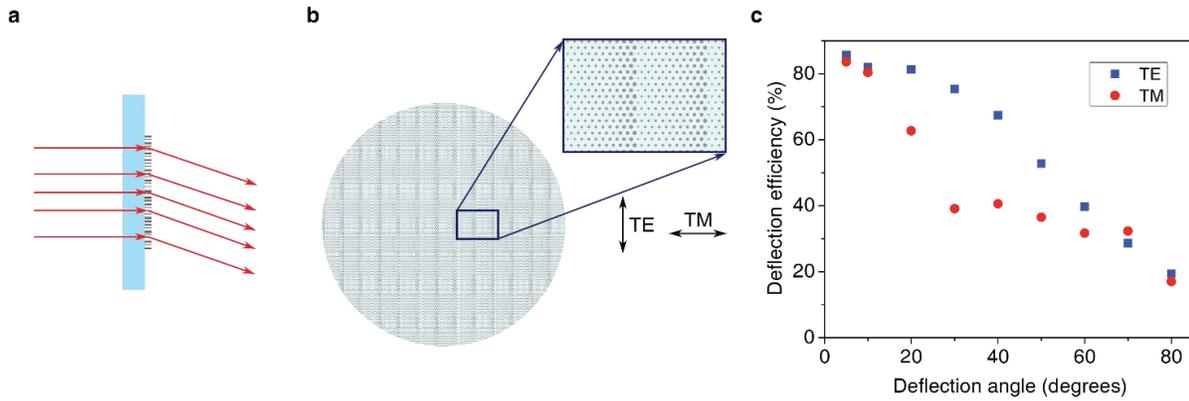


Supplementary Figure 1 | Optimum transmissive mask design for shaping an incident light to a desired tangential form. (a) The light from the sources and scatterers in the half space (1) passes through the transmission mask $t(x, y)$, and its tangential electric field on the target plane is represented by \mathbf{E}_{tan} . (b) The equivalent magnetic surface current density \mathbf{M}_s emits the same tangential electric field \mathbf{E}_{tan} on the target plane. (c) The magnetic field $\mathbf{H}_d^{(2)}$ is emitted by an electric surface current density \mathbf{J}_s which is located on the target plane.



Supplementary Figure 2 | Measurement results of HCTA beam deflectors. (a) Schematic illustration of an ideal beam deflector. (b) The HCTA pattern that implements a uniform phase ramp along the horizontal direction. The polarization direction for the TE and TM polarizations are also shown. (c) Measured deflection efficiency of a set of beam deflectors for the transverse electric and magnetic polarized incident beam as a function of deflection angle.

Supplementary Note 1: Optimum transmissive mask design

We consider the general case of using a transmissive mask to modify the optical wavefront emitted by given sources to a desired form. The light is generated by sources located in the half space (1) as shown schematically in Supplementary Fig. 1a. The tangential component of the electric field of the incident light just before passing through the transmissive mask is represented by $\mathbf{E}_i^{(1)}$. The incident field might be, for example, a diverging beam from a semiconductor laser or a collimated Gaussian beam. The desired output wavefront can be chosen arbitrarily (examples include a beam that is matched to a mode of an optical fiber, a Bessel beam, or a tightly focused beam). Since the propagation is governed by the Maxwell's equations, the desired output beam is fully described by the tangential components of its electric field on a target plane parallel to the transmissive mask. We represent this desired tangential component by \mathbf{E}_d (as shown in Supplementary Fig. 1a).

The output beam formed by the transmissive mask is in general different from the desired beam, and its tangential electric field on the target plane \mathbf{E}_{tan} is an approximation for the desired tangential electric field. Our main objective is to determine the transmissive mask $t(x, y)$ such that \mathbf{E}_{tan} is the best possible approximation for \mathbf{E}_d . A useful measure for quantifying the accuracy of the approximation is the norm of the projection integral defined as

$$|\langle \mathbf{E}_{\text{tan}}, \mathbf{E}_d \rangle| = \left| \int \mathbf{E}_{\text{tan}} \cdot \mathbf{E}_d^* ds \right|, \quad (1)$$

where $*$ represents the complex conjugate operation; and the surface integral is evaluated over the target plane. For the best approximation, this norm should be maximized.

To relate the projection integral to the transmission mask we use the equivalence principle and the reciprocity theorem. According to the definition of a transmissive mask, the tangential electric field of the light just after it passes through the transmissive mask is given by $t\mathbf{E}_i^{(1)}$. Using the equivalence principle, a magnetic surface current density $\mathbf{M}_s = 2t\mathbf{E}_i^{(1)} \times \hat{\mathbf{z}}$ located at the output plane of the transmissive mask and emitting in vacuum, will generate the same beam in the region (1) as the original sources and the transmissive mask (see Supplementary Fig. 1b). Next, we consider an electric surface current density with $\mathbf{J}_s = \mathbf{E}_d^*$ on the target plane which is emitting in

vacuum. We denote the magnetic field emitted by \mathbf{J}_s at the output plane of the transmissive mask by $\mathbf{H}_d^{(2)}$. Using the reciprocity theorem [1], we can write

$$\int \mathbf{E}_{\text{tan}} \cdot \mathbf{J}_s \, ds = \int \mathbf{H}_d^{(2)} \cdot \mathbf{M}_s \, ds. \quad (2)$$

From (1) and (2) we obtain

$$|\langle \mathbf{E}_{\text{tan}}, \mathbf{E}_d \rangle| = 2 \left| \int t \mathbf{H}_d^{(2)} \cdot (\mathbf{E}_i^{(1)} \times \hat{\mathbf{z}}) \, ds \right|, \quad (3)$$

Using (3), we see that the best approximation to a desired output beam is achieved when $|t| = 1$ and

$$\angle t = -\angle \left(\mathbf{H}_d^{(2)} \cdot (\mathbf{E}_i^{(1)} \times \hat{\mathbf{z}}) \right) \quad (4)$$

In other words, the best transmissive mask is a phase mask, and we can determine its phase profile as follows. First, we find the tangential component of the incident light at the location of the transmissive mask ($\mathbf{E}_i^{(1)}$); this can be done either analytically or numerically depending on the type of the excitation. Next, we consider an electric surface current density with the same spatial distribution as the complex conjugate of the desired tangential electric field at the target plane which is emitting in the free space, and we find the magnetic field emitted by this current at the location of the transmission mask ($\mathbf{H}_d^{(2)}$). Since \mathbf{J}_s is planar and is radiating in free space, its fields can be obtained using a simple method such as the plane wave expansion technique [2]. Finally, we obtain the phase profile of the transmission mask using (4). To achieve a lens with tight focus, the desired field should be set to a uniform electric field confined inside a circle with deep subwavelength radius located at the plane of focus.

Supplementary Note 2: Under-sampling of the phase profile

Considering a micro-lens in the geometrical optics picture offers an intuitive understanding of how the under-sampling of the phase profile affects the lens performance. According to the geometrical optics, a ray that is propagating parallel to the lens's optical axis is deflected by the lens toward its focal point. The rays that are propagating farther away from the optical axis are

deflected by larger angles, and the lens's NA represents the sine of the largest deflection angle. Therefore, a lens can be considered as a deflector whose local deflection angle gradually increases from zero at the center of the lens to its maximum (which is given by $\sin^{-1}(\text{NA})$) at the perimeter of the lens. We show that, due to the under sampling of the phase profile by the HCTA lattice, the deflection efficiency of the HCTA deflectors decreases by increasing their deflection angle. The lower deflection efficiency at larger deflection angles leads to the lower focusing efficiency of the micro-lenses with higher NA.

A uniform deflector functions similar to a blazed grating and deflects a monochromatic normally incident light by a fixed angle (as shown in Supplementary Fig. 2a). The uniform deflector has a linearly varying phase profile whose slope is proportional to the sine of its deflection angle. A schematic illustration of an HCTA uniform deflector, with phase profile varying linearly along the horizontal direction, is depicted in Supplementary Fig. 2b. Nine $400\ \mu\text{m}$ diameter uniform deflectors with different deflection angles were fabricated using the same family of periodic HCTAs used for the high NA micro-lenses and the same fabrication process.

The deflectors were illuminated with a linearly polarized collimated laser beam with beam radius of approximately $100\ \mu\text{m}$. The deflected power was measured using a photodetector located 10 cm away from the deflector along the expected deflection direction. The deflection efficiency was obtained by dividing the measured deflected power by the incident power. The measured deflection efficiencies for two linear orthogonal polarizations are depicted in Supplementary Fig. 2c. The directions of polarization for the TE and TM polarized lights are shown in Supplementary Fig. 2b. The TE polarization corresponds to the transverse electric polarized deflected light, while the deflected light is transverse magnetic polarized for the TM polarization.

As Supplementary Fig. 2c shows, the deflection efficiency of an HCTA deflector decreases as its deflection angle increases. The deflection efficiency drop is faster for the TM polarized incident light compared to the TE one. We attribute the efficiency reduction to the under sampling of the phase profile of the deflectors with large deflection angles. The desired phase profile of a deflector with deflection angle of θ is sampled by $n = \lambda/(a \sin(\theta))$ unit cells over 2π phase variation, where a is the lattice constant of the HCTA. For the HCTA used in this study, the lattice constant is roughly equal to a half of a wavelength; therefore, for a 40° deflector, the full phase range of 2π is sampled by approximately three unit cells. Similar diffraction efficiency reduction due to

phase sampling and quantization error is encountered in the design and implementation of Fresnel lenses with a limited number of levels [3]. One approach to increase the efficiency of a high NA micro-lens is to use an HCTA with a smaller lattice constant. This leads to a finer sampling of the desired micro-lens profile. For example, as Fig. 1b in the main manuscript shows, we could use the lattice constant of 650 nm instead of 800 nm and still achieve the full 2π transmission phase range by changing the post diameters.

Supplementary References

- [1] Harrington, R. F. Time Harmonic Electromagnetic Fields (Wiley-IEEE Press, 2001).
- [2] Born, M. & Wolf, E. Principles of Optics (Cambridge University Press, Cambridge, 1999), 7th edn.
- [3] Wyrowski, F. Efficiency of quantized diffractive phase elements. *Optics Communications* **92**, 119–126, (1992).