Supplementary Information

An unsupervised deep learning algorithm for single-site reconstruction in quantum gas microscopes

Alexander Impertro,[1,2,3,*] Julian F. Wienand,[1,2,3] Sophie Häfele,[1,2,3] Hendrik von Raven,[1,2,3] Scott Hubele,[1,2,3] Till Klostermann,[1,2,3] Cesar R. Cabrera,[1,2,3] Immanuel Bloch,[1,2,3] and Monika Aidelsburger[1,2,†]

[1] *Department of Physics, Ludwig-Maximilians-Universität München, Schellingstr. 4, D-80799 Munich, Germany*
[2] *Munich Center for Quantum Science and Technology (MCQST), Schellingstr. 4, D-80799 Munich, Germany*
[3] *Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Strasse 1, D-85748 Garching, Germany*

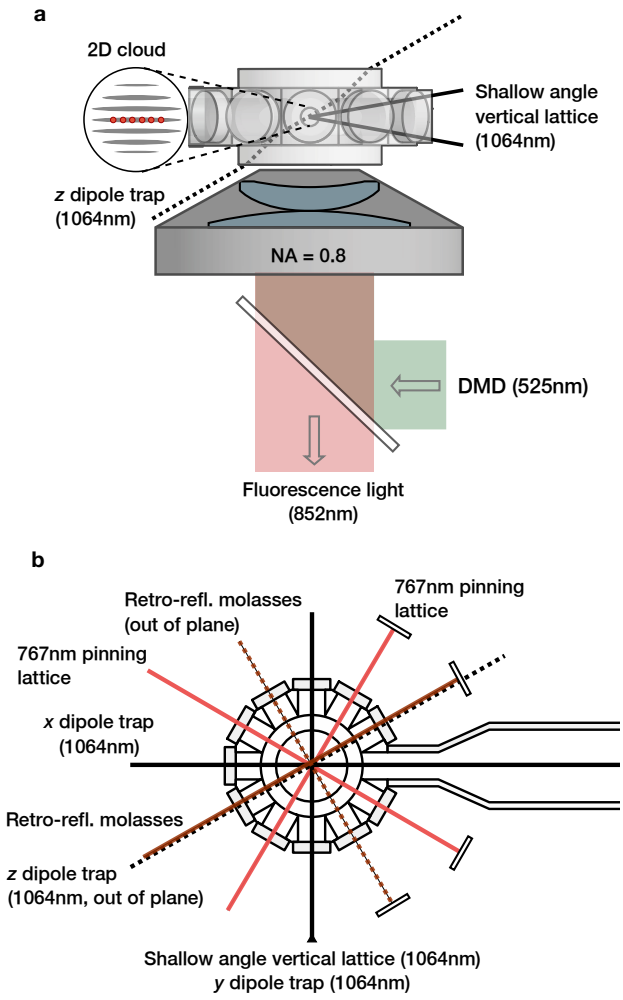
In this supplementary material we present further details about the experimental setup and describe how we generate fluorescence images of homogeneously distributed atoms at different fillings so that they can be used as training data for the neural network (Section 'Experimental details'). We also elaborate on important details of the reconstruction process (Section 'Reconstruction scheme'), including the extraction of the lattice sites in experimental images and other data preprocessing steps, and demonstrate how the autoencoder-based reconstruction performs with simulated fluorescence images (Section 'Evaluation on simulated data'). Finally, we present the mathematical details behind estimating the experimental fidelity from taking two images of the same cloud (Section 'Fidelity estimation from double imaging').

SUPPLEMENTARY NOTE 1: EXPERIMENTAL DETAILS

A detailed discussion of the experimental setup used to generate a Bose-Einstein condensate (BEC) in the science chamber has been introduced in Refs [1–3]. In the following, we focus on the ingredients needed for the preparation of a Mott insulator (MI) and the fluorescence imaging technique.

Setup overview

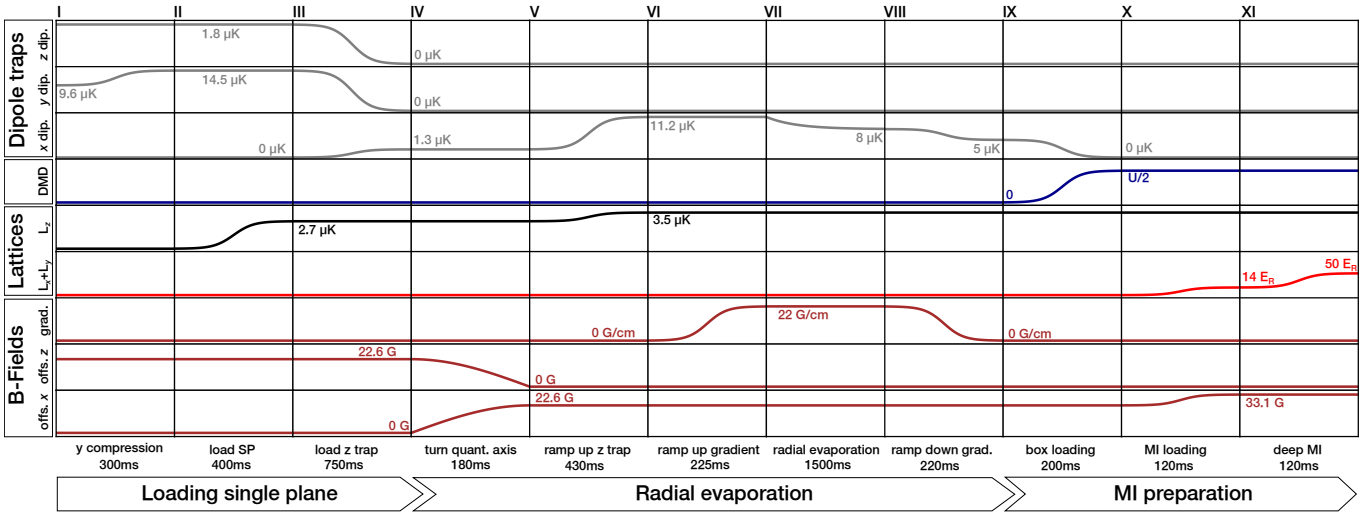
Glass cell.– The science chamber is a glass cell of dodecagon shape with 11 side windows (12 mm diameter) as well as a top and a bottom window (30 mm diameter). Our high-resolution objective (NA = 0.8) is mounted directly underneath the glass cell, optically accessing the atoms through the bottom window (see Supplementary Fig. 1a).



Supplementary Figure 1. **Experimental Setup.** Dipole traps, lattices and optical molasses in the science chamber, **a** side view, **b** top view. Solid lines indicate beam paths in the horizontal plane, dashed lines indicate that the beam path is out-of-plane.

Dipole Traps.– The science chamber is equipped with three dipole traps and three lattices, providing loading and pinning confinement in all spatial directions. As indicated in Supplementary Fig. 1b, the x and y dipole traps (100 µm × 100 µm and 370 µm × 110 µm waist, respec-

* a.impertro@lmu.de
† monika.aidelsburger@physik.uni-muenchen.de



Supplementary Figure 2. **Mott insulator preparation.** Full experimental sequence for loading the BEC into a single plane of the vertical lattice and turning it into a large Mott insulator with high filling. Note, that the horizontal axis is not to scale.

tively) are perpendicular to each other in the horizontal plane, forming a crossed dipole trap at the position of the atoms at the center of the glass cell. The z dipole trap enters the glass cell under a vertical angle of 60° through the top window, closely passing by the objective (Supplementary Fig. 1a). Its elliptical waist (110 µm × 53 µm) is designed to create a (projected) round confinement in the horizontal plane.

Lattices.– The atoms are trapped in a single plane of a vertical lattice with a spacing of 8 µm. It is the result of two 1064 nm beams (170 µm × 35 µm waist) interfering under a small angle of 3.6°. The horizontal lattice confinement is provided by two perpendicular retro-reflected beams (150 µm×45 µm waist) with a wavelength of $\lambda = 767$ nm. All three lattices are used both for physics experiments (e.g. to prepare the Mott insulating initial state) and for pinning during fluorescence imaging. The maximum lattice depth is around 400 µK in the x, y and z-direction.

Optical molasses.– For fluorescence imaging we perform molasses cooling on the D2 line ($\lambda = 852$ nm) using two retro-reflected molasses beams (2 mm diameter) and a repumper. The first beam lies in the horizontal plane and shares an axis with the projected beam path of the (out-of-plane) z dipole trap. The second beam enters the glass cell under a vertical angle of 60° through the top window (similar to the z dipole trap beam but from a direction that is perpendicular to the first molasses beam). The retro-mirrors of the molasses beams are mounted on piezoelectric actuators which are modulated at around 100 Hz in order to wash out the polarization gradient lattice and achieve a more homogeneous fluorescence signal.

Imaging Path.– After passing the objective (25 mm effective focal length), the fluorescence photons are focused by a tube lens (1000 mm focal length, Thorlabs ACT508-1000-B) onto a sCMOS camera (Teledyne Photometrics Kinetix). The magnification of this imaging system is 40. For projecting custom potentials onto the atoms in the horizontal plane, we use a DMD (digital micromirror device, Texas Instruments DLP6500, interface by bbs Bild- und Lichtsysteme GmbH) that reflects light generated by a 5 W multimode laser (Wavespectrum WSLX-525-005-400M-H) running at 525 nm. The DMD imaging path demagnifies the mask displayed on the DMD chip by a factor of 160 and projects it onto an area of about 50 µm × 90 µm in the atom plane. A dichroic mirror is used to separate the light from the DMD going into the objective and the fluorescence light coming from the objective (cf. Supplementary Fig. 1a).

Coils.– The science chamber is surrounded by multiple pairs of coils in close vicinity (not shown in Supplementary Fig. 1) that allow the generation of offset and gradient fields in all spatial directions. Further, the table on which the experiment is mounted is placed inside a coil cage that is used for compensating environmental magnetic field fluctuations at the position of the atoms (cf. subsection 'Deterministic preparation of different fillings').

Mott insulator preparation

In Supplementary Fig. 2 we show the full experimental sequence used for preparing a Mott insulator in a box potential with about 2500 atoms in the $n = 1$-plateau and up to 98 % filling. Starting point is the BEC in the crossed dipole trap, formed by the x and y dipole trap beams [1].

Single plane loading.– As a first step, the BEC is loaded into a single plane of the vertical lattice, making the system two-dimensional. We use the y dipole trap to compress the degenerate gas vertically (I) and, subsequently, ramp up the vertical lattice potential (II). The compres-

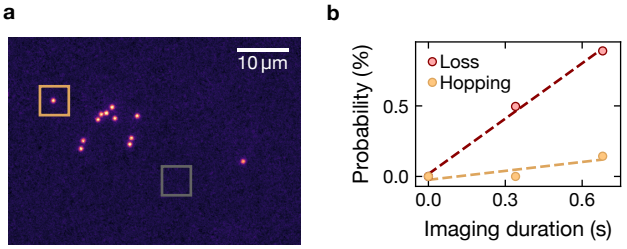
sion allows us to achieve a higher single-plane loading efficiency. We then transfer from the crossed dipole trap into the z dipole trap (III) by ramping down the former and, simultaneously, removing the latter. The z dipole trap provides a more isotropic harmonic confinement and makes it easier to achieve efficient evaporation.

Radial evaporation.– As a next step, we perform in-plane evaporative cooling by applying a gradient in the horizontal plane. We turn the quantization axis in our system (IV) by ramping down the z offset field and, at the same time, ramping up the x offset field, while keeping the total field strength near the three-body-loss-minimum [4, 5]. In preparation for the evaporation process, we increase the harmonic confinement (V) and apply a horizontal gradient force (VI) which, at this point, is not yet strong enough for pulling the atoms out of the single plane in the horizontal direction. We then evaporate by slowly reducing the horizontal confinement provided by the z dipole trap using an exponential ramp profile (VII). Finally, the gradient is adiabatically turned off and the z dipole trap power is reduced to its final value (VIII). During evaporation any remaining population of neighboring planes of the vertical lattice becomes negligible and we end up with a low-temperature sample in a single plane.

Mott insulator loading.– In order to load the $n = 1$ Mott insulator state into a box potential we start by transferring the evaporation-cooled superfluid from the z dipole trap to a blue-detuned box potential provided by the DMD (IX). The box potential has a 50×50 sites large quadratic trapping region, delimited by a 4 µm thick wall whose barrier height linearly drops to zero with increasing distance from the trapping region [the slope is $U/(24\,\mu\text{m})$]. Next, we simultaneously ramp up both horizontal lattices in two consecutive ramps, the first one designed to reach the SF-to-MI transition at $U/J = 16$ (X) and the second one ending in the deep MI regime with $U/J > 50$ (XI), making sure that the SF-to-MI transition is crossed slowly. Together with the first ramp, we also increase the offset field to 33.1 G. Thanks to the broad Feshbach resonance in cesium, this increases the scattering length from $280\,a_0$ to $580\,a_0$ and results in onsite interactions of $U \approx h \times 900$ Hz. The wall height of the box potential is chosen to match $U/2$, such that excess atoms spill out of the box without forming an $n = 2$ shell. Those atoms which have spilled over gather in the surroundings of the box due to the remaining horizontal harmonic confinement of the red-detuned vertical lattice beams (see Fig. 2a in the main text).

Fluorescence imaging

Sequence.– After preparing the MI state, we image the atoms by performing molasses cooling on the D2 line ($\lambda = 852$ nm) and collecting the fluorescence photons using an NA = 0.8 objective and a sCMOS camera. For imaging, all lattices (both horizontal lattices and the vertical lattice) are quenched to their maximum depth of 400 µK. Immediately after pinning, all offset and gradient fields are switched off instantly and the molasses beams are turned on. Their detuning is set to 70 MHz with respect to the $F = 4 \leftrightarrow F' = 5$ transition. Additionally, a repumper, resonant with the $F = 3 \leftrightarrow F' = 4$ transition, is switched on to prevent the atoms from occupying the $F = 3$ ground state. Shortly (10 ms) after beginning the cooling process, we start exposing the sCMOS camera to the fluorescence light for 300 ms.



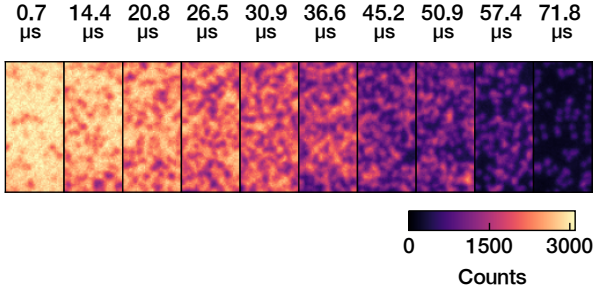
Supplementary Figure 3. **Imaging characterization. a** Fluorescence image of a dilute cloud of atoms used for analyzing the PSF and determining the SNR by comparing the counts in crops with one atom (orange) and without any atom (gray). **b** Loss and hopping probability as a function of imaging time, extracted from multiple images of the same dilute cloud of atoms.

Point spread function (PSF) and signal-to-noise ratio (SNR).– In order to characterize our imaging resolution, we take several hundred images of very dilute clouds of atoms, as shown in Supplementary Fig. 3a. We analyze these images by cutting crops with exactly zero atoms or one atom inside. The crops containing exactly one atom are used to obtain the average experimental PSF shown in Fig. 1c of the main text. A Gaussian fit yields a width that corresponds to a Rayleigh resolution of about 850 nm. From the count statistics of the crops, as shown in Fig. 1d of the main text, we can determine the experimental SNR (as defined in the main text).

Hopping and loss rates.– In order to determine the loss and thermal hopping rates during the imaging process, we image very dilute clouds of atoms (as shown in Supplementary Fig. 3b) three times in a row, with an exposure time of 300 ms per image and a small delay of 40 ms between the exposures, yielding a total imaging time of 1020 ms. In this dilute regime where clustering of atoms is very unlikely, the reconstruction fidelity is assumed to be 1. From comparing the site occupations of the second and third image with that of the first image, we obtain loss and hopping probabilities, as shown in Supplementary Fig. 3b. Using linear fits, we extract a hopping rate of 2.1(1.2) mHz and a loss rate of 13(1) mHz. Note that an atom hopping to a position outside the region of interest (ROI) is also counted as a loss event.

Deterministic preparation of different fillings

For training the neural network and for evaluating its performance as a function of filling, we prepare experimental images with homogeneously distributed atoms at various fillings. We start by preparing a Mott insulator of about 40×40 lattice sites. At the end of the MI preparation sequence and before fluorescence imaging, the atoms are in the $F = 3, m_F = 3$ state. In order to reduce the filling from $n = 1$ in a controlled fashion, we coherently drive the $F = 3, m_F = 3 \leftrightarrow F = 4, m_F = 4$ transition at 9.2 GHz using a resonant microwave pulse of variable duration. All atoms transferred to the $F = 4, m_F = 4$ state are subsequently removed by a blowout pulse resonant with the $F = 4 \leftrightarrow F' = 5$ transition. Supplementary Fig. 4 shows the fluorescence signal from a cropped region inside the MI plateau as a function of the pulse duration. The filling decreases gradually during the first half of the first Rabi oscillation period, as expected. In order to be able to drive the $F = 3, m_F = 3 \leftrightarrow F = 4, m_F = 4$ transition coherently, we rely on very stable magnetic field conditions, as the transition frequency shifts with 2.45 MHz/G. We reduce the environmental magnetic field fluctuations from up to 5 mG to below 100 µG using a commercially available electromagnetic interference (EMI) compensation system (IDE MK5). The magnetic field noise is measured using sensors placed in the direct vicinity of the glass cell and compensated by applying a feedback current to a coil cage surrounding the experiment [3].



Supplementary Figure 4. **Preparation of different fillings.** Cropped fluorescence images of the Mott insulator plateau undergoing Rabi oscillations on the $F = 3, m_F = 3 \leftrightarrow F = 4, m_F = 4$ transition. We image the $F = 3, m_F = 3$ state. The microwave pulse duration is varied to cover the first half of the first oscillation period, yielding a variety of different fillings.

SUPPLEMENTARY NOTE 2: RECONSTRUCTION SCHEME

Lattice extraction

Extraction of lattice vectors.– The real space lattice vectors are extracted from images with low filling fraction. For this, isolated atoms are fitted with a two-dimensional Gaussian in order to determine their position with sub-pixel accuracy. A two-dimensional Fourier transform of these coordinates reveals peaks corresponding to the reciprocal lattice vectors, from which the lattice orientation and spacing are determined (see Ref. [2] for further information).

Image rotation and phase extraction.– The reconstruction scheme based on the autoencoder architecture introduced in the main text (Fig. 1a) requires that the lattice axes are roughly aligned with the image edges due to the square occupation matrix in the bottleneck layer. Since our lattice axes enclose an angle on the order of 30° with the edges of the image, we rotate the full images prior to reconstruction, using only linear interpolation to enable sub-pixel shifts. Additionally, our lattices deviate by $\lesssim 2°$ from a perfect square configuration, which however does not seem to degrade the reconstruction performance noticeably. The lattice vectors in the rotated frame are determined using the aforementioned procedure with rotated dilute images.

In order to fully determine the location of the lattice sites, we also require the lattice phase with respect to the image origin. This phase is obtained in each image individually by fitting the positions of isolated atoms that are always present in the vicinity of the MI.

Data pre-processing

Cropping of lattice segments.– To process fluorescence images of arbitrary size, we need to crop smaller segments containing exactly 16×16 lattice sites to be used as the input for the autoencoder network. Since the lattice spacing is not commensurate with the camera pixel spacing, we first up-sample the images by a factor of 20 without interpolation in order to avoid large discretization errors when cutting. This ensures that the lattice sites inside a crop are consistently located at the same positions relative to the origin. Finally, the cropped images are down-sampled to a size of 256×256 pixels using linear interpolation to fit the input layer shape of the autoencoder, where each lattice sites appears as 16×16 pixels.

Scaling of the training data.– Scaling the input data before processing it using a neural network is essential for successful training. As a general rule, the magnitude of the input values should be on the order of one and symmetrically centered around zero. We re-scale the pixel values z_i of the 16×16 site crops to the range $[-1, 1]$

according to

$$z_i' = 2\frac{z_i - z_{\min}}{z_{\max} - z_{\min}} - 1, \tag{S1}$$

where $z_{\min}$ and $z_{\max}$ are the minimum and maximum pixel values in the entire training set, respectively. The scaling parameters $z_{\min}$ and $z_{\max}$ are determined from the training set and later used for re-scaling all new data subject to the reconstruction process.

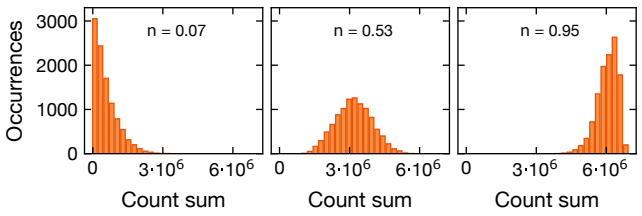
Since non-linear activation functions are used throughout the architecture, the classification result becomes sensitive to the absolute scale of the input data. It is hence recommended to correct for varying counts by, for instance, measuring the counts of isolated atoms or monitoring the shape of the deconvolved count distributions and re-scaling the input data accordingly.

Implementation of the network and training details

We implement the autoencoder architecture in Python using the software packages TensorFlow[6] and Keras[7].

For training and evaluating the network, we record a comprehensive dataset containing around 200 fluorescence images. This is sufficient for generating 10^5 crops that we use for training as well as further data for validation and testing. Given our cycle time of roughly 20 s, this amounts to slightly more than one hour of measurement time.

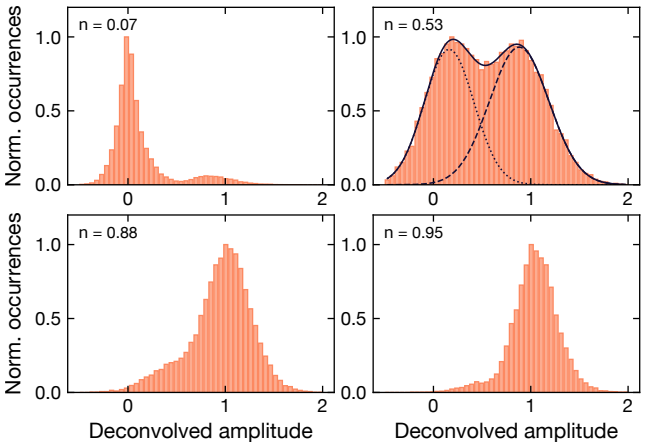
Site counts in unprocessed fluorescence images



Supplementary Figure 5. **Unprocessed count sum distributions.** Distribution of the count sums within each lattice site for images at three exemplary average filling values. There is no discernible separation, making a direct threshold discrimination impossible.

In case of a smaller resolution ratio β, it is possible to directly discriminate empty and occupied lattice sites by means of a simple threshold reconstruction. In this method, the pixel sums of the lattice sites exhibit a bimodal distribution which allows to perform a threshold discrimination. As the ratio β is increased, the spillover of a PSF onto the neighboring sites causes the bimodal distribution to wash out, rendering a threshold discrimination impossible. For $\beta = 2.2$ in our experiment, the pixel sum distributions (examples in Supplementary Fig. 5) do not exhibit a discernible separation for any filling, highlighting the need for a powerful deconvolution algorithm.

Comparison to a deconvolution with the point spread function



Supplementary Figure 6. **Distribution of deconvolved PSF amplitudes.** The amplitudes are obtained from deconvolving experimental fluorescence images at four different average fillings with the measured PSF. The same data as in Fig. 4 of the main text was analyzed. In the panel for half filling, the solid line is a bi-modal Gaussian fit, capturing the individual distributions of empty (dotted line, left peak) and occupied (dashed line, right peak) lattice sites, respectively. The occurrences are normalized to the respective maximum in each plot, and the average filling of each bin was determined using the autoencoder.

To highlight the advantages of our method, we apply one common deconvolution technique to our dataset and discuss the deconvolution result in light of the reconstruction presented in the main text. Specifically, we deconvolve a fluorescence image with the experimentally measured PSF by fitting subsections of the image to an array of PSFs located at pre-determined lattice sites $\mathbf{x}_n$ [8]. A trial image is computed as

$$I_{\text{trial}}(\mathbf{x}, \alpha) = \sum_n \alpha_n \text{PSF}\left(\mathbf{x} - \mathbf{x}_n\right), \tag{S2}$$

where $\mathbf{x}$ is the position in the image, and α_n denotes the deconvolved amplitude of a PSF at site n. During the fitting procedure, the α_n are optimized with the goal to minimize the squared difference to the input fluorescence image. After successful deconvolution, the α_n should follow a bi-modal distribution, and the occupation of the lattice sites can be determined by threshold discrimination.

We perform this deconvolution using our experimentally measured PSF (Fig. 1c of the main text) on subsections of 20×20 lattice sites and use the stochastic optimizer ADAM [9] with an initial learning rate of 0.5 and a fixed number of 200 iterations. The resulting distributions of deconvolved PSF amplitudes for four different average fillings are shown in Supplementary Fig. 6, where the raw data and the respective average filling is the same as analyzed with the autoencoder in Fig. 4 of the main text. While the simple PSF deconvolution is able to create a noticeable separation for the lowest filling, the separation quickly drops towards higher fillings and vanishes completely for $n \gtrsim 0.5$. In analogy to the autoencoder reconstruction, we fit two Gaussians to the deconvolved amplitudes at half filling in order to estimate the reconstruction fidelity from the overlap area. For the PSF deconvolution, we obtain an estimated fidelity of $\mathcal{F} \sim 90\%$. Comparing this estimate as well as the separation between the distributions in Supplementary Fig. 6 to the autoencoder analysis (Fig. 4 of the main text), it is evident that the autoencoder approach significantly outperforms a simple PSF deconvolution at all filling fractions.

Visualization of the encoder network

Convolutional operations.– The autoencoder architecture introduced in the main text (Fig. 1a of the main text) is constructed from several convolutional layers. The basic working principle is illustrated in Supplementary Fig. 7a. An input image is traversed by a kernel, advancing by a given step width (stride). At each position, an inner product between the kernel and the input patch is computed, yielding one entry of the output matrix. A stride of one (together with appropriate padding) retains the shape of the input image, while a stride larger than one reduces the output shape. There are usually several kernels in each convolutional layer, resulting in a set of outputs.

Visualization of the learned kernels.– Since the kernels encode the transformation learned by the network, a visualization of the kernels as well as the outputs of each layer can give further insight into the inner workings of the neural network. However, an interpretation of kernels in deeper layers is not straightforward as they are high-dimensional and operate on abstract features that have been pre-processed by previous layers. The development of methods to interpret neural networks is still an active field of research [10].

We investigate the transformation applied in the encoder by tracing the evolution of an experimental crop through the network (Supplementary Fig. 7b). We show a selection of kernels and outputs of the individual convolutional layers (Supplementary Fig. 7c). Since the first layer operates directly on the input image, a visualization of the kernels is straightforward. Here, we observe smooth structures in all eight learned kernels. The eight corresponding outputs show that no significant transformation happens at this stage yet, and the majority of kernels for this specific input image are suppressed. In the following three layers, we observe how a combination of both large scale and increasingly small-scale features is extracted. The last step before arriving at the deconvolved counts in the bottleneck layer (rightmost image in Supplementary Fig. 7b) is a convolution with one 3D kernel at a depth of 64, corresponding to one 2D kernel for each output of the previous layer. Interestingly, these kernels exhibit structures where one central entry has a strong weight and the nearest neighbors are oppositely weighted. We interpret this shape to serve the purpose of reverting the spillover of counts from atoms on neighboring sites, which is a major factor complicating reconstruction at large resolution-to-spacing ratios.

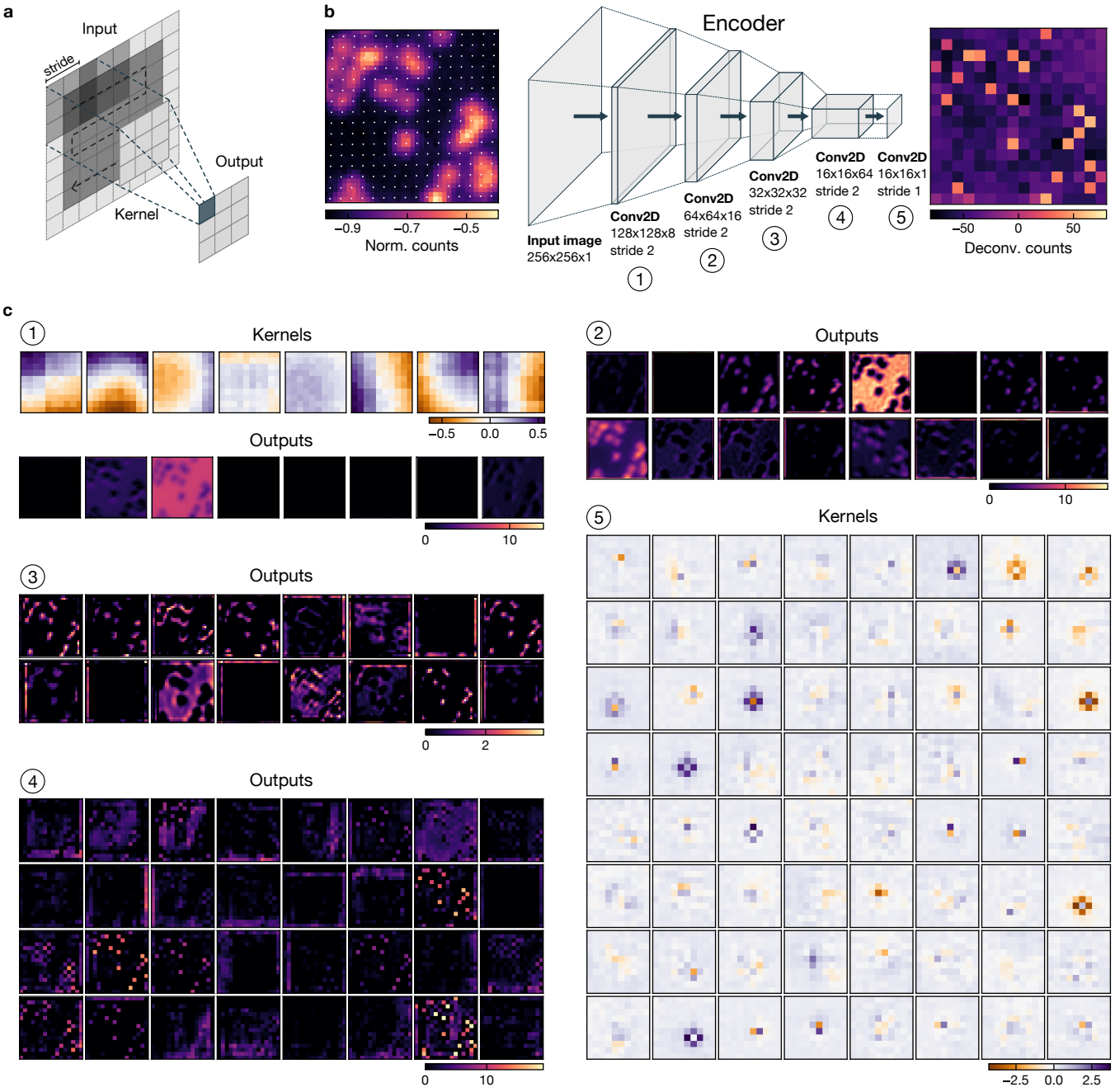
SUPPLEMENTARY NOTE 3: EVALUATION ON SIMULATED DATA

A central feature of the autoencoder architecture is the possibility of unsupervised training directly with experimental data, independent of a possibly imperfect simulation. Nevertheless, an evaluation of simulated fluorescence images can provide useful insights into the reconstruction algorithm due to the available knowledge about the actual occupation. Here, we conduct a basic fluorescence image simulation that reproduces our resolution, lattice spacing and the measured signal-to-noise ratio, but neglects further effects such as superradiance or losses during imaging. This is sufficient to verify the working principle qualitatively, but not accurate enough to be applicable to the experiment on a quantitative level. Specifically, we use the extracted lattice vectors and the measured PSF (Fig. 1c of the main text), and we choose the photon counts per atom (c_{atom}, σ_{atom}) and background noise (σ_{bg}) to match the measured SNR (Fig. 1d of the main text).

Simulation procedure.– To create a simulated fluorescence image, we proceed as follows: (1) A grid of site coordinates is spanned according to the measured lattice vectors. (2) For each occupied lattice site, we draw a number of scattered photons from a normal distribution with mean and variance (c_{atom}, σ_{atom}). (3) The positions of the scattered photons are sampled according to the experimentally measured PSF, capturing asymmetries and long-range aberrations. (4) Photons are then binned according to the camera pixel grid and (5) the calibrated background noise level σ_{bg} is added per pixel.

Before training or reconstruction, simulated images are pre-processed as described in section 'Data preprocessing' in the exact same way as experimental images.

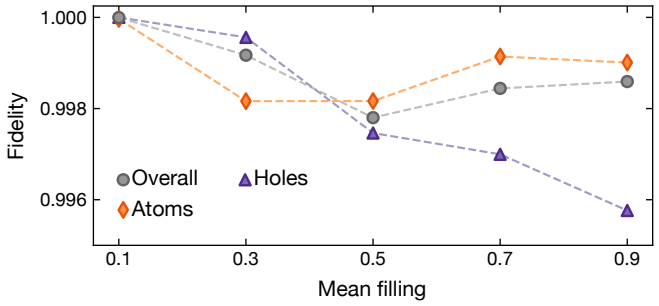
Architecture and training.– We use the simulated fluorescence images to train the network as described in the main text and using the composite loss function [Eq. (1) of the main text]. Similar to the case of experimental



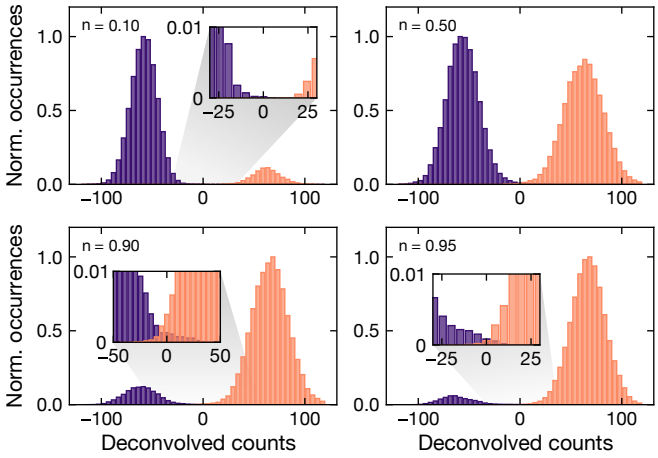
Supplementary Figure 7. **Visualization of the convolutional operations in the encoder network. a** Schematic illustration of a convolution operation. A kernel with entries learned during training is moved across the input data, advancing at a fixed step width (stride). At each position, an inner product between the kernel and the input patch is computed. Shown here for a stride of two. **b** Visualization of the transformation applied by the encoder network for an experimental input image. **c** Selection of kernels and outputs of the different encoder layers. Note that we only show one half of the outputs for layers three and four due to the large number of kernels.

data it is essential that the training dataset contains images at all fillings. For training, we use 70,000 simulated lattice sections.

Fidelity.– The fluorescence simulation enables a straightforward quantitative evaluation of the reconstruction fidelity by comparing the network prediction to the true occupation labels. As in the experimental case, this is done using a previously unseen test set. We define the overall reconstruction fidelity as the ratio of correctly classified sites to the



Supplementary Figure 8. **Fidelity on simulated data as a function of the mean filling.** The fidelity has been evaluated on a test dataset not seen during training with about 1,000 crops per filling bin. The overall fidelity is the fraction of correctly classified sites, while the individual (hole) atom fidelities are the fractions of correctly classified (holes) atoms. Dashed lines are a guide to the eye.



Supplementary Figure 9. **Deconvolved site count distributions for different mean fillings.** Based on the true occupation labels, purple and orange mark empty and occupied lattice sites, respectively. Across all fillings, one finds a bimodal distribution with negligible overlap.

total number of sites $\mathcal{F} = N_{\text{correct}}/N_{\text{tot}}$. Additionally, it is interesting to separately evaluate the atom and hole fidelity to gain information on systematic errors in the reconstruction process, where we define $\mathcal{F}_{\text{holes}} = (\#\,\text{correctly predicted holes})/(\#\,\text{true holes})$ and accordingly for atoms (occupied sites).

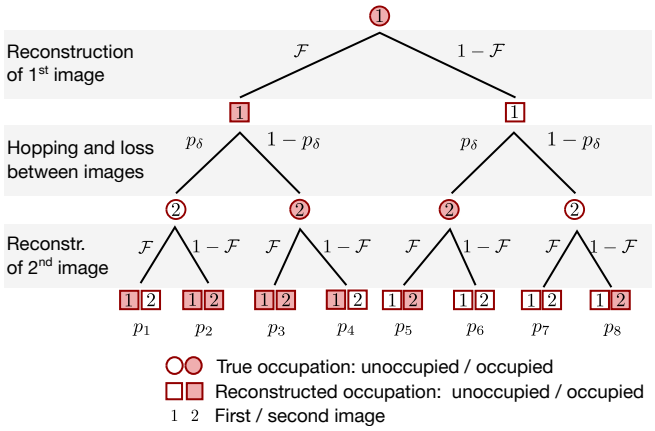
Supplementary Fig. 8 shows the reconstruction fidelities as a function of the average filling per image. While the overall fidelity has a minimum around half filling, the hole fidelity monotonically decreases toward higher fillings. This indicates that it is most difficult, even for simulated data, to identify individual holes in high-density regions. All fidelities stay well above 99 % for all fillings, therefore exceeding the experimentally determined value in the main text (Fig. 5 of the main text). The striking difference to the experimental results most likely comes from additional effects that have not been considered in the simulation. Examples are hopping and loss, density-dependent effects due to superradiance as well as other imaging imperfections such as spatially inhomogeneous imaging performance.

Deconvolved count statistics.– In addition to the quantitative fidelity evaluation, the simulation also provides deeper insights into the interpretation of the deconvolved count distributions. Similar to the evaluation of experimental data in the main text, we extract the values of the bottleneck layer before binarization via the activation function and show the distributions for different average fillings. In Supplementary Fig. 9 one finds well-separated bimodal distributions for all fillings. The two peaks can be attributed to occupied and unoccupied sites based on the known occupation labels. In the vicinity of the intersection point where the two distributions overlap, one finds that the tails extending into the respective other class decay rapidly. This suggests that the overlap area can be used as an approximation to estimate the reconstruction fidelity also in the case where the underlying distribution is not exactly known. Compared to the experimental case (Fig. 4 of the main text), the qualitative shape of the distributions obtained here is much closer to a normal distribution. The strong skewing in Fig. 4 of the main text is thus likely to be related to hopping and loss as well as density-dependent effects, which are not included in our simulation.

SUPPLEMENTARY NOTE 4: FIDELITY ESTIMATION FROM DOUBLE IMAGING

One of the two ways introduced in the main text for evaluating performance of the neural network reconstruction is imaging the same atom distribution twice, similar to the procedure for determining the loss and hopping rate, as described in the section 'Fluorescence imaging'. In the following, we will elaborate on how an experimental reconstruction fidelity can be derived from comparing the reconstruction results of the two images. In doing so, we assume that any errors in the reconstruction are caused by a finite SNR in presence of (random) noise and disregard systematic errors which entail a reconstruction bias toward either occupied or unoccupied sites.

Bernoulli trial.– Assuming that the hopping and loss probabilities are not too large, we can envision the process of taking two fluorescence images of the same cloud as a three-step process which can be thought of as a Bernoulli trial: (I) taking and reconstructing the first image, (II) hopping and loss causing actual differences between the two images and (III) taking and reconstructing the second image. Let us consider an arbitrary single site. The reconstruction fidelity $\mathcal{F}$ is the probability that a site in the first image is reconstructed correctly, while there is a probability of $(1-\mathcal{F})$ that the reconstruction yields the wrong occupation. Due to thermal hopping



Supplementary Figure 10. **Double imaging probability tree.** Probability tree modeling the reconstruction and hopping/loss processes as a three-step Bernoulli trial that depends on the reconstruction fidelity $\mathcal{F}$ and the hopping/loss probability p_δ. Quadratic and round symbols (red/white color) show the true and reconstructed occupation (occupied/unoccupied), respectively, of a site in the first or second image (1,2). The probabilities of the eight outcomes of the probability tree are labeled $p_1, \ldots, p_8$.

and loss events, the true occupation of the same site in the second image has a chance of p_δ to be different from that in the first image. The probability p_δ is proportional to both the average density in the image n and the hopping and loss probabilities in the exposure time window (p_{loss} and p_{hop}, respectively):

$$p_\delta(n) = n\,(p_{\text{loss}} + 2\,p_{\text{hop}}) \tag{S3}$$

The factor of two in front of the hopping term accounts for the fact that a hopping event always alters the occupation of two sites, leading to a two-fold higher difference between the first and the second image as compared to a loss event. Finally, the probability that the site occupation (altered by loss and hopping) is reconstructed correctly in the second image is again given by the reconstruction fidelity $\mathcal{F}$.

The probabilities governing this three-step process can be summarized in a tree diagram as shown in Supplementary Fig. 10. At the bottom of this diagram all eight outcomes with probabilities $p_1, \ldots, p_8$ are shown. As an example, $p_1 = \mathcal{F}\,p_\delta\,\mathcal{F}$ is the probability for the case where both images are reconstructed correctly but the two occupations are different due to a hopping or loss event. The probability $p_2 = \mathcal{F}\,p_\delta\,(1-\mathcal{F})$ corresponds to the case where the two reconstructed occupations match even though there was a hopping or loss event, as the second reconstruction result is incorrect.

Fidelity estimation.– Among the eight outcomes shown at the bottom of the probability tree only four (those numbered 1, 4, 5 and 8) cause a measured occupation difference between the first and the second image. Thus, the probability δ that the measured occupation of a site in the second image is different compared to the first image is:

$$\begin{aligned}\delta &= p_1 + p_4 + p_5 + p_8 \\ &= \mathcal{F}^2\,p_\delta(n) + \mathcal{F}\,(1-\mathcal{F})\,(1-p_\delta(n)) \\ &+ (1-\mathcal{F})\,\mathcal{F}\,p_\delta(n) + (1-\mathcal{F})^2\,(1-p_\delta(n)).\end{aligned} \tag{S4}$$

Solving this equation for $\mathcal{F}$ yields:

$$\mathcal{F} = \frac{1}{2}\left(1 + \sqrt{\frac{1-2\delta}{1-2\,p_\delta(n)}}\right). \tag{S5}$$

For vanishing hopping and loss rates ($p_\delta = 0$), this expression simplifies to:

$$\mathcal{F} = \frac{1}{2}(1 + \sqrt{1-2\delta}). \tag{S6}$$

If no difference between the first and the second image is measured ($\delta = 0$), we obtain a perfect reconstruction fidelity of $\mathcal{F} = 1$. In contrast, when measuring the maximum possible difference $\delta = 0.5$ (which is equivalent to comparing two random images), Eq. (S5) yields $\mathcal{F} = 0.5$, the lowest possible value for the reconstruction fidelity.

SUPPLEMENTARY REFERENCES

[1] Klostermann, T. *et al.* Fast long-distance transport of cold cesium atoms. *Phys. Rev. A* **105**, 043319 (2022).

[2] Klostermann, T. M. *Construction of a caesium quantum gas microscope.* Ph.D. thesis, Ludwig-Maximilians-Universität München (2022).

[3] von Raven, H. *A new caesium quantum gas microscope with precise magnetic field control.* Ph.D. thesis, Ludwig-Maximilians-Universität München (2022).

[4] Chin, C. *et al.* Precision Feshbach Spectroscopy of Ultracold Cs_2. *Phys. Rev. A* **70**, 032701 (2004).

[5] Weber, T., Herbig, J., Mark, M., Nägerl, H.-C. & Grimm, R. Bose-Einstein Condensation of Cesium. *Science* **299**, 232–235 (2003).

[6] Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems (2015). Software available from tensorflow.org.

[7] Chollet, F. *et al.* Keras (2015). Software available from keras.io.

[8] Yamamoto, R., Kobayashi, J., Kuno, T., Kato, K. & Takahashi, Y. An ytterbium quantum gas microscope with narrow-line laser cooling. *New J. Phys.* **18**, 023016 (2016).

[9] Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980* (2017).

[10] Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901* (2013).