can be positioned anywhere inside the wafer. The bits represented by these volume elements were read with IR transmission microscopy and the information was numerically decoded with 96 % accuracy with a simple thresholding algorithm (Fig. S16b).

Afterwards, the laser-written structures could be erased nearly completely, following exposure to high temperatures (1100 °C) in an oven for 2 hours. The same digital readout protocol that correctly detected 96% of the written dots, false-detected only 3% of erased dots (Fig. S16c), highlighting the prospects for erasable and rewritable information storage and holography. In particular, the areal storage density with the available parameters imply storage densities surpassing bluray densities, on the order of 10 Gbits/mm^2.

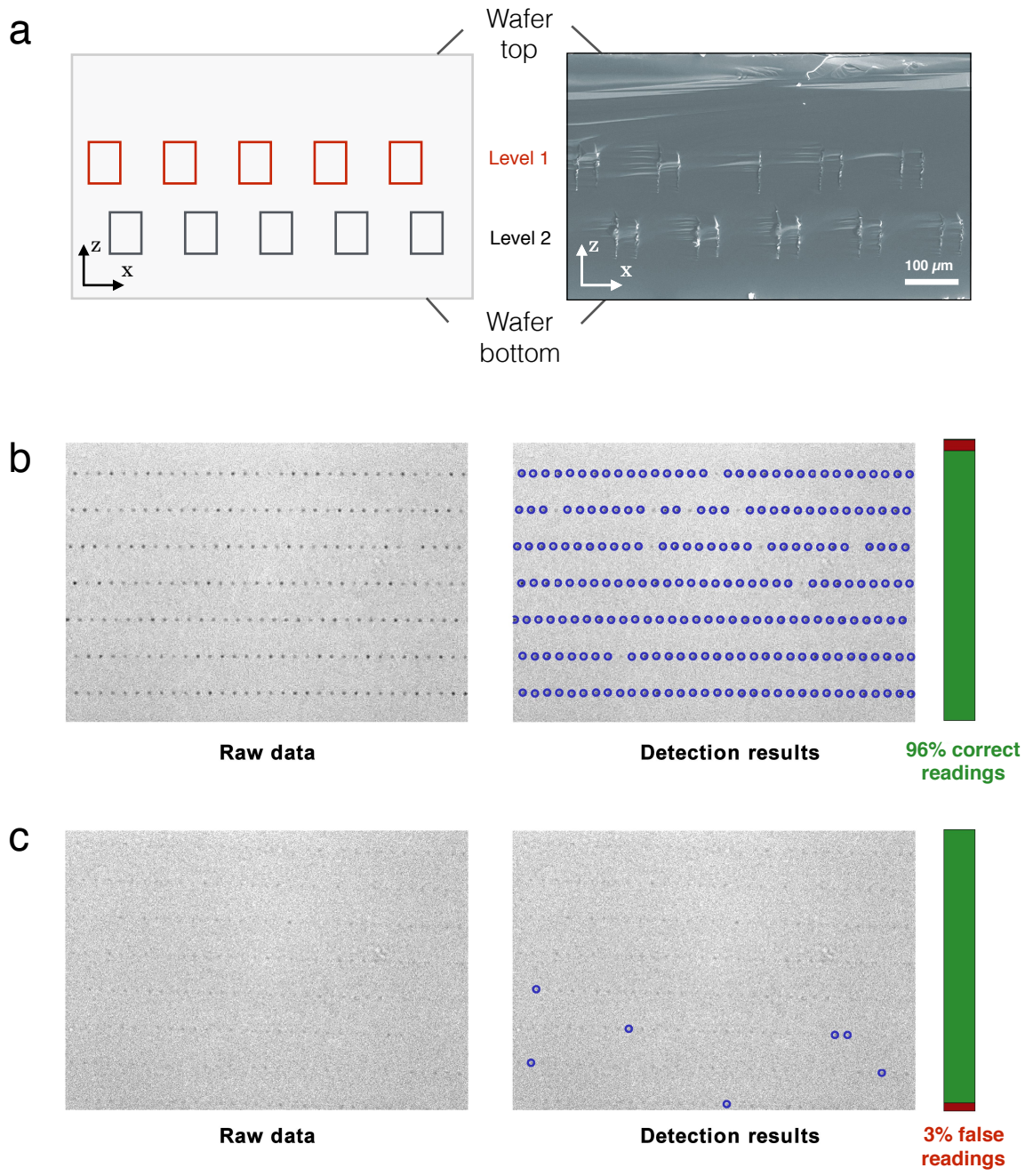

Figure S16: 3D data storage and erasure in Si. a, Illustration and representative SEM image of multi-level subsurface structures in Si. Longer structures are chosen in SEM, for ease of imaging, **b**, IR microscope image of buried "dots" and the result of numerical decoding to detect binary data stored in Si. The algorithm detected 96% of the bits correctly. **c**, The sample in (b) was exposed to heat (1100 °C, 2 hours), and subsequently the same decoding algorithm is applied to the IR image, with 3% false-detection rate. Binary information was encoded in p-doped Si (1 Ω.cm).

6. In-chip Computer Generated Holograms

We demonstrate a plethora of optical devices inside Si exploiting spatial amplitude or phase control. Through amplitude control, we demonstrate multilevel subsurface data encoding/decoding. Through spatial phase control inside Si, we extend the diffractive optics and holography capabilities into Si, whereby we construct the first functional optical elements in Si, such as buried Fresnel Zone Plates, gratings and further demonstrate in-chip holographic components. Since Si is transparent in the 1.2-7 μm regime, such diffractive components may find use in spectroscopy, sensing, filtering, anti-stealth, anti-counterfeiting, wavefront correction and imaging applications[40,41] for near- and mid-IR wavelengths[42].

6.1 Silicon as a medium for holography

Holograms are diffractive optical elements that can steer light by simultaneously or separately modulating phase and amplitude at the hologram plane, such that the hologram controls amplitude and/or phase at the image space[41]. Computer Generated Holograms (CGHs) were first developed for spatial filtering[43], and have since been used in a diverse set of applications including beam shaping for imaging and microscopy[44], optical manipulation[45], planar photonics[46] , and manipulating atomic beams[47]. Since the first demonstration of CGHs which were printed on photographic films[43], there has been intense research on recording media, and holograms have been recoded on photo-refractive materials[48,49], liquid crystal spatial light modulators[50,51], and most recently on graphene[52] and metamaterials[53–55]. In these applications, the recording media impose various limitations. For instance, while metasurface holograms provide high efficiency, they require circularly polarised illumination. Reflective holograms on metasurfaces are harder to align in comparison to transmission holograms. Metamaterial holograms are limited by the high losses of noble materials, and by their incompatibility with CMOS fabrication. Many of these limitations, such as reliance on circular polarisation, alignment problems and CMOS incompatibility issues can be overcome with subsurface Si CGHs. In addition, Si holograms may be used in conjunction with Si-photonics.

Holograms can be fabricated on surfaces or can be embedded in the bulk of materials. The latter type volume holograms have been demonstrated in glasses[56,57], photrefractive crsytals[58], and in photopolymers[59]. Similar subsurface Si holograms are highly desired, but have not been possible so far. Here,

we report the first functional holograms created within the bulk of silicon wafers. These buried holograms are enabled by Nonlinear Laser Lithography (NLL), which provides previously unattainable phase control inside the wafer. In comparison to amplitude holograms, phase holograms allow fewer pixel counts to provide the same image quality. They can be used in conjunction with amplitude holograms to produce complex holograms. Further, the durability of embedded holograms is expected to be longer, in comparison to surface holograms. We finally note that, since the refractive index modulation is quite low in glass embedded holograms[53], phase-type holography has not been viable in those materials.

We designed and implemented Fourier-type and Fresnel-type CGHs inside Si, to demonstrate 2D wavefront structuring and also 3D image reconstruction. First, digital synthesis algorithms are developed for Si holography. Then, CGHs are constructed with these algorithms, encoded in Si with our NLL method, and finally optically reconstructed images from the CGHs are compared with the simulated images. In the next section, we present the details of the developed algorithms.

6.2 Generation of CGHs with a modified iterative Fourier algorithm

Hologram diffraction can be in the far-field or near-field zones, corresponding to Fourier and Fresnel holograms, respectively. Fourier holograms, in general, require less computational power to synthesise in comparison to Fresnel holograms, and also need fewer pixels to reconstruct the same 2D images. In contrast, Fresnel holograms can create 3D images with superior depth perception. In practice, one can realise Fourier holograms with projected images within their near field by implementing a Fourier lens, which optically transforms the associated diffraction equation from Fresnel to Fraunhofer regime[40]. With these considerations, Fourier-type CGHs are selected for 2D wavefront structuring, while Fresnel-type holograms are chosen for 3D image reconstruction.

Computer generated holograms are usually designed with algorithms such as Iterative Fourier Transform Algorithms (IFTAs), Detour-Phase, and Lee algorithms[40,41]. These algorithms rely on the modulation mode of the hologram, *i.e.*, phase or amplitude. In order to take full advantage of phase-type Si subsurface holograms, a modified IFTA is used. To this end, we modified the spectral condition of the Adaptive-

Additive IFTA[60], which, in turn, can be considered as an improvement on the better-known Grechberg-Saxton algorithm[61,62]. Moreover, we expand the noise space of the iterative generation process to give the algorithm more freedom for improving the quality of the reconstructed image, which will be explained later in detail. This customised algorithm allows creating binary phase holograms (*i.e.*, binary Kinoforms) in Si, which produce binary and greyscale images with high fidelity.

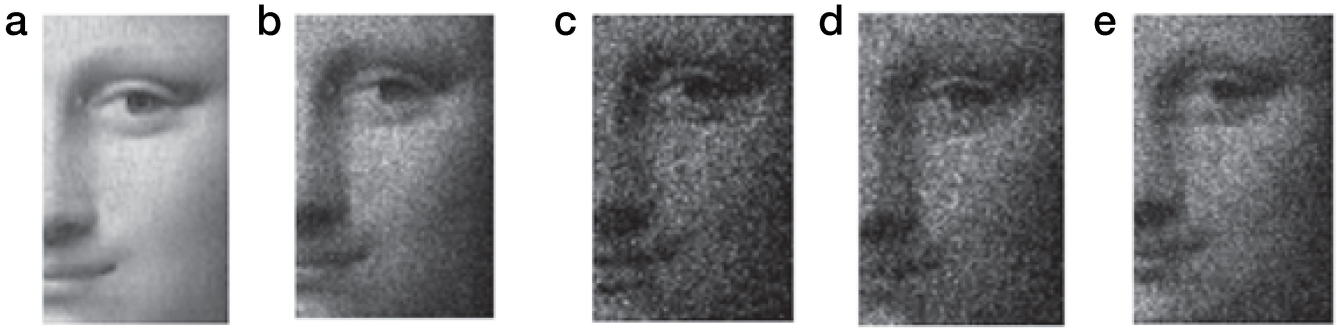

Figure S17: Simulations for optimising the Fourier binary phase hologram generation algorithm. **a**, A detail of the original image. **b**, The simulation result of the reconstructed image of greyscale kinoform (256 Levels) generated by adaptive-additive IFTA algorithm. **c**, The simulation result of the reconstructed image of binary kinoform generated without IFTA algorithm. **d**, The simulation result of the reconstructed image of binary kinoform generated with binarized adaptive-additive IFTA algorithm. **e**, The simulation result of the reconstructed image of binary kinoform generated with binarized adaptive-additive IFTA algorithm after increasing noise space.

In general, image quality of binary holograms is lower than that of greyscale holograms[51]. This problem is overcome with the modified Adaptive-Additive IFTA, by taking advantage of the expanded noise space in the iterative process, such that the generated kinoform improves the image quality while reducing speckles at the target plane. The improvement afforded by the developed algorithm is presented in Fig. S17, where various versions of CGH algorithm are compared. For reference, a portion of the original image is given in Fig. S17a. In addition, a greyscale (256 levels) kinoform is generated with an additive-adaptive IFTA and the reconstructed image is given in Fig. S17b for reference. The next set of images progressively illustrate the improvement in the algorithm. First, a binary phase-type kinoform is generated without IFTA (Fig. S17c). Then, a binary kinoform is generated with the binarized adaptive-additive IFTA algorithm and the simulation result of the reconstructed image is given in Fig. S17d. Finally, the modified adaptive-additive

IFTA algorithm is used to generate the binary phase-type hologram and the corresponding reconstructed image is given in Fig. S17e. Note that the improved binary phase-type algorithm (Fig. S17e) performs similarly to a greyscale hologram (Fig. S17b).

To gain more insight on how the CGHs are generated, and to shed light on their performance, we present a conceptual schematic (Fig. S18) explaining the adaptive-additive IFTA and modifications implemented into the algorithm. The figure incorporates the adaptive-additive IFTA with typical initial steps, where random phase is added to the target image (Steps 1 and 2). This computational addition on the target image can be regarded as a process that compensates for the elimination of amplitude distribution of Fourier complex field at the hologram plane (Step 5). Next, 3 modifications are performed to take advantage of the various constraints and degrees-of-freedoms of IFTAs. In the case of kinoforms, it is customary to impose constraints on the amplitude distributions at both the hologram and image planes, while the algorithm is free to distribute phase over both these planes. Thus, phase spaces can be considered as noise space from the perspective of the final image, which can be computationally exploited as a degree-of-freedom. To take advantage of the noise space, first a DC bias is applied to the amplitude distribution of the source image (Step 3). This ensures that the phase is not cancelled in parts where it is weighted by zero amplitude. The bias value should be chosen carefully to improve image quality without sacrificing image contrast and hologram efficiency. Next, a frame of random amplitude and phase is padded around the target image (Step 4), such that the algorithm can freely vary both the amplitude and phase within the enlarged noise space. If the size of this frame is chosen judiciously, reconstructed image improves in quality, while artifacts can be kept at experimental noise levels. Finally, two constraints are imposed on the hologram plane, that is the amplitude distribution is flattened and the phase distribution is binarized (Step 5). This latter constraint somewhat limits the ability of the hologram to manipulate light at its imaging plane, which explains why binary-phase holograms generally project lower quality images in comparison to greyscale holograms.

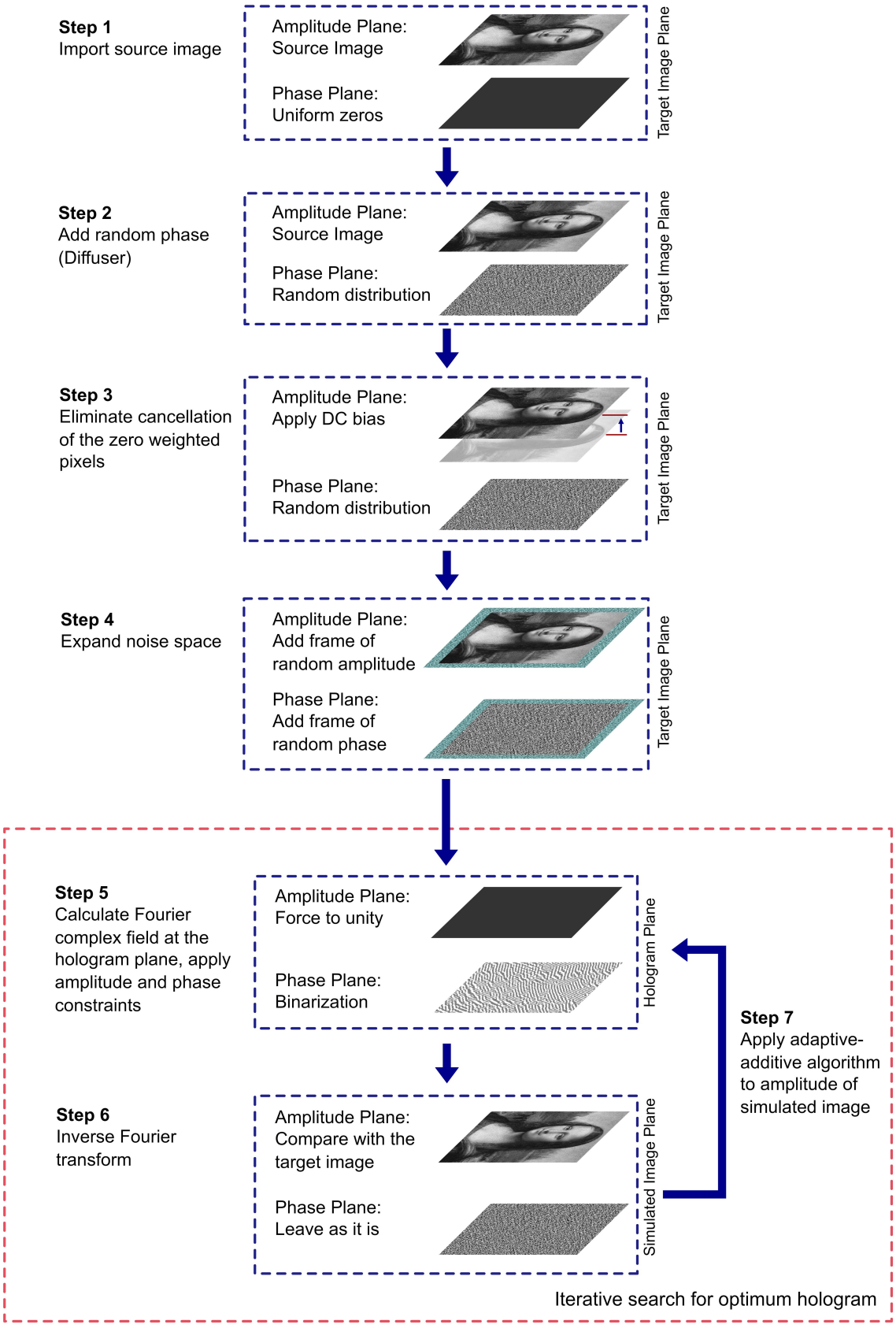
Step 1
Import source image
Amplitude Plane:
Source Image
Phase Plane:
Uniform zeros
Target Image Plane
Step 2
Add random phase
(Diffuser)
Amplitude Plane:
Source Image
Phase Plane:
Random distribution
Target Image Plane
Step 3
Eliminate cancellation
of the zero weighted
pixels
Amplitude Plane:
Apply DC bias
Phase Plane:
Random distribution
Target Image Plane
Step 4
Expand noise space
Amplitude Plane:
Add frame of
random amplitude
Phase Plane:
Add frame of
random phase
Target Image Plane
Step 5
Calculate Fourier
complex field at the
hologram plane, apply
amplitude and phase
constraints
Amplitude Plane:
Force to unity
Phase Plane:
Binarization
Hologram Plane
Step 6
Inverse Fourier
transform
Amplitude Plane:
Compare with the
target image
Phase Plane:
Leave as it is
Simulated Image Plane
Step 7
Apply adaptive-
additive algorithm
to amplitude of
simulated image
Iterative search for optimum hologram

Figure S18: The modified adaptive-additive iterative Fourier algorithm employed to generate Si subsurface kinoforms. First, the source image is imported (Step 1) and random phase distribution is added (Step 2). Next, a small DC bias is applied on the greyscale amplitude of the source image (Step 3). Then, a frame of both random phase and amplitude is added around the source image (Step 4). The frame is added to increase the noise space so as to give more degrees of freedom to the algorithm. Step 4 concludes the preparations for the iterative routine. In the iterative part, a uniform amplitude is imposed on the amplitude plane of the hologram's complex field. Further, the phase distribution is binarized (Step 5). The simulated image of the hologram is compared to the source image (Step 6). Here, the normalised mean square error is used to decide whether to continue or terminate the iteration. If continued, the algorithm reiterates from Step 5, where the input now is taken as the simulated image in Step 6, but after modifying its amplitude distribution using the amplitude distribution of the source image (adaptive-additive equation[63]) (Step 7). The converged hologram is implemented inside Si with NLL.

The remaining steps are common to IFTA methods. In step 6, the optical image of the generated hologram is simulated by inverse Fourier transform. Here the mean square error is used to compare the simulated image with the source image, and if they are similar enough, the iterative loop ends. Otherwise, the algorithm reiterates from Step 5. In each iteration (Step 7), the input is taken as the simulated image in Step 6, but after modifying its amplitude distribution using the amplitude distribution of the source image (adaptive-additive equation[63]).

The preceding modifications to IFTA in Steps 3 and 4, give more degrees of freedom to the algorithm, and thus binary phase holograms perform similar to greyscale holograms in simulations (Fig. S17). Using the developed algorithm, we are able to synthesise binary Fourier holograms and write them in Si, such that they can project binary or greyscale images with high fidelity.

We further modified the algorithm to enable 3D image reconstruction by exploiting Fresnel holography. The preceding kinoform technique is exploited to generate these Fresnel holograms, without the need to directly solve the Fresnel diffraction equation. First, a stack of binary Fourier holograms are generated, each of which is able to generate the image slice of a 3D object. Then the hologram stack is used to create a single binary hologram, with appropriate normalisation and binarization steps, along with superposed Fresnel Zone Plates (FZPs). Here, the function of the phase-type FZPs are to shift each image to a different

plane, such that individual images collectively form a 3D object. The generated Fresnel hologram was also tested on a phase-only Spatial Light Modulator (SLM, Hamamatsu X10468-03) which helps in the design of 3D holograms.

We thus demonstrate light steering and wavefront structuring with four types of functional optical elements inside Si: a binary Fresnel zone plate, diffractive gratings, two types of binary Fourier phase holograms for high quality binary- and greyscale image projection, and a binary Fresnel phase hologram that can produce 3D images with high fidelity. Next, implementation of Si CGHs are described in detail.

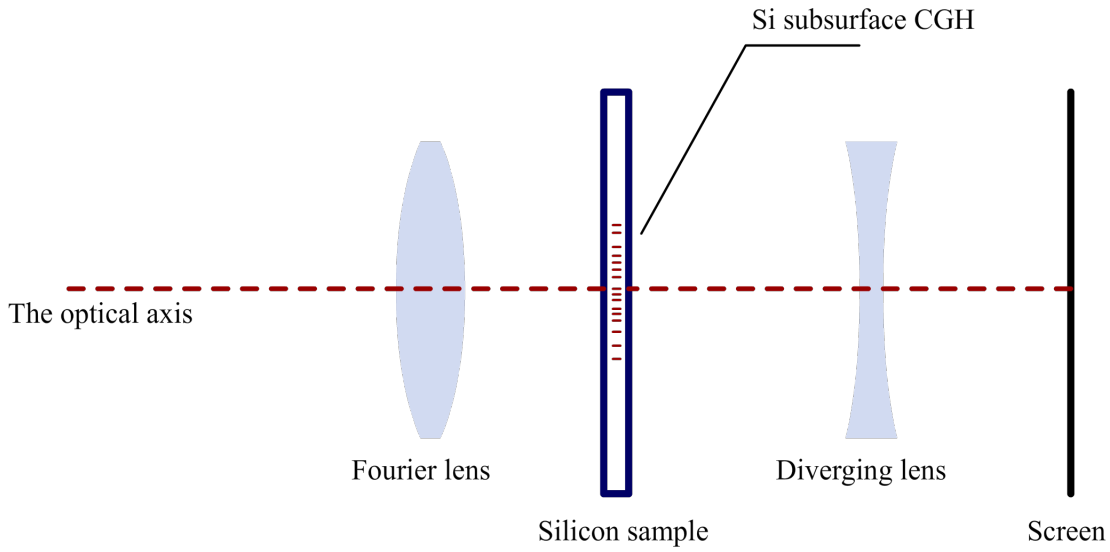

Figure S19: Optical setup used to reconstruct images of subsurface silicon Fourier kinoforms. The optical reconstruction setup can be kept simple thanks to the fact that CGHs are transmission based.

6.3 Implementation of the holograms

We demonstrate wavefront-structuring capability with two types of Fourier holograms. The first type of hologram proves that high-frequency spatial frequencies can be reproduced, while the second type hologram reproduces greyscale images with high fidelity. For the first case, a binary image involving high-frequency spatial components is chosen. For the second case, to illustrate wavefront structuring capability, a greyscale image (Leonardo da Vinci's *Mona Lisa*) is selected. Experimental reconstruction of

the holograms is performed with an optical setup (Fig. S19), which is coupled to a collimated output of a 1030 nm laser. A lens is included to form the Fourier image in the Fraunhofer regime, which is placed before the hologram. A diverging lens is used to expand the reconstructed image further, for better image recording. A spatial filter is included to eliminate higher order images.

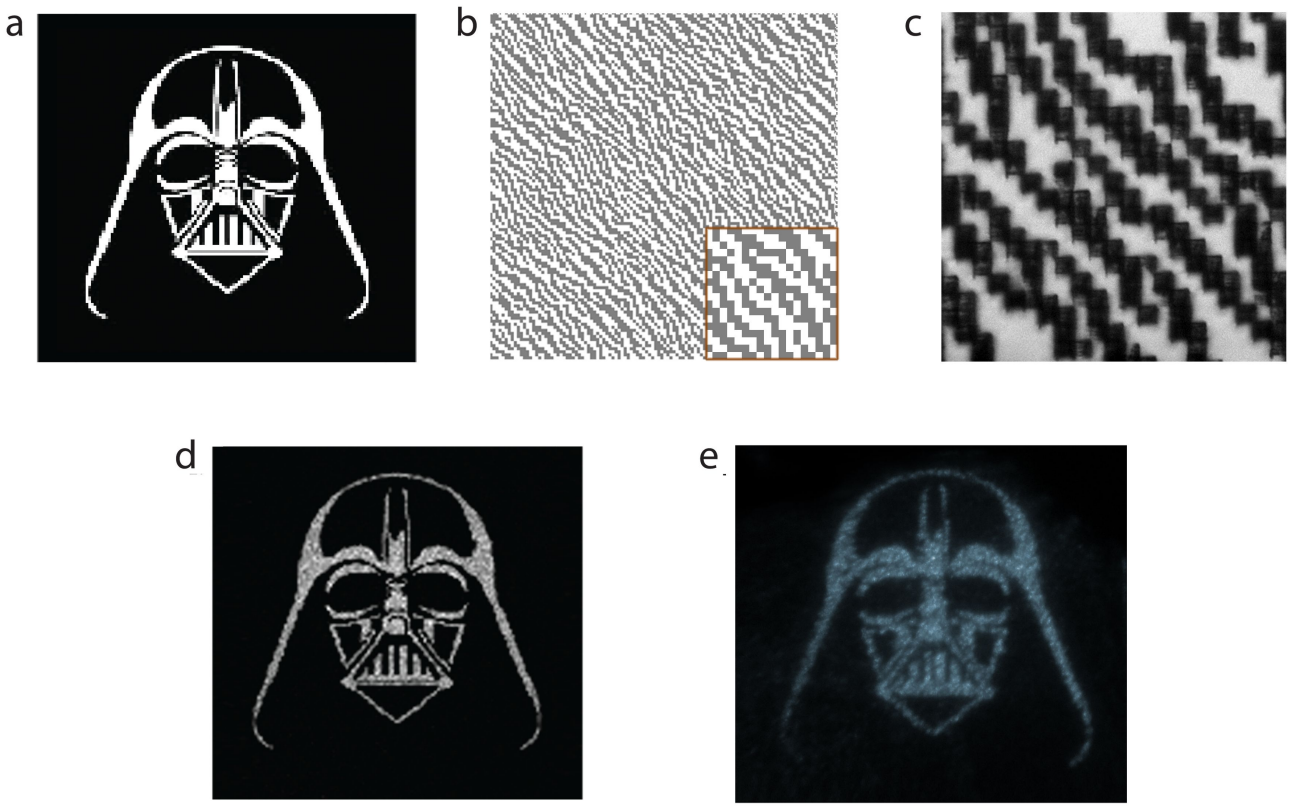

Figure S20: The digital synthesis of the binary hologram and experimental reconstruction of the corresponding binary image. a, The original image, including high-frequency spatial components. **b**, The laser written CGH in Si. The inset shows an zoomed portion of the hologram. **c**, IR microscope image of a portion of the fabricated hologram in Si. **d**, The simulated image from the digital synthesis algorithm (first order only). **e**, Optical reconstruction of the image (first order) from the CGH in Si.

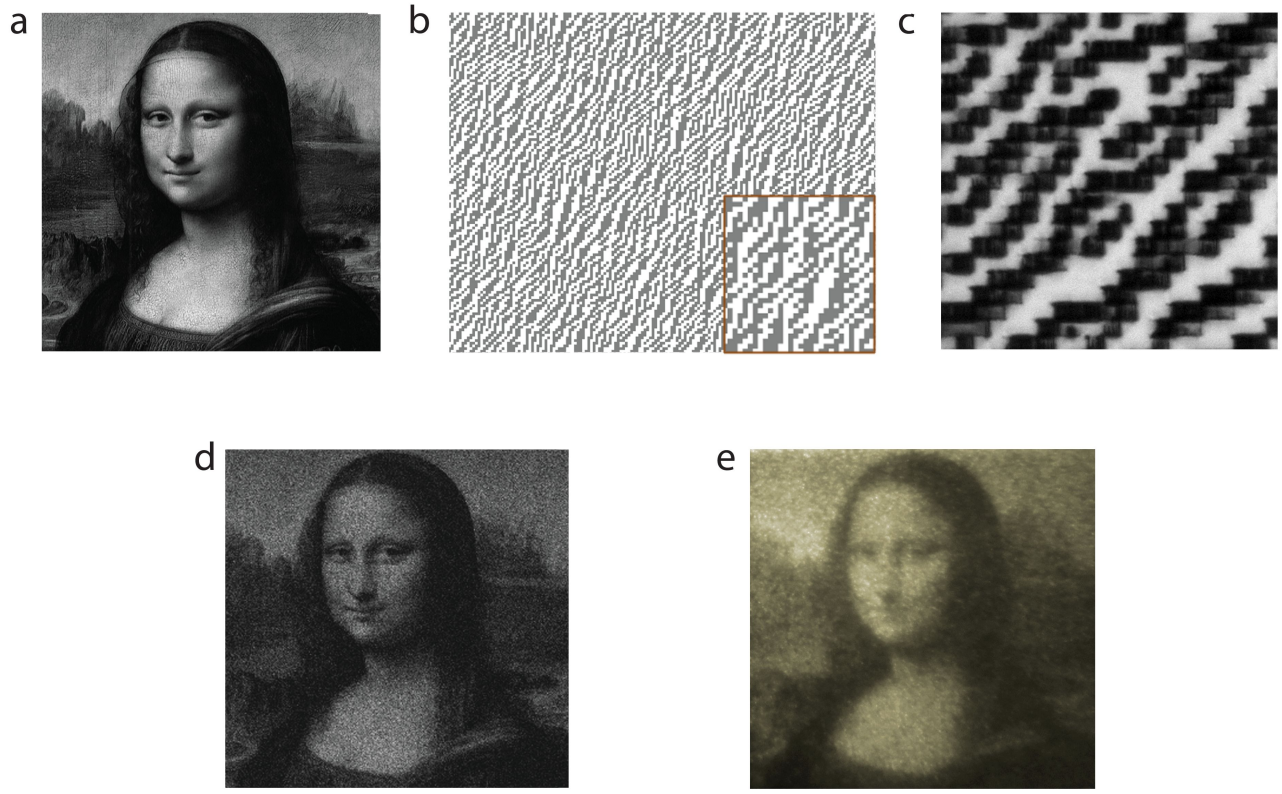

Figure S21: The digital synthesis and experimental images of the *Mona Lisa* hologram. **a**, The original greyscale image. **b**, The 800 × 600 CGH for Mona Lisa. The inset shows an zoomed portion of the hologram. **c**, IR microscope image of a portion of the fabricated hologram. **d**, The simulated image from the digital synthesis algorithm (first order only). **e**, Optical reconstruction of the image (first order) from the CGH in Si.

The reconstructed images on the screen are recorded with a digital camera (Canon, SX710 HS) after removing its infrared filter. Fig. S20 and Fig. S21 show optical reconstructions of the binary image and *Mona Lisa*, and their corresponding digitally synthesised holograms. Simulations and experiments are seen to be in very good agreement. We note that the images have less speckle when observed by eye in comparison to the case with a digital camera, which is a common case in Kinoform imaging, due to parasitic interference on the protective layer of CMOS sensors. The hologram size in Fig. S20 is 256 x 256 pixels, generated from a source image of 128 x 128 pixels. The hologram size in Fig. S21 is 800 x 600 pixels, generated from a source image of 595 x 595 pixels. The pixel size is 10 μm in both cases, which is

about 10 times the imaging wavelength and thus scalar diffraction theory is applicable. We note that the minimum spacing between the structures we can create is 2 μm, which is likely to allow the fabrication of metasurface holograms operating in the mid-IR wavelength range of 5-10 μm.

In order to demonstrate 3D image reconstruction capability of in-chip Si holograms, we employ a binary Fresnel hologram. The hologram size is 800 x 600 pixels with a pixel size of 10 μm. The projected image by the hologram follows the outline of a twisted rectangular prism in 3D, with rotated and scaled rectangular cross-sections at different planes. The Fresnel hologram's 2D reconstructed images at 4 locations are presented in Fig. S22. The length of the twisted rectangular prism is 15 cm affording $\pi/2$ rotation. The 3D projection can clearly be seen as rotation when the screen is moved at a constant speed (Supplementary Video 2) along the optical axis. The optical reconstruction of the Fresnel hologram and the corresponding video is created with the setup given in Fig. S22.

6.4 Hologram efficiency

The hologram efficiency is assessed by the ratio of power in the first order to the power in the zeroth order[64]. In order to evaluate efficiency, we fabricated a Si subsurface phase-type grating, with 10-μm pixel size, 50 lines/mm and over a 2 mm x 2 mm area. For this diffracting element, efficiency is given as[64], $R = F(4/\pi^2)sin(\Delta\phi/2)^2/(1 - F + Fcos(\Delta\phi/2)^2)$, where $\Delta\phi$ is the phase modulation depth and F is the filling factor. With 1.03 μm linearly polarised laser illumination, the first-to-zero order ratio was measured 150 %. This corresponds to $\Delta\phi = 0.69\pi \pm 0.04\pi$, which matches to the value of $\Delta\phi = 0.69\pi \pm 0.01\pi$ measured with interferometry.

We finally comment on some properties of subsurface Si structures and CGHs relevant for diffractive optical elements. Si subsurface holograms can perform with both linear and circular polarisations. This feature simplifies hologram operation, with potential applications in communications and optical information processing[40]. We observed birefringence effects in the subsurface structures, which may pave the way

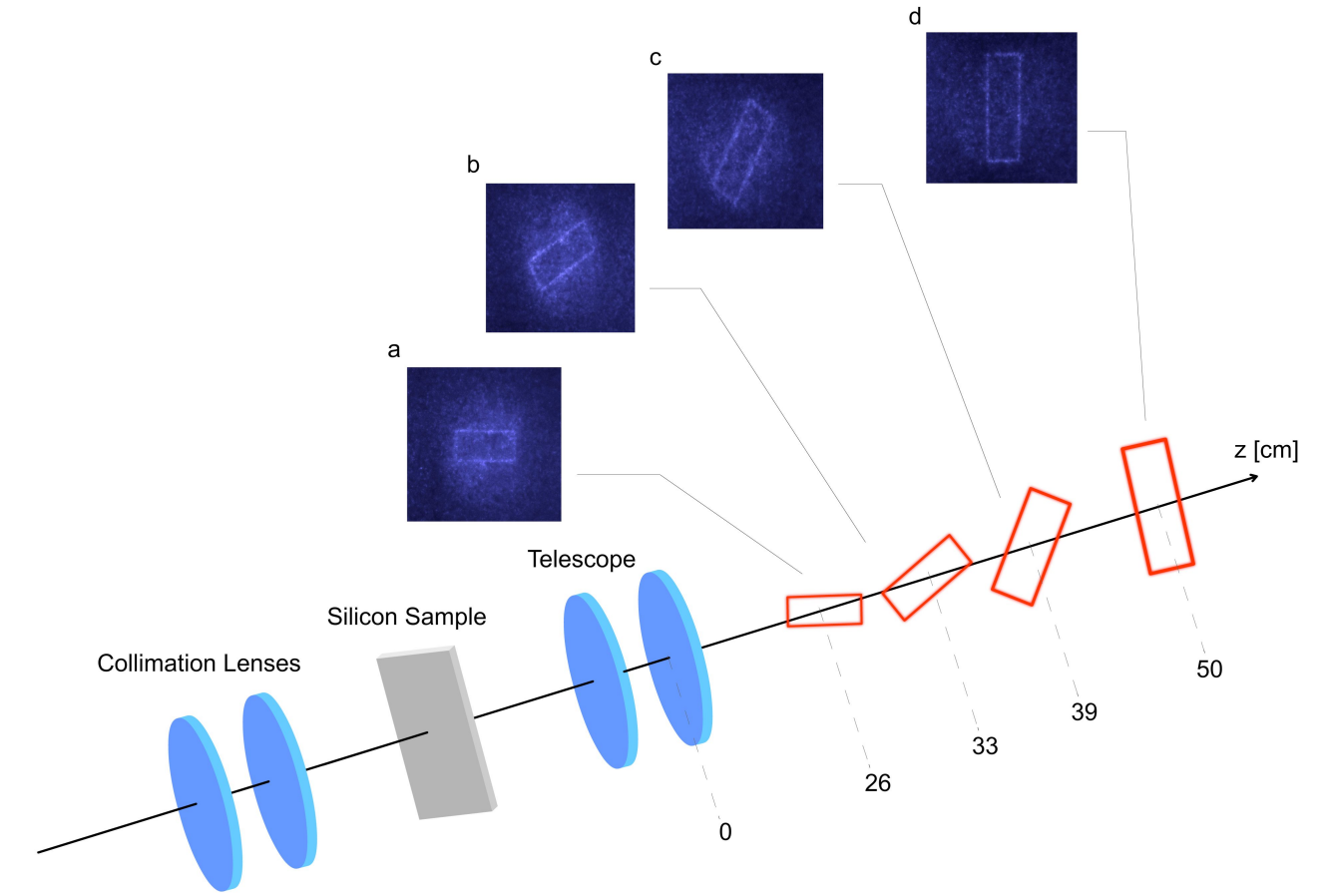

Figure S22: The optical setup used to reconstruct Fresnel images of Si subsurface holograms. A collimated laser beam passes through the buried hologram in Si. A 6X telescope was used to facilitate the use of camera in capturing the 3D projection. The numbers indicate the relative locations of the images.

towards producing polarisation-dependent optical elements in Si, such as polarisers or wave plates. As we have shown, multilevel subsurface structures are possible in Si, which may be used to create multilevel or greyscale holograms. In addition, combined with surface holograms, complex holograms may be possible for controlling both the amplitude and phase at the imaging plane.

7. 3D Sculpting in Si with Preferential Etching

7.1 Chemical etching process

In order to selectively remove the laser modified subsurface regions, a reliable and repeatable wet-chemical-etching process is needed. The etching solution should ideally have selective reactive behaviour to defects, operate at room temperature, be stable over long times and CMOS-compatible.

Towards these goals, we developed a highly-selective wet-chemical-etching process and achieved successful removal of the laser-modified Si regions with minimal damage to crystalline Si. This enabled sculpting in laser-treated Si to create various 3D architectures, including micropillar arrays and microchannels. The etching solution is comprised of a combination of copper(II) nitrate ($Cu(NO_3)_2$), hydrofluoric acid (HF), nitric acid (HNO_3) and acetic acid (CH_3COOH) with 0.05 M, 10 M, 4 M and 3.5 M concentrations, respectively. The chemical process corresponds to the dissolution of Si through oxidation[65]. In the solution, HNO_3 is used as an oxidising agent while HF dissolves the oxidised Si[66,67]. CH_3COOH is used as diluent, and participates in reducing the reactant concentration. CH_3COOH also maintains better wetting of Si surface through relieving solution's surface tension[68], thus resulting in better surface profile for etched Si. Finally, $Cu(NO_3)_2$ is used as a catalyst to increase the selectivity with relatively low activation energy[69,70].

Chemical etching of semiconductor materials is an electrochemical process taking place through local currents flowing between local anode and cathode sites formed on surfaces by oxidation-reduction reactions[67,68]. Consequently, the chemical mechanism of etching process includes production of excess holes and electrons, which facilitate the charge transfer between induced local electrodes. Various possible anode and cathode reactions have been proposed to characterise anodic etching of Si in HF/HNO_3 based solutions. In our case, since H_2 gas is generated during Si dissolution with the developed etchant, direct dissolution of Si in divalent state is suggested[65]. Further, the colour gain of etchant from light blue to green during etching indicates the existence of nitric oxide (NO)[70]. The production of NO signifies that the reduction of HNO_3 to nitrous acid (HNO_2) continues with further reduction to nitric oxide[66,67,71].

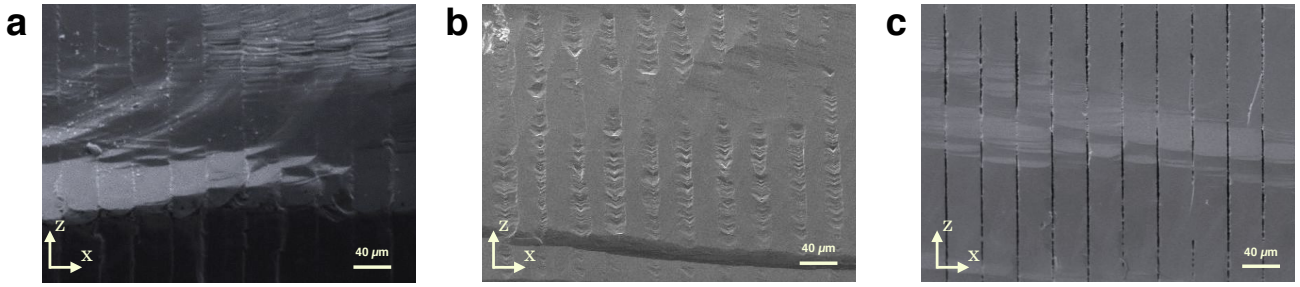

Figure S23: Chemical etching of laser-processed wafers, evaluated with Scanning Electron Microscopy (SEM). a, Cross-section of the wafer after laser treatment, but before etching. **b**, Cross-section of the processed area after etching for 10 minutes in KOH solution. **c**, Cross-section of the processed area after etching for 10 minutes in the developed etching mixture. The developed solution predominantly removes the laser-modified areas, allowing for controlled removal of these areas, and thus for controlled 3D sculpting of Si.

Considering these observations, the anode and cathode reactions for the $Cu(NO_3)_2$/HF/HNO_3/CH_3COOH etchant is:

Cathode reactions:

$$Cu^{2+} + 2e^-_{VB} \rightarrow Cu^0, \tag{36}$$

$$HNO_3 + 3H^+ \rightarrow NO + 2H_2O + 3h^+, \tag{37}$$

Anode reaction:

$$Si + 4HF_2^- \rightarrow SiF_6^{2-} + 2HF + H_2 + 2e^-. \tag{38}$$

The oxidizing HNO_3 agent and Cu^{2+} ions not only provide reduced material at the local electrodes, but also modify the carrier densities at the etching interface[67,68]. Required holes for current flow through local electrodes are provided by means of the cathodic reduction of oxidising agent and electron capturing of Cu^{2+}. HNO_3 in solution prevents the aggregation of Cu atoms and nanoparticle formation[72]. Further, any solution related contamination can be cleaned without adversely interfering CMOS production[73].

The selective behaviour of the developed etching solution on the laser-modified regions is evaluated in comparison to the commonly used isotropic and anisotropic etchant solutions for Si, including KOH and other HNO_3 based oxidising etchants. As a representative case, the performance of the developed etchant is compared with KOH solution, and SEM images of etch profiles of these etchants on subsurface patterns

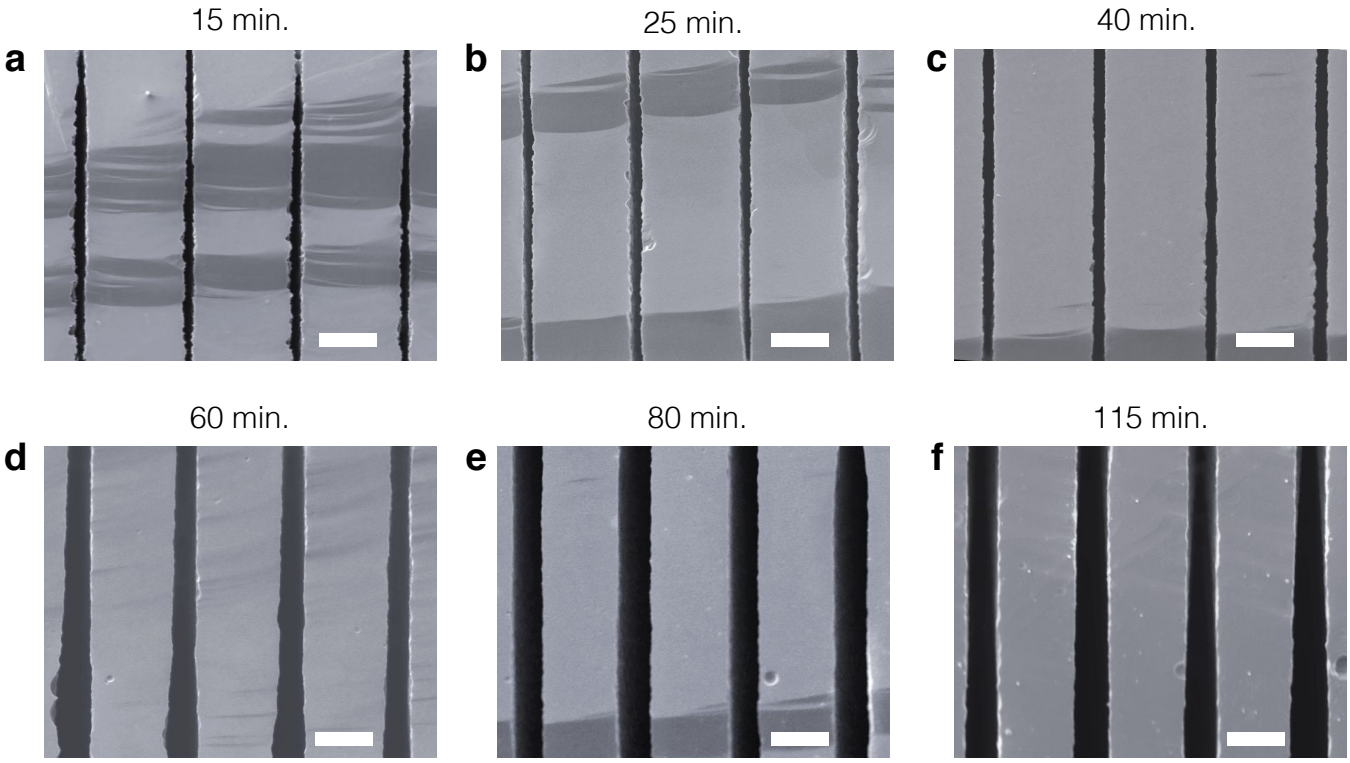

Figure S24: Characterisation of the chemical removal of the laser-modified areas with the developed etchant. a – f, The SEM images show the wafer cross-section after etching in the developed mixture for various durations. All scale bars indicate 20 μm.

are given in Fig. S23.

This selective behaviour can be explained considering the anodic reaction at the boundaries between crystalline Si and the modified regions. These areas are expected to have electronically active defects and different dopant concentrations in comparison to crystalline Si, producing variations in the Cu^{2+} ion concentration. The etch rate difference between the interface and crystalline Si can be attributed to this uneven profile of Cu^{2+} ions, which have a higher electronegativity in comparison to Si (electronegativity for Cu is 1.9 and for Si is 1.8)[65,74]. This results in the selective etching behaviour at the defect sites.

The laser modified regions are etched away at a much faster rate, which varies depending on the extent of laser modification and the surface area available for chemical reaction. A representative etching sequence is illustrated in Fig. S24. We note that as the optimised etchant is strictly defect selective, the performance

of the etching process correlates with the crystal quality of the wafer.

7.2 Micropillar arrays as example of regularity and repeatability

Laser-induced subsurface modifications in Si can be selectively removed with the developed chemical etching method to create large-area-covering 3D complex structures. Prior to etching, access to modified areas is gained through polish removal (for *x*-*y* plane of attack) or dicing (for *x*-*z* plane of attack). For the former case, a large-area periodic set of pillars is revealed in 20 minutes. The pillars have 20 μm by 30 μm top surfaces, and 500 μm heights, which can conceivably be used for MEMS applications.

7.3 Through-Si vias

Through-Si vias are widely used for three-dimensional large-scale chip integration. The developed method can be used for creating through-Si vias as illustrated in Fig. S25. First, in-chip cylinders to be used for vias are created with the laser. Each cylinder is created with a single circular motion of the laser with respect to the sample (Fig. S25a). The laser-induced in-chip cylinders are illustrated in Fig. S25b. Then, the sample is polished for access to the modified areas and treated with the developed chemical ethant (Fig. S25c). The array of vias that cut through the entire chip are revealed within a few minutes (Fig. S25d). The length of vias along *z* axis can be controlled with the number of laser pulses.

7.4 Cantilever-like structures

Complex 3D structures can be created directly in Si with scanning. As a further example of 3D control, we demonstrate the creation of cantilever-like structures (Fig. S26). First, the laser is raster-scanned to create subsurface walls (Fig. S26a). Each wall is created with a single scan. The laser-induced in-chip walls are illustrated in (Fig. S26b). Then, the sample is polished and treated with the developed chemical etchant (Fig. S26c). During the chemical etching process, unprocessed areas are removed (Fig. S26d), and the cantilevers are revealed within a few minutes (Fig. S26e).

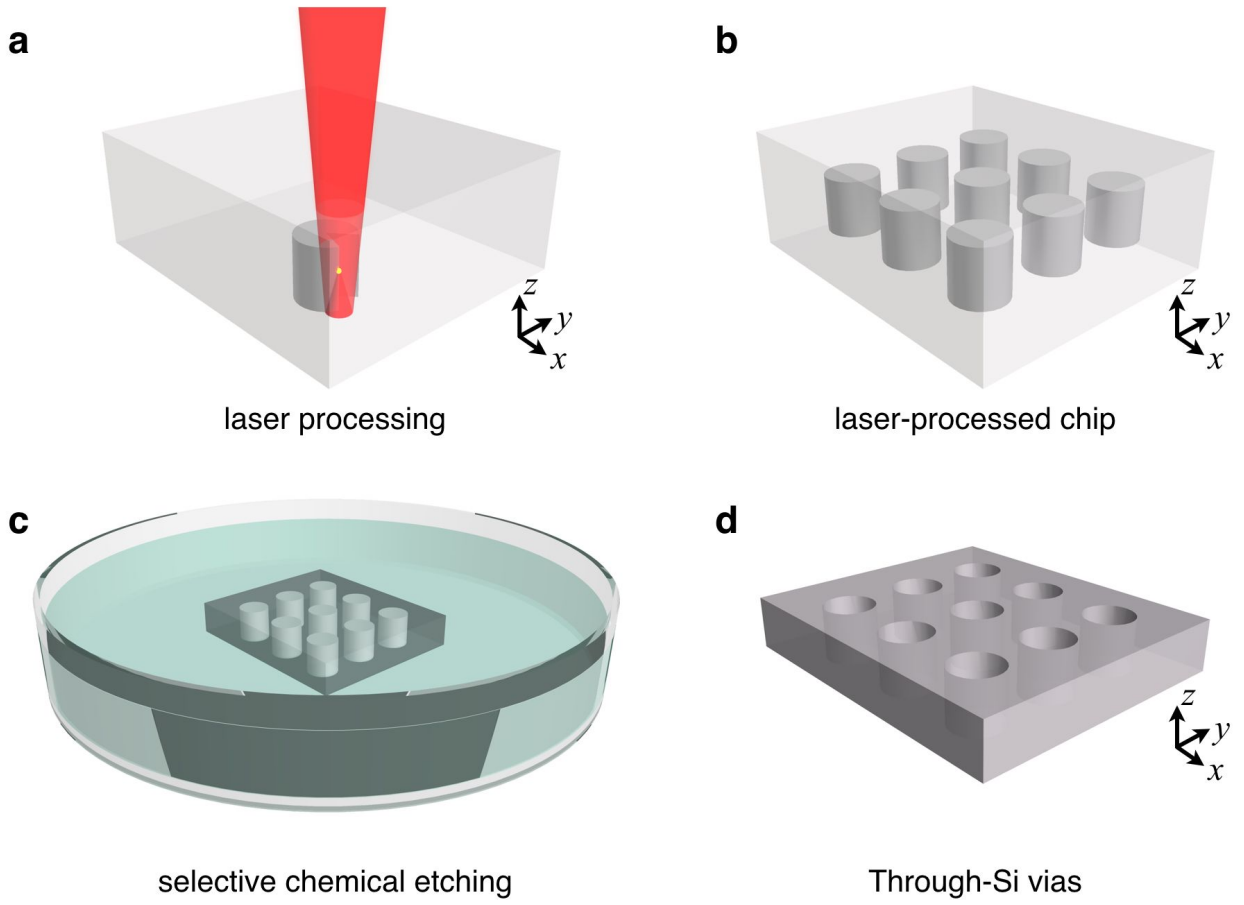

Figure S25: Creation of through-Si vias. **a**, A 3D structure in the form of a cylinder is created in Si with a single circular motion of the sample in the *x*-*y* plane. **b**, An array of cylinders created in Si after laser processing. **c**, The modified areas are exposed with polishing, and Si is treated with the developed chemical. **d**, A uniform array of through-Si vias are created within a few minutes.

7.5 Silicon slicing

Silicon thinning is another potential application, which may find use in photovoltaics. Si-based solar cells are currently leading the market. In these devices, the active region of the wafer is formed by the top few micrometres, while the rest of the wafer is wasted. A potential solution to this loss is to develop technologies for creating crystal thin films[75]. While chemical methods[76] or conventional methods, such as reactive ion etching[77] exist for producing thin films, to our knowledge, there is no laser based method for Si slicing. We demonstrate this capability by slicing a wafer into thin plates (≈ 30 μm). In Fig. S26d, the slicing process is illustrated. The unmodified areas bounded by the laser-induced raster lines effortlessly

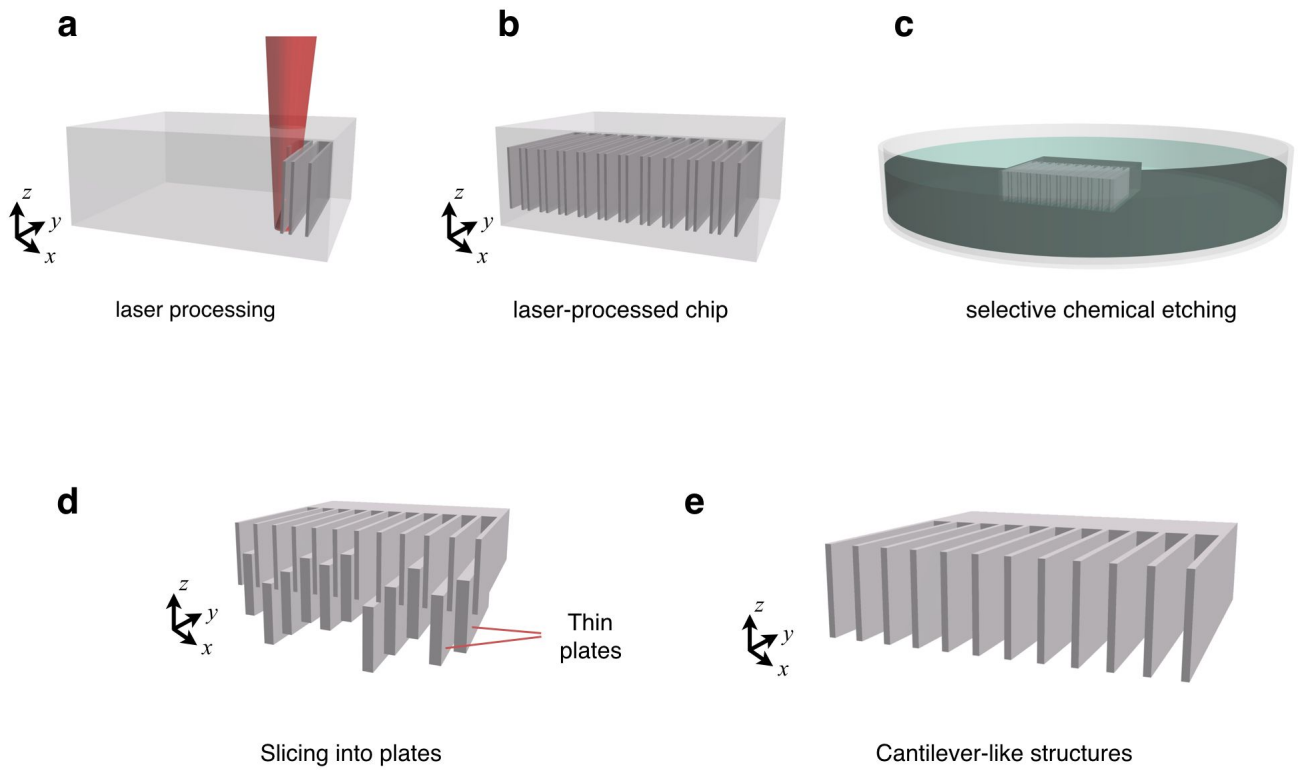

Figure S26: Creation of cantilever-like structures and silicon slicing. **a**, A 3D architecture is created in Si by scanning the laser. **b**, Finished array of subsurface walls after laser processing. **c**, Si is polished, and then treated with the developed chemical. **d**, During etching the unmodified areas located between the subsurface walls effortlessly separate from the chip. **e**, A uniform array of cantilever-like structures are revealed after etching.

separate from the rest of the wafer. In addition, thanks to the nature of the fabrication technique, created geometries are not limited to plates, but other shapes such as curved plates are within possibility.

8. Description of the Laser System

We used a home-built all-fibre master-oscillator power-amplifier (MOPA) system that works at 1.55 μm, producing 5-ns pulses with 60-μJ pulse energy at 150-kHz repetition rate. The laser system is coupled to a high-resolution computer-controlled 3-axis motorised stage (Aerotech, ANT130-XY, ANT95-L-Z) which is used for laser alignment and sample processing. The laser system, which includes a coupled processing station is shown schematically in Fig. S27.

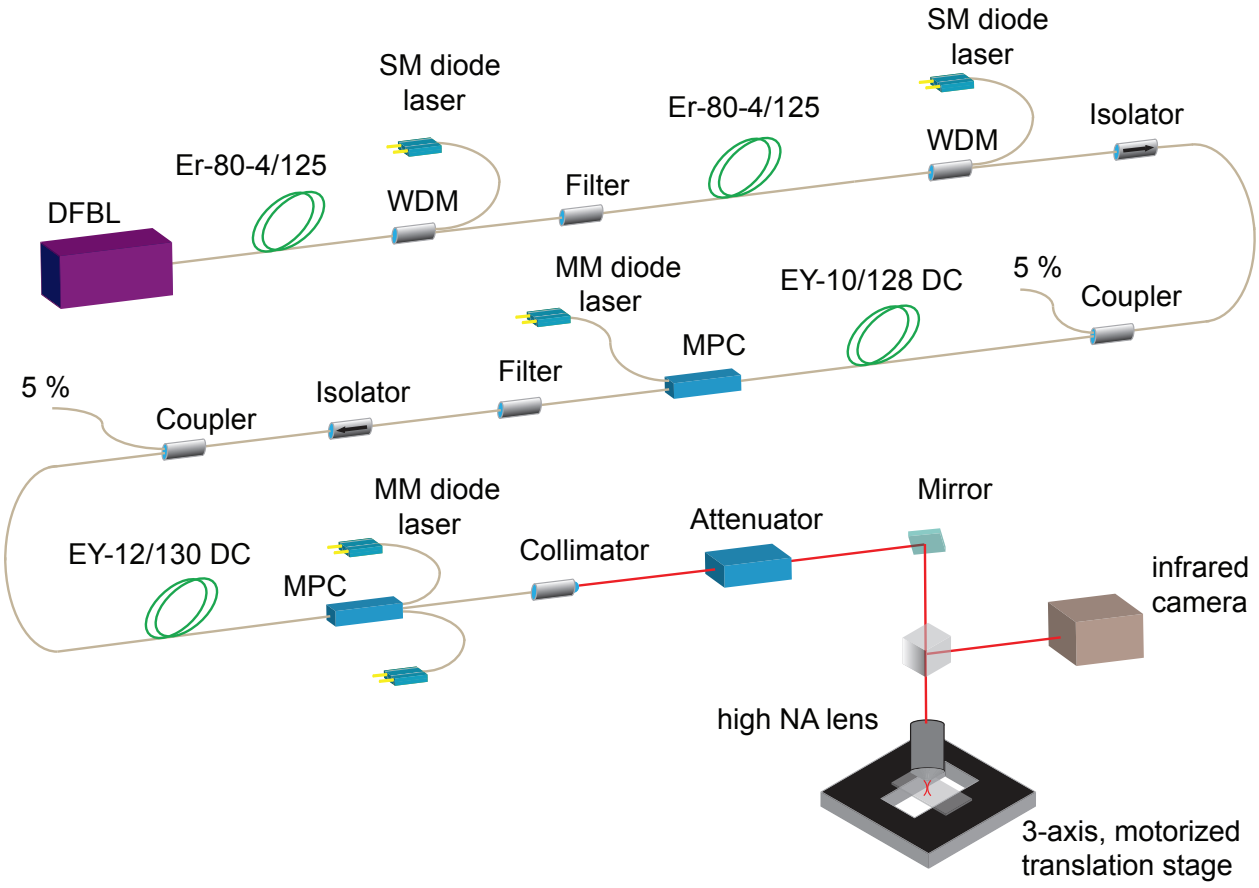

Figure S27: The laser system used for nonlinear laser lithography inside Si. DFBL: Distributed feedback laser, Er: erbium, SM: single mode, DC: double-clad, MM: multimode, MPC: multimode pump combiner.

The seed pulses which will be amplified are produced by a single-frequency distributed feedback semiconductor laser (DFBL, EM4 Inc.) operating at 1550 nm. The modulated DFBL produces 15-ns

pulses at 150-kHz repetition rate, with a peak power ≈ 100 mW, corresponding to 0.2 mW of average power. The system requires several amplification stages due to the low repetition rate and low peak power. The multi-stage amplification scheme gradually increases the average power to fully saturate each stage and thus prevents amplified spontaneous emission (ASE). The first two preamplifier stages use Er-doped fibres (Er-80 4/125). These gain fibres are backward-pumped through 1550 nm / 980 nm wavelength division multiplexers.

The last two amplifier stages are based on double-clad Er-Yb co-doped fibres (EY-10/128 DCF and EY-12/130 DCF). These two stages are clad pumped by multimode high-power pump diodes by multimode pump combiners. After the final amplifier stage, the system produces up to 9 W of average power which corresponds to a pulse energy of 60 μJ. The output pulses are shortened to 5.5 ns due to pulse steepening during amplification. At the end of the last amplifier stage, a single-mode fibre is used to collect the system output. An attenuator allows for polarisation and power control. Finally, the laser light is directed to the sample, where alignment is done with the help of an IR camera.

For single-pulse or low-repetition-rate (< 100 kHz) experiments, pulsed-pumping scheme was used. For this purpose, the currents of the first two preamplifiers were reduced to 30% - 50% of their nominal values, and the last two amplifier stages were pulsed pumped. The durations and intensities of the pump pulses, and the delays between the seed and pump pulses were optimised to minimise the ASE content. In single mode operation, we could obtain more than 30 μJ of pulse energy with less than 50% of ASE. With these settings, high-aspect-ratio subsurface structures were observed between 2 kHz and 300 kHz repetition rates.

9. Bibliography

[1] Nejadmalayeri, A. H. *et al.* Inscription of optical waveguides in crystalline silicon by mid-infrared femtosecond laser pulses. *Optics Letters* **30**, 964–966 (2005).

[2] Leyder, S. *et al.* Multiphoton absorption of 1.3 μm wavelength femtosecond laser pulses focused inside Si and SiO 2. In Dreischuh, T. N. & Daskalova, A. T. (eds.) *Seventeenth International School on Quantum Electronics: Laser Physics and Applications*, 877004–8 (SPIE, 2013).

[3] Mouskeftaras, A. *et al.* Self-limited underdense microplasmas in bulk silicon induced by ultrashort laser pulses. *Applied Physics Letters* **105** (2014).

[4] Kononenko, V. V., Konov, V. V. & Dianov, E. M. Delocalization of femtosecond radiation in silicon. *Optics Letters* **37**, 3369–3371 (2012).

[5] Grojo, D., Mouskeftaras, A., Delaporte, P. & Lei, S. Limitations to laser machining of silicon using femtosecond micro-Bessel beams in the infrared. *Journal of Applied Physics* **117** (2015).

[6] Mori, M. *et al.* Tailoring thermoelectric properties of nanostructured crystal silicon fabricated by infrared femtosecond laser direct writing. *Physica Status Solidi (a)* **212**, 715–721 (2015).

[7] Ito, Y. *et al.* Modification and machining on back surface of a silicon substrate by femtosecond laser pulses at 1552 nm. *Journal of Laser Micro Nanoengineering* **9**, 98–102 (2014).

[8] Pavlov, I., Dulgergil, E., Ilbey, E. & Ilday, F. Ö. 10 W, 10 ns, 50 kHz all-fiber laser at 1.55 µm. *Conference on Lasers and Electro-Optics 2012 (2012), paper CTu2M.5* (2012).

[9] Tokel, O. *et al.* Laser-writing in silicon for 3D information processing. *arXiv.org* (2014). `1409.2827v1`.

[10] Verburg, P. C., Römer, G. R. B. E. & Huis in t Veld, A. J. Two-photon–induced internal modification of silicon by erbium-doped fiber laser. *Optics Express* **22**, 21958–21971 (2014).

[11] Bristow, A. D., Rotenberg, N. & van Driel, H. M. Two-photon absorption and kerr coefficients of silicon for 850-2200nm. *Applied Physics Letters* **90**, 191104 (2007).

[12] Chambonneau, M., Li, Q., Chanal, M., Sanner, N. & Grojo, D. Writing waveguides inside monolithic crystalline silicon with nanosecond laser pulses. *Optics Letters* **41**, 4875–4878 (2016).

[13] Li, Q., Chambonneau, M., Chanal, M. & Grojo, D. Quantitative-phase microscopy of nanosecond laser-induced micro-modifications inside silicon. *Applied Optics* **55**, 9577–9583 (2016).

[14] Topol, A. W. *et al.* Three-dimensional integrated circuits. *IBM Journal of Research and Development* **50**, 491–506 (2006).

[15] Emma, P. G. & Kursun, E. Is 3d chip technology the next growth engine for performance improvement? *IBM Journal of Research and Development* **52**, 541–552 (2008).

[16] Motoyoshi, M. Through-silicon via (tsv). *Proceedings of the IEEE* **97**, 43–48 (2009).

[17] Tsibidis, G. D., Barberoglou, M., Loukakos, P. A., Stratakis, E. & Fotakis, C. Dynamics of ripple formation on silicon surfaces by ultrashort laser pulses in subablation conditions. *Physical Review B* **86**, 115316 (2012).

[18] Liu, P., Jiang, L., Hu, J., Han, W. & Lu, Y. Direct writing anisotropy on crystalline silicon surface by linearly polarized femtosecond laser. *Optics Letters* **38**, 1969–1971 (2013).

[19] Öktem, B. *et al.* Nonlinear laser lithography for indefinitely large-area nanostructuring with femtosecond pulses. *Nature Photonics* **7**, 897–901 (2013).

[20] Nilsson, N. G. Band-to-band auger recombination in silicon and germanium. *Physica Scripta* **8**, 165 (1973).

[21] Dziewior, J. & Schmid, W. Auger coefficients for highly doped and highly excited silicon. *Applied Physics Letters* **31**, 346–348 (1977).

[22] van Driel, H. M. Kinetics of high-density plasmas generated in si by 1.06- and 0.53- μm picosecond laser pulses. *Phys. Rev. B* **35**, 8166–8176 (1987).

[23] Wood, R. F. & Giles, G. E. Macroscopic theory of pulsed-laser annealing. i. thermal transport and melting. *Phys. Rev. B* **23**, 2923–2942 (1981).

[24] Shanks, H. R., Maycock, P. D., Sidles, P. H. & Danielson, G. C. Thermal conductivity of silicon from 300 to 1400°k. *Phys. Rev.* **130**, 1743–1748 (1963).

[25] Downer, M. C. & Shank, C. V. Ultrafast heating of silicon on sapphire by femtosecond optical pulses. *Phys. Rev. Lett.* **56**, 761–764 (1986).

[26] Soref, R. & Bennett, B. Electrooptical effects in silicon. *IEEE Journal of Quantum Electronics* **23**, 123–129 (1987).

[27] Timans, P. J. Emissivity of silicon at elevated temperatures. *Journal of Applied Physics* **74**, 6353–6364 (1993).

[28] Bekenstein, R., Schley, R., Mutzafi, M., Rotschild, C. & Segev, M. Optical simulations of gravitational effects in the Newton-Schrodinger system. *Nature Physics* **11**, 872–878 (2015).

[29] Cocorullo, G. & Rendina, I. Thermo-optical modulation at 1.5 μm in silicon etalon. *Electronics Letters* **28**, 83–85 (1992).

[30] Watts, M. R. *et al.* Adiabatic thermo-optic mach-zehnder switch. *Optics Letters* **38**, 733–735 (2013).

[31] Brodeur, A. *et al.* Moving focus in the propagation of ultrashort laser pulses in air. *Optics Letters* **22**, 304–306 (1997).

[32] Agrawal, G. *Nonlinear Fiber Optics* (Academic Press, Boston, 2013).

[33] Couarion, A., & Mysyrowicz, A. Femtosecond filamentation in transparent media. *Physics Reports* **441**, 47–189 (2007).

[34] Verburg, P. C. *et al.* Crystal structure of laser-induced subsurface modifications in Si. *Applied Physics A: Materials Science and Processing* **120**, 683–691 (2015).

[35] Domnich, V. & Gogotsi, Y. Phase transformations in silicon under contact loading. *Reviews on Advanced Materials Science* **3**, 1–36 (2002).

[36] Leuthold, J., Koos, C. & Freude, W. Nonlinear silicon photonics. *Nature Photonics* **4**, 535–544 (2010).

[37] Verdeyen, J. *Laser Electronics (2nd Edition)* (Prentice-Hall International Edition, 1989).

[38] Saleh, E. A., Bahaa & Teich, M. C. *Fundamentals of Photonics*. Wiley Series in Pure and Applied Optics (Wiley, 2007).

[39] Beresna, M., Gecevičius, M. & Kazansky, P. G. Ultrafast laser direct writing and nanostructuring in transparent materials. *Advances in Optics and Photonics* **6**, 293–339 (2014).

[40] Goodman, J. W. *Introduction to Fourier Optics* (Roberts and Company, 2005).

[41] Benton, S. A. & Michael Bove, J. V. *Holographic Imaging* (John Wiley and Sons, 2008).

[42] Soref, R. Mid-infrared photonics in silicon and germanium. *Nature Photonics* **4**, 495–497 (2010).

[43] Brown, B. & Lohmann, A. Complex spatial filtering with binary masks. *Applied Optics* **5**, 967–969 (1966).

[44] Rosen, J. & Brooker, G. Non-scanning motionless fluorescence three-dimensional holographic microscopy. *Nature Photonics* **2**, 190–195 (2008).

[45] Grier, D. G. A revolution in optical manipulation. *Nature* **424**, 810–816 (2003).

[46] Kildishev, A. V., Boltasseva, A. & Shalaev, V. M. Planar photonics with metasurfaces. *Science* **339**, 1232009 (2013).

[47] Fujita, J. *et al.* Manipulation of an atomic beam by a computer-generated hologram. *Nature* **380**, 691–694 (1996).

[48] Pugliese, L. & Morris, G. M. Computer-generated holography in photorefractive materials. *Optics Letters* **15**, 338–340 (1990).

[49] Jolly, S., Smalley, D. E., Barabas, J. & Bove, V. M., Jr. Direct fringe writing architecture for photorefractive polymer-based holographic displays: analysis and implementation. *Optical Engineering* **52**, 055801–055801 (2013).

[50] Reichelt, S. *et al.* Full-range, complex spatial light modulator for real-time holography. *Optics Letters* **37**, 1955–1957 (2012).

[51] Makey, G., El-Daher, M. S. & Al-Shufi, K. Utilization of a liquid crystal spatial light modulator in a gray scale detour phase method for fourier holograms. *Applied Optics* **51**, 7877–7882 (2012).

[52] Li, X. *et al.* Athermally photoreduced graphene oxides for three-dimensional holographic images. *Nature Communicatons* **6** (2015).

[53] Larouche, S., Tsai, Y.-J., Tyler, T., Jokerst, N. M. & Smith, D. R. Infrared metamaterial phase holograms. *Nature Materials* **11**, 450–454 (2012).

[54] Huang, L. *et al.* Three-dimensional optical holography using a plasmonic metasurface. *Nature Communications* (2013).

[55] Meinzer, N., Barnes, W. L. & Hooper, I. R. Plasmonic meta-atoms and metasurfaces. *Nature Photonics* **8**, 889–898 (2014).

[56] Li, Y., Dou, Y., An, R., Yang, H. & Gong, A. Q. Permanent computer-generated holograms embedded in silica glass by femtosecond laser pulses. *Optics Express* **13**, 2433–2438 (2005).

[57] Guo, Z., Qu, S. & Liu, S. Generating optical vortex with computer-generated hologram fabricated inside glass by femtosecond laser pulses. *Optics Communications* **273**, 286–289 (2007).

[58] Heanue, J. F., Bashaw, M. C. & Hesselink, L. Volume holographic storage and retrieval of digital data. *Science* **265**, 749–752 (1994).

[59] Blanche, P.-A. *et al.* Holographic three-dimensional telepresence using large-area photorefractive polymer. *Nature* **468**, 80–83 (2010).

[60] Dufresne, E., Spalding, G., Dearing, M., Sheets, S. & Grier, D. Computer generated holographic optical tweezer arrays. *Review of Scientific Instruments* **72**, 1810–1816 (2001).

[61] Gerchberg, R. W. & Saxton, W. O. A practical algorithm for the determination of phase from image diffraction plane pictures. *Optik* **35**, 237–246 (1972).

[62] Fienup, J. R. Phase retrieval algorithms: a personal tour. *Applied Optics* **52**, 45–56 (2013).

[63] Soifer, V. K. & Doskolovich, L. *Iterative Methods for Diffractive Optical Elements Computation* (Taylor and Francis, 1997).

[64] Loewen, E. G. & Popov, E. *Diffraction Gratings and Applications* (Marcel Dekker Inc., 1997).

[65] Huang, Z., Geyer, N., Werner, P., de Boor, J. & Gosele, U. Metal-assisted chemical etching of silicon: a review. *Advanced Materials* **23**, 285–308 (2011).

[66] Schwartz, B. & Robbins, H. Chemical etching of silicon. *Journal of the Electrochemical Society* **123**, 1903–1909 (1976).

[67] Turner, D. On the mechanism of chemically etching germanium and silicon. *Journal of the Electrochemical Society* **107**, 810–816 (1960).

[68] Shih, S. *et al.* Photoluminescence and formation mechanism of chemically etched silicon. *Applied Physics Letters* **60**, 1863–1865 (1992).

[69] Abbadie, A., Hartmann, J.-M. & Brunier, F. A review of different and promising defect etching techniques: From si to ge. *ECS Transactions* **10**, 3–19 (2007).

[70] Chandler, T. Memc etch- a chromium trioxide-free etchant for delineating dislocations and slip in silicon. *Journal of The Electrochemical Society* **137**, 944–948 (1990).

[71] Kooij, E., Butter, K. & Kelly, J. Silicon etching in hno 3/hf solution: charge balance for the oxidation reaction. *Electrochemical and solid-state letters* **2**, 178–180 (1999).

[72] Wang, Y. *et al.* Maskless inverted pyramid texturization of silicon. *Scientific Reports* **5**, 10843 (2015).

[73] Renard, V. T. *et al.* Catalyst preparation for cmos-compatible silicon nanowire synthesis. *Nature Nanotechnology* **4**, 654v657 (2009).

[74] Lu, Y.-T. & Barron, A. R. Anti-reflection layers fabricated by a one-step copper-assisted chemical etching with inverted pyramidal structures intermediate between texturing and nanopore-type black silicon. *Journal of Materials Chemistry A* **2**, 12043–12052 (2014).

[75] Shah, A., Torres, P., Tscharner, R., Wyrsch, N. & Keppner, H. Photovoltaic technology: the case for thin-film solar cells. *Science* **285**, 692–698 (1999).

[76] Tong, Q.-Y. & Goesele, U. *Semiconductor wafer bonding: science and technology* (John Wiley, 1999).

[77] Mizushima, I., Sato, T., Taniguchi, S. & Tsunashima, Y. Empty-space-in-silicon technique for fabricating a silicon-on-nothing structure. *Applied Physics Letters* **77**, 3290–3292 (2000).