nature materials

SUPPLEMENTARY INFORMATION

DOI: 10.1038/NMAT3289

Sparsity-based single-shot subwavelength coherent diffractive imaging

1 Experiment description

In our microscope, a laser beam at $\lambda = 532$ nm was launched onto the specimen using a NA=0.8 (40×) water immersion microscope objective. The transmitted light was collected using a NA=1.0 (60×) water immersion microscope objective and projected on a camera using a single optical lens. A schematic representation of the experimental setup is shown in Figure 1.

In order to observe both real plane and Fourier plane of the specimen, the image was taken at two different camera positions. We recall that the Fourier plane intensity was used for actual reconstruction, while the real plane intensity was used only for rough support estimation.

The specimen mask, whose transmission function corresponds to the optical information superimposed on the laser beam, is fabricated as follows: as substrate material we chose fused silica, because it is a high quality transparent material at optical frequencies, and because its processing technology is well developed. In order to to create a mask containing the optical image, we deposit opaque material on the substrate and make several patterned holes in it, such that the holes pass the light while the opaque material blocks it. For this purpose, we sputter a chromium layer onto the surface of the substrate. Chromium is a metal, which absorbs light at optical frequencies. Nevertheless, the thickness of the chromium layer has to be larger than the skin depth at optical frequencies, to avoid undesired transmission through that layer. Thus we select a thickness of 100 nm as suitable compromise between high quality optical behavior and fabrication considerations. The structures in the chromium layer are nano-holes, drilled in the chromium by a beam of focused gallium ions from a liquid metal ion source [1, 2] (Zeiss Neon 60). With this technology, it is feasible to mill the desired structures into the

DOI: 10.1038/NMAT3289

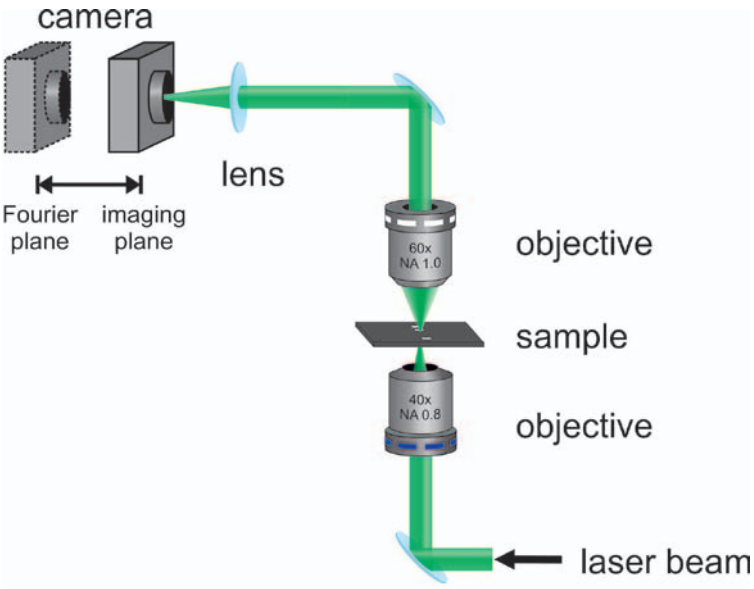
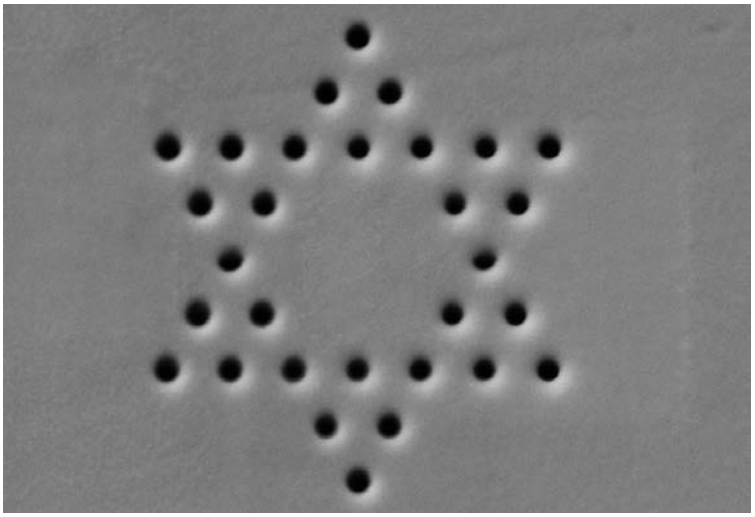
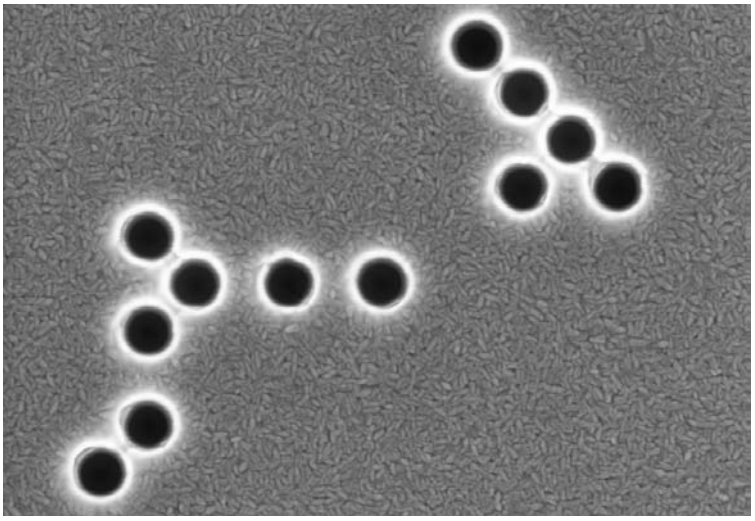


Figure 1: Schematic representation of the experimental setup.

chromium layer directly and efficiently, without any additional lithography process. Utilizing a convenient set of parameters, it is possible to imprint the designed structures into the metal layer, without significantly affecting the substrate material, and with high spatial accuracy. We fabricated two different samples yielding a two-dimensional sub-wavelength optical structure: (a) a Star of David (SOD) image, consisting of 30 holes, with 100 nm diameter each, spaced by 100 nm; and (b) a "random" image comprised of 12 circular holes of 100 nm diameter each, placed in a random order on a grid, as defined by (1). The Scanning Electronic Microscope (SEM) images of the samples are shown in Figure 2. Note that the SEM images are not in proportion as, in reality, the holes are of the same size and their diameter is equal to the spacing between holes. Generally, we use this approach throughout the paper: all images are shown in some abstract units that are, however, proportional to the corresponding physical quantities. The correspondence can be established using the fact that all holes are of diameter 100 nm.



(a)



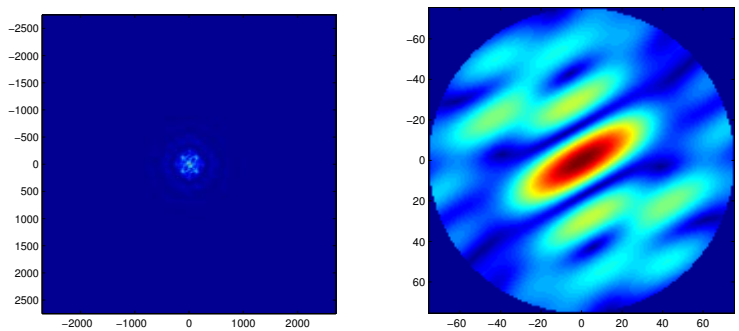
(b)

Figure 2: SEM images of the samples: (a) Star of David, (b) "random".

DOI: 10.1038/NMAT3289

2 CDI sparsity-based reconstruction

Under the experimental conditions described in the previous section, our problem amounts to reconstruction of a signal from the magnitude of its Fourier transform, assuming furthermore that this information is known only over a small interval of low frequencies as shown in Figure 3. The discussion below is general and applies to both examples given in the paper (and, of course, to a very large class of optical images). However, in order to make the explanation more succinct, we demonstrate most of the results on the "random" image, because it has no implicit symmetries.



(a) (b)

Figure 3: Fourier domain magnitude of the "random" image: (a) the full spectrum (simulated, without noise) needed to reconstruct the image precisely (by a simple application of the inverse Fourier transform), (b) the low-frequency part (actual measurements, in the presence of experimental noise).

Of course, when the majority of the frequencies are lost, precise reconstruction is not possible, unless we have, or may assume, some additional information about the sought signal. In fact, the problem is even more difficult as the measurements contain noise. In a manner similar to [3], we assume that the EM field in the object domain (u, v) can be represented precisely, or approximated adequately (hereinafter, this relation is denoted by $\approxeq$) by means of a known generating function $g(u, v)$. That is

$$E(u,v) \approxeq \sum_m \sum_n x_{mn} g(u - m\Delta_u, v - n\Delta_v) \ , \tag{1}$$

where x_{mm} are unknown signal coefficients in the basis defined by the shifted versions of $g(u, v)$. Note that the set $\{(m\Delta_u, n\Delta_v)\}$ defines a rectangular grid where the shifted versions of the generating function are located. Hence, for example, by choosing $g(u, v)$ to be the Dirac delta function we can obtain the sampled version of the continuous EM field distribution, where Δ_u, and Δ_v define the sampling interval. As another classical example: all bandwidth limited signals can be represented precisely in this form where g is chosen to be the jinc(ρ) function. For more examples see [4] and references therein. Of course, the generator must be chosen in a way that corresponds to the signal in question (although, in the most general case of 2D information, the generator could simply be rectangular pixels). In this section and in Section 3, where we compare our algorithm with other methods, we assume that the basis function is chosen in a way that allows a perfect reconstruction of the sought signal, namely g represents a circle of a priori known diameter (100 nm). We assume also that $\Delta_u = \Delta_v = 100nm$. That is, we assume that the sought signal is comprised of non-overlapping circles of known diameter. The grid $\{(m\Delta_u, n\Delta_v)\}$ containing all possible locations (144) is shown in Figure 4. Note that the exact placement of the grid is unimportant as our measurements are insensitive to shifts. A more detailed explanation of this property is presented in Section 4, where we discuss the implications of the grid assumption along with the impact of the basis function on the reconstructed signal.

Before proceeding on, there are two points we would like to stress. First, the assumption of an underlying grid is natural in many situations arising in digital signal processing. A prominent example are digital images that are comprised of pixels located on a rectangular grid. Just like in digital images, the grid in our case defines the resolution of a digitized version of the sought signal (see Section 4 for details). Second, it is important to note that all our comparisons with other methods are done under exactly the same assumptions, including a grid, basis functions, etc. As is evident from the experiments presented in Section 3, our algorithm outperforms other methods.

DOI: 10.1038/NMAT3289

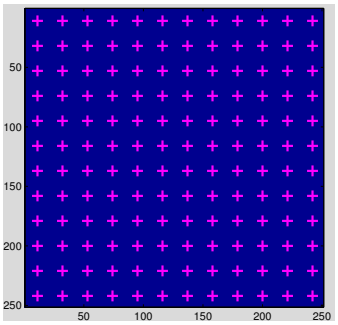
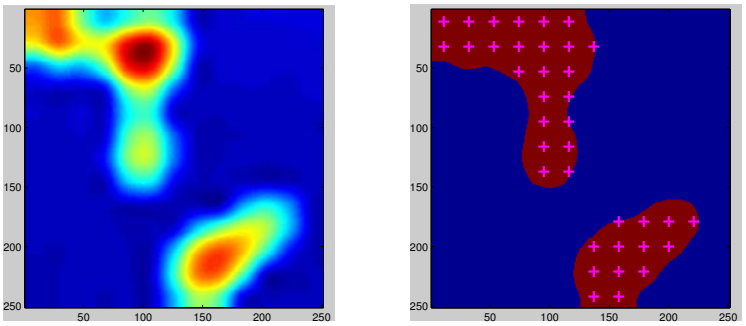


Figure 4: The full grid.

We emphasize that even if the correct number of circles were known (12 circles, in this example) there would be $\binom{144}{12} > 10^{17}$ possible variants to choose from for the signal support. To limit the search space, we use the blurred version of the signal as shown in Figure 5.



(a) (b)

Figure 5: Support restriction by the low-resolution image: (a) blurred image magnitude, (b) grid restricted by the blurred image.

However, even after this restriction, there still remain $\binom{37}{12} > 1.85 \times 10^9$ variants. More importantly, even after this restriction, the image cannot be reconstructed precisely unless additional information is available (see Section 3). Below we present our method that provides excellent reconstruction results based on the knowledge that the total number of circles in the image

is small, that is, the image is *sparse* in the basis associated with the circles, as defined by (1).

In our method, we reconstruct the support and the magnitude of the circles in the sought signal, simultaneously. To this end, we seek the sparsest x (x being a column vector comprised of the image coefficients x_{mn} as defined by (1)), that yields a good agreement with the measurements. Mathematically, we try to solve the following optimization problem

$$\begin{aligned} \min \quad & \|x\|_0 \\ \text{subject to} \quad & \| |LFCx| - r \|_2^2 \leq \epsilon \,, \\ & x \geq 0 \,. \end{aligned} \tag{2}$$

Here $\|\cdot\|_0$ denotes the l_0 norm: $\|x\|_0 = \sum_i |x_i|^0$, that is, $\|x\|_0$ equals the number of elements of x that are not zero. The measured (noisy) magnitude in the Fourier domain is denoted by r. Note that the operators and inequalities, like $|\cdot|$, and $\geq$ are applied element-wise. The matrix C represents all possible shifts of the generator function (a circle) so that Cx is the actual image that we reconstruct, F stands for the Fourier transform operator, and L represents the low-pass filter. That is, L is obtained from the identity matrix of appropriate size by removing most of its rows while keeping only those that correspond to the low frequencies of its operand, as shown in Figure 3. Physically, L is the low-pass filter associated with the cutoff spatial frequency of the optical system, which, for microscopes with NA=1, corresponds to the diffraction limit. Note that, due to errors in the measurements, the discrepancy in the Fourier domain is allowed to be up to some small value ϵ (> 0). A short discussion about the precise value of ϵ and whether it must be known a priori will follow. The last requirement $x \geq 0$ is valid because the optical information is generated by illuminating the sample with a plane wave, that is, a plane of equal amplitude and phase. Hence, the phase is the same across the whole image. Therefore, without loss of generality, we may assume that the phase is zero everywhere, since the absolute phase is unimportant. We do not assume that all circles have the same magnitude.

To solve (2) we developed an iterative method whose basic iteration contains the following two steps:

Step 1: Solve the minimization problem:

$$\begin{aligned} \min \quad & \| |LFCx| - r \|_2^2 \\ \text{subject to} \quad & x \geq 0 \,. \end{aligned} \tag{3}$$

DOI: 10.1038/NMAT3289

In practice, we use an unconstrained formulation whose solution is approximated by the L-BFGS method [5] (note that the problem is not convex and therefore its exact solution cannot in general be found).

Step 2: After a solution x to Step 1 is found, set to zero the entry of x with minimal value. Once set to zero the entry remains so forever.

A schematic representation of our method is given in Figure 6 below.

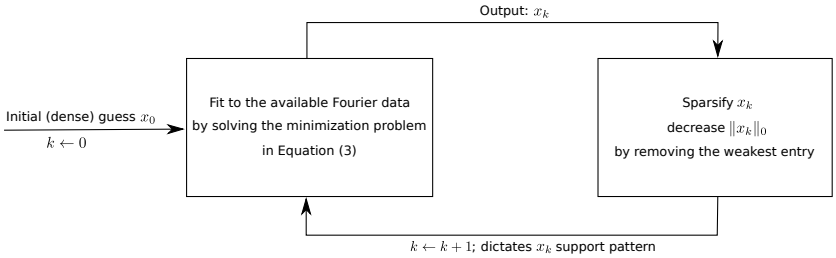


Figure 6: Flowchart diagram of our reconstruction method.

In theory, the iterations should be repeated so long as the constraint $\||LFCx| - r\|^2 \leq \epsilon$ is satisfied. It is often argued that the value of ϵ is known a priori or can be estimated from physical constraints (in the case of the "random" image, the difference between the measured Fourier magnitude r and its ideal variant r^* is $\|r - r^*\|^2 = 1.7434$, which corresponds to a signal-to-noise ratio of $\|r^*\|/\|r - r^*\| = 1/0.041$). However, it is an important question whether the best value of $\|x\|_0$ (the true number of circles in the image) can be determined *automatically*. Consider the different stages of our method as shown in Figure 7. Is there any way to recognize that the correct number of circles is 12 without knowing ϵ? It turns out that the answer to the above question is affirmative in many cases. As is evident from Figure 8, there is often a big jump in the objective function value when the number of circles dips below the correct value of 12. Hence, even without knowing the noise bound ϵ one can easily identify that the smallest number of circles that "explains" well the measurements is 12 (this is, of course, correct as long as the circles have large enough amplitude). The result of our reconstruction and the true image are shown in Figure 9. Note that some circles have low magnitude so they are invisible in the color images. We therefore, place the

'+' sign at the center of all circles in the image (x's entries that are not zeros).

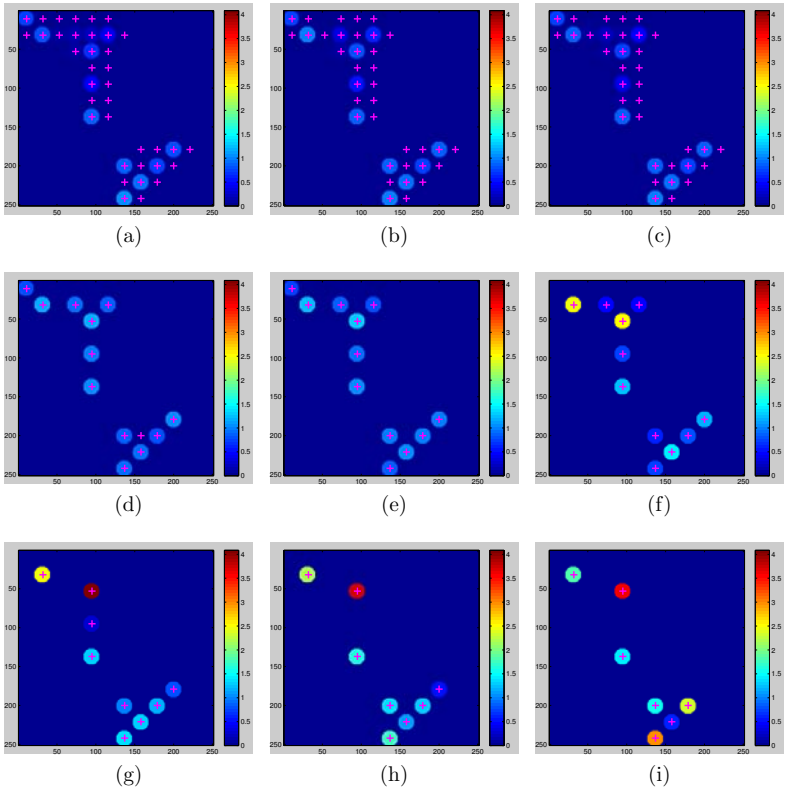


Figure 7: Reconstruction stages for the example of the "random" image. Each stage (iteration) corresponds to a certain number of circles (non-zero entries in x): (a) 37 circles, (b) 36, (c) 35, (d) 13, (e) 12, (f) 11, (g) 9, (h) 8, (i) 7.

DOI: 10.1038/NMAT3289

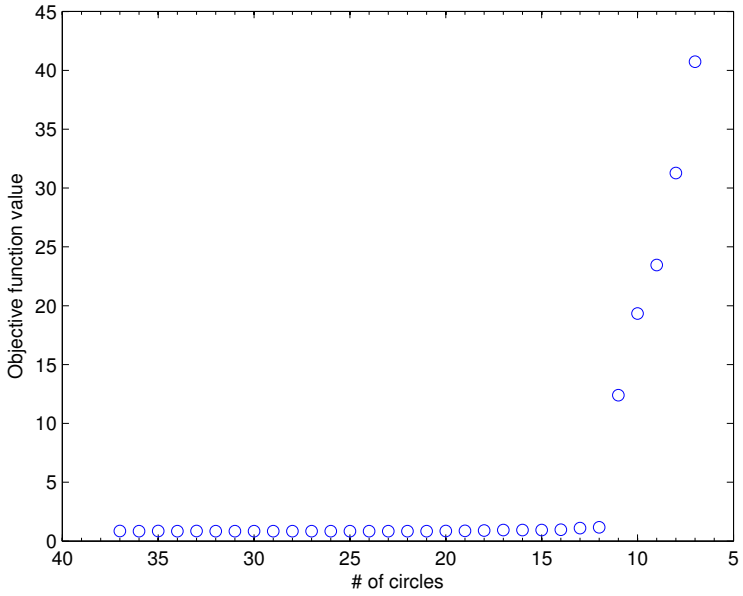
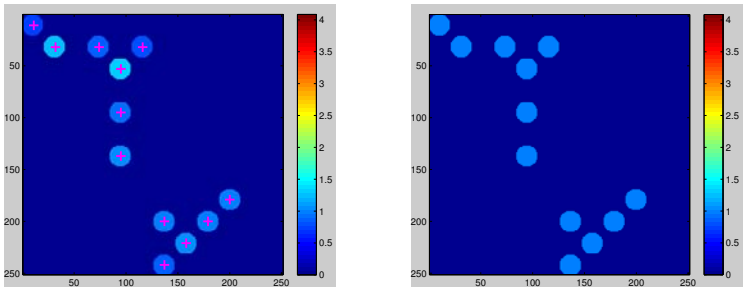


Figure 8: "Random" image: objective function value (Fourier domain discrepancy) versus the number of circles in the solution—a sharp jump occurs when the number of circles dips below the correct value of 12.



(a) (b)

Figure 9: Reconstruction result for the "random" image (a), and the true (original) image (b).

Very similar behavior is observed for the second image (SOD) whose results are shown in Figures 10 and 11.

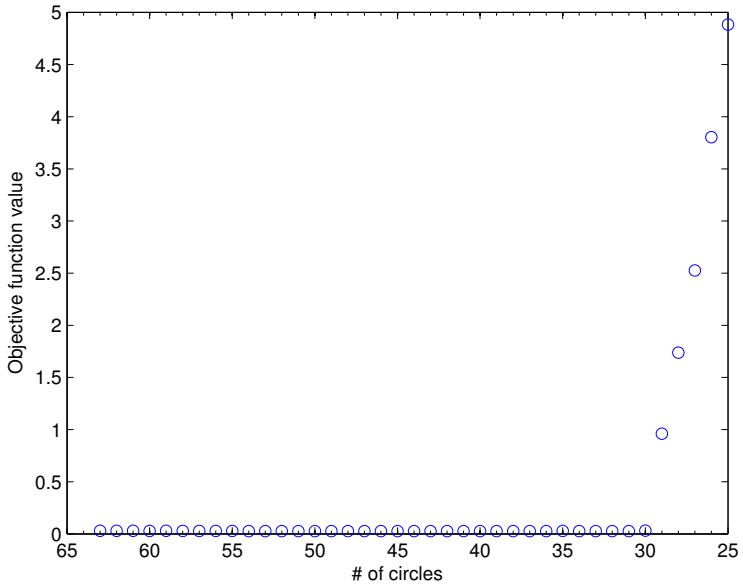
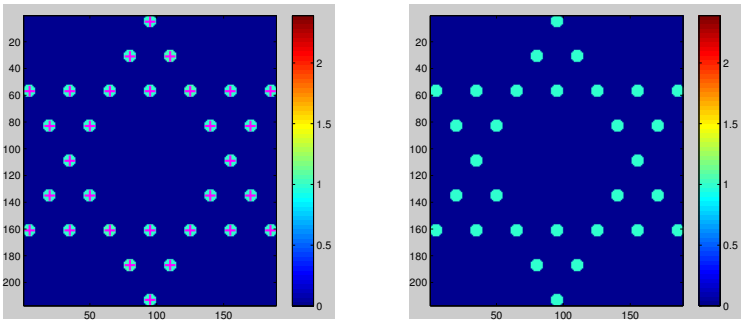


Figure 10: SOD image: objective function value (Fourier domain discrepancy) versus the number of circles in the solution—a sharp jump occurs when the number of circles dips below the correct value of 30.

DOI: 10.1038/NMAT3289



(a) (b)

Figure 11: Reconstruction result for the SOD image (a), and the true (original) image (b).

In Section 4 we demonstrate that choosing an "incorrect" basis function, even one whose shape does not allow perfect representation of the sought signal, results, nevertheless, in a reasonable reconstruction. Furthermore, we also demonstrate that the grid's cell size can be determined automatically.

3 Comparison with other methods

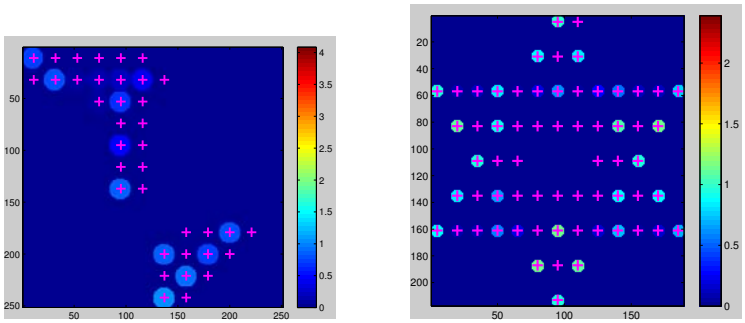
We would like to stress again that our algorithm is successful because we exploit the *sparsity* of the sought signal. To demonstrate this, we present a comparison with some classical reconstruction methods, and discuss the relation between our setup and classical compressed sensing.

3.1 Without a regularization

Our sparsity-based technique minimizes the l_0 norm subject to additional constraints. This formulation resembles closely a regularization imposed on x. Hence, the most naive approach would be to abandon the regularization altogether and to try to find x that minimizes the discrepancy in the measurements. That is, we might attempt to solve the following problem:

$$\begin{aligned} \min \quad & \| |LFCx| - r \|_2^2 \\ \text{subject to} \quad & x \geq 0 \ . \end{aligned} \tag{4}$$

Note that this is exactly the problem we solve in the first iteration of our method. However, using this approach as the full reconstruction process has a number of drawbacks. First, the problem of image reconstruction from the magnitude of its Fourier transform (also called phase retrieval) is known to be particularly tough for continuous optimization techniques (for explanation and further details see [6]). To the best of our knowledge, the most widely used method for phase retrieval without additional information is the Hybrid Input-Output algorithm [7]. A more detailed investigation of this method will follow in Section 3.3.2. Here, we present the results obtained by our optimization routine. As mentioned earlier, this formulation is equivalent to performing only one iteration of our method. Hence, the result is as shown in Figure 12. Note that the reconstruction contains many superfluous circles, and even if the correct number of the circles were known, a simple thresholding would yield an incorrect reconstruction.



(a) (b)

Figure 12: Reconstruction without regularization on x: (a) "random" image, (b) SOD image.

3.2 Replacing l_0 with another norm

Using l_2 regularization has long been a favorite among engineers due to its simplicity and the ability to obtain closed-form solutions in linear cases. In the non-linear case, these benefits are lost, of course. However, for us it is more important that the l_2 norm does not promote sparsity (actually, some papers claim that it usually results in the most dense solution possible [8]).

DOI: 10.1038/NMAT3289

To demonstrate that this regularization is not suitable for bandwidth extrapolation of sparse signals, we solved the following problem

$$\begin{aligned} \min \quad & \|x\|_2 \\ \text{subject to} \quad & \||LFCx| - r\|_2^2 \leq \epsilon \ , \\ & x \geq 0 \ . \end{aligned} \tag{5}$$

The problem was solved by transforming it into an unconstrained optimization problem and choosing the weights of the penalty function terms so as to get the discrepancy in the measurements close to the true values. That is, assuming that the true ϵ is known ($\epsilon = 1.74$ in the case of "random" image, and $\epsilon = 0.0329$ in the case of SOD image). To solve (5) we used exactly the same routine (L-BFGS) as in our main algorithm. The results are shown in Figure 13.

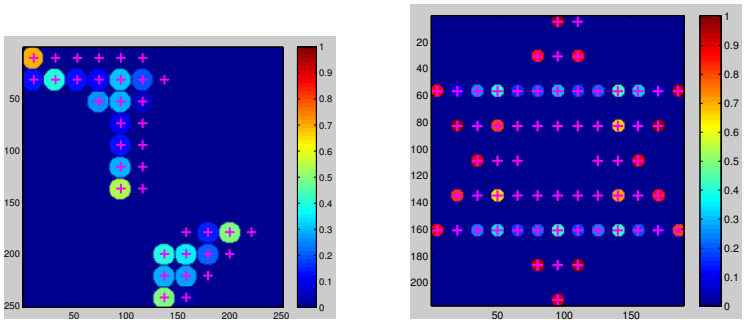


Figure 13: Reconstruction using l_2 regularization: (a) "random" image, and (b) SOD image.

It is obvious that the reconstruction quality is poor. Moreover, even if the correct number of circles were known, a simple thresholding would still produce an incorrect result.

Another viable alternative would be using the l_1 norm. A discussion on this norm is postponed to Section 3.4.

3.3 Methods based on alternating projections

A popular approach for phase recovery [9, 7] and bandwidth extrapolation [10, 11] is based on on a simple and elegant idea of alternating projections.

The current signal estimate is transformed back and forth between the object and the Fourier domains. In each domain, all available information is used to form the next estimate. Here we consider two major methods of this type: Gerschberg type method (often referred to as Gerschberg-Saxton or Gerschberg-Papoulis) and Fienup's Hybrid Input-Output method. The former is a classical method of alternating projections where all available information in the current domain is imposed upon the current estimate. In the latter approach the object domain information is not directly imposed on the current estimate; instead a more complex update rule is used as we explain later.

3.3.1 Gerschberg type methods

As mentioned before, Gerschberg type methods are "pure" projection techniques. The idea is to transform the current signal estimate back and forth between the signal and the Fourier domain performing a "projection" in each of the domains, that is, replacing the current estimate x_{cur} with the nearest one that satisfies the constraints in the relevant domain (x_{new}). Hence, in each domain the following optimization problem is solved

$$\begin{aligned} \min_{x_{new}} \quad & \|x_{cur} - x_{new}\|_2^2 \\ \text{subject to} \quad & x_{new} \in \mathcal{S} \ , \end{aligned} \tag{6}$$

where $\mathcal{S}$ denotes the set of all admissible signals in the current domain. In our case the estimate is first Fourier transformed. Then its (wrong) magnitude is replaced with the measured (correct) magnitude in the low-frequency regime. The resulting signal is back-transformed into the object domain (the result denoted by x') where it is converted into an image comprised of circles (denoted by x_{new}) in the following manner. Recall that the image model is of the following form $E = Cx$. Hence to find a projection we must solve the following problem

$$\begin{aligned} \min_{x_{new}} \quad & \|Cx_{new} - x'\|_2^2 \ , \\ \text{subject to} \quad & x_{new} \geq 0 \ . \end{aligned} \tag{7}$$

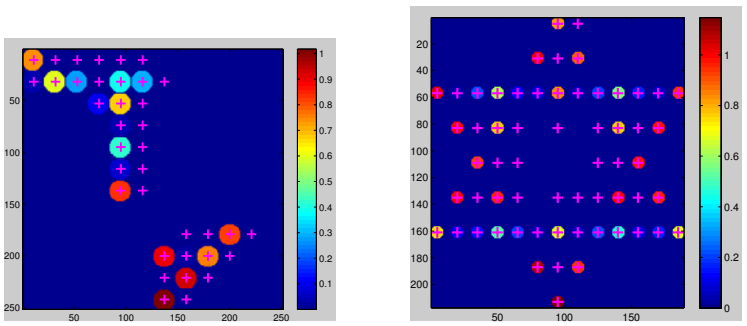
This problem is convex and therefore can be solved efficiently. In practice however, we used a two-step approximation instead of a full solution.

DOI: 10.1038/NMAT3289

Step 1 Solve $\min_{x_{new}} \|Cx_{new} - x'\|_2^2$. Note that this problem has a closed form solution: $x_{new} = C^{\dagger}x'$, where $C^{\dagger}$ denotes the Moore-Penrose pseudo-inverse of C.

Step 2 Set all entries of x_{new} that are negative to zero.

In general, this is not a true projection. However, it is a projection, if the vector x_{new} obtained after the first step is non-negative. This is indeed the case we observe in all our experiments. The results obtained after 5000 iterations of this method are shown in Figure 14. Usually, the correspondence between these and the true image falls considerably behind our sparsity-based reconstruction method.



(a) (b)

Figure 14: Gerschberg type method: results of reconstruction (a) "random" image, (b) SOD image.

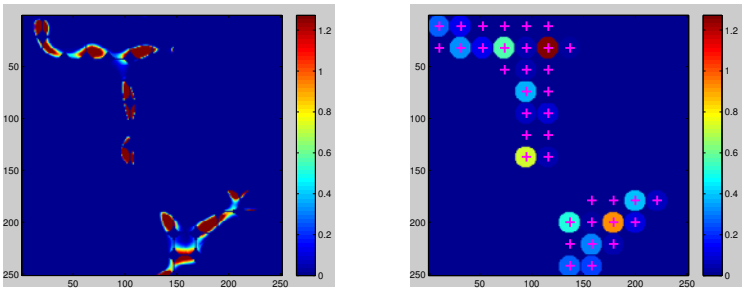
From the results above, it is evident that the reconstruction is poor and even if the correct number of circles were known a simple thresholding would still result in incorrect images.

3.3.2 Fienup's Hybrid Input-Output method

The Hybrid Input-Output method was developed by Fienup for the phase retrieval problem [7]. Although based on alternating projections, HIO does not enforce the object domain constraints, that is, the image is allowed to be non-zero in the off-support areas and the values may be negative. To the best of our knowledge, HIO is the most successful numerical method for

signal reconstruction from the magnitude of its Fourier transform. However, this algorithm only achieves good results when all or most of the Fourier spectrum is available. Judging by the result shown below, the method is not suitable for the situation where the Fourier magnitude is available only for a small fraction of the frequencies.

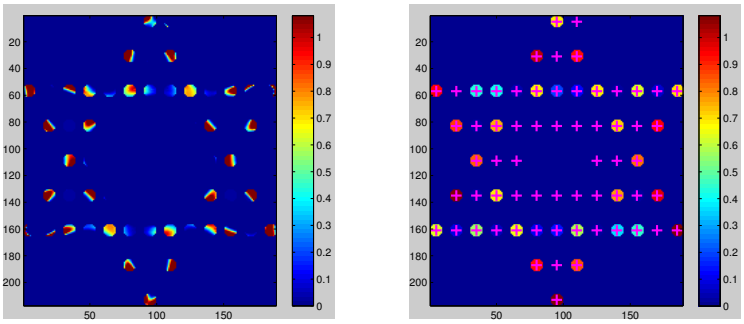
In our tests we applied the method in its original form, using only the Fourier domain magnitude and support information in the object domain (along with non-negativity). We did not try to enforce a constant value across every circle or zero values in the off-support areas, as the original method does not do that. As a post-processing step, the result returned by HIO was zeroed in the off-support areas (shown in Figures 15a and 16a) and then the values across each circle were averaged (shown in Figures 15b and 16b). As is evident from the results, the method is not capable of correct reconstruction of the signals. They cannot be recovered even if the correct number of circles is known: a simple thresholding will result in an incorrect reconstruction.



(a) (b)

Figure 15: Fienup's HIO method: "random" image results of reconstruction: (a) as produced by the method, (b) after enforcing a constant value across every circle.

DOI: 10.1038/NMAT3289



(a) (b)

Figure 16: Fienup's HIO method: SOD image results of reconstruction: (a) as produced by the method, (b) after enforcing a constant value across every circle.

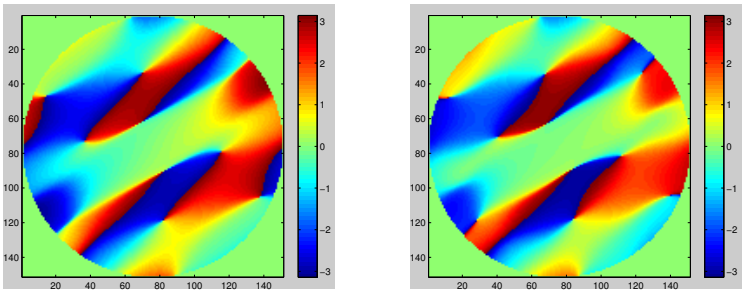
3.4 Relation to compressed sensing

Compressed sensing (CS) is an emerging field in image processing that performs signal reconstruction from a small number of its projections [12, 13, 14]. Conceptually, CS techniques and the corresponding mathematical theory are based heavily upon the sparsity of the sought signals. In its classical form, CS deals with measurements that are linear with respect to the unknown signal and generally assume random (or random-like) sampling distributed throughout the measurement domain. By contrast, in our current case of sub-wavelength CDI the measurements are: (1) nonlinear with respect to the sought signal, (2) taken only in a small (low-frequency) region of the measurement domain, where (3) they are taken in a periodic fashion (dictated by the pixels' arrangement of a digital camera sensor). Still, our reconstruction method relies on sparsity. As such, conceptually, our approach can still be viewed as CS in a broader sense.

Clearly, for the reasons stated above, many theoretical results and reconstruction methods of classical CS are not applicable to our problem. For example, the Matching Pursuit (MP) method [15] cannot be applied in its original form. Another popular method Basis Pursuit (BP) [8], could, in principle, be applied here (considering BP as a general approach based on replacing the l_0 with the l_1 norm, rather than a specific algorithm). How-

ever,in contrast to the linear case, in our nonlinear problem—using the l_1 norm still does not lead to a convex problem.

Besides the standard CS methodsit is instructive to consider other sparsity-based approaches which are related to CS, in the broader sense. One of these is based on division of the reconstruction process into two stages: at the first stage the missing Fourier phase is reconstructed using Fienups HIO algorithm (or Gerchberg-type method); at the second stage this phase is combined with the measured Fourier magnitude to form complete measurements that are linear with respect to the unknown signal. Once these linear measurements are available, one can use methods from classical CS (like, for example, BP) or our previously proposed method NLHT [3], which is aimed at recovering data from linear low-pass measurements. We find, however, that this approach does not produce high quality results. This failure is, probably, attributed to inability of the projection-based methods to reconstruct the phase precisely, as shown in Figure 17 below.



(a) (b)

Figure 17: Fourier phase of the "random" image: (a) the true phase, (b) the phase obtained after 5000 iterations of HIO.

Recently, several works have considered CS with quadratic nonlinear measurements [16, 17]. In both papers the resulting nonlinear constraints are relaxed to semidefinite constraints using matrix lifting and an appropriate sparsity promoting objective is used. The work of [17] considers phase retrieval assuming the availability of several diffraction patterns obtained from multiple structured illuminations, which is not relevant to our problem. In contrast, the scenario considered in [16] is much closer to our current case. Namely, simultaneous phase retrieval and bandwidth extrapolation from a

DOI: 10.1038/NMAT3289

single-shot power-spectrum measurement. In fact, our present problem can be viewed as a special case of that addressed in [16]. However, the algorithm suggested in [16] is targeting a more general problem, and hence its computational complexity is high. With this reasoning in mind, we devised the new sparsity-based approach and algorithm described in this Section 2, which is tailored for the specific problem of sub-wavelength CDI.

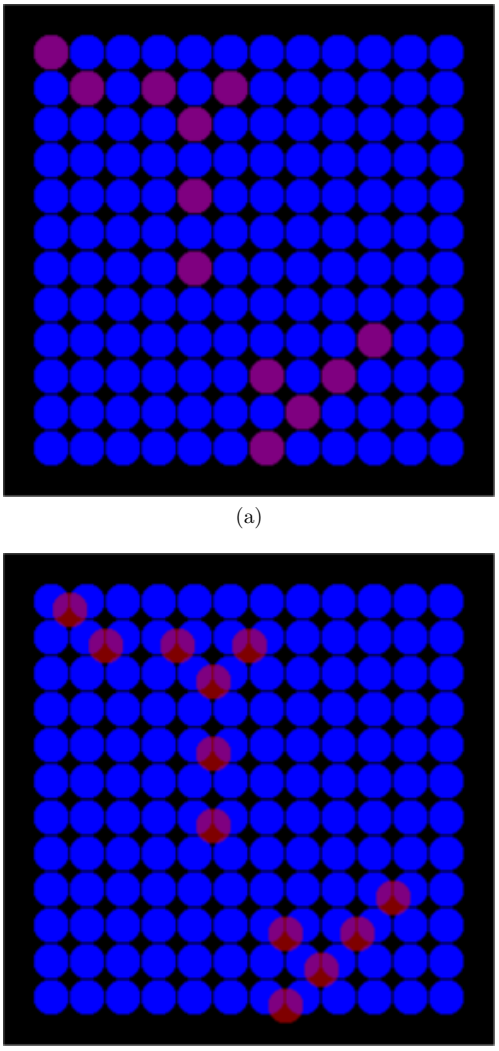
4 Automatic grid determination, and the (un)-importance of the basis function

In this section we would like to discuss the implications of our assumption regarding the existence of a grid that, in fact, defines a discrete set of allowed locations where the chosen basis function can be placed. In many cases, especially when the optical information represents experimental data, introducing such a grid is highly justified. For example, a digital image is obtained from a continuous intensity distribution by sampling it with a sensor that physically is an array of square pixels arranged in rows and columns. Hence, naturally, the gird is rectangular and the basis functions are squares whose size is equal to the grid's cell size. Likewise, our reconstruction provides a digitized version of the true signal as if it were performed by a sensor whose pixels' shape corresponds to the chosen basis function (circular in our experiments above). Hence, the grid used in our reconstruction algorithm essentially defines the resolution of the reconstructed image. This is especially true when the spatial extent of the basis function is smaller than (or equal to) the grid's cell size. An example of such a sensor with circular pixels is shown in Figure 18.

However, there is an important dissimilarity between our case and the regular sampling in the object domain. Since our measurements contain only the Fourier magnitude and no information is available about the phase, we cannot distinguish between all the shifted versions of the original signal. Namely, if $E(u, v)$ represents the original signal, our best hope is to reconstruct a shifted version of it, that is, $E(u - \Delta u, v - \Delta v)$ for some Δu, and Δv. Which version (shift) of the original signal is reconstructed depends, of course, on the reconstruction method. Because our approach seeks the sparsest solution, we obtain the digitization that corresponds to the perfect alignment shown in Figure 18a and not the "misaligned" version shown in

Figure 18b. In the latter case each circle in the original image "switches on" two pixels in the sensor, in contrast to one pixel per circle in the aligned case. Hence, one does not need to manually align the grid with respect to the sought signal as the best alignment is typically obtained automatically with our reconstruction method. The only remaining concern regarding the grid alignment is related to the placement of the blurred image that we use for loose support estimation. Fortunately, the solution to this problem is easy: the blurred image must be placed in a way that guarantees maximal grid coverage, that is, we shall keep as many allowed locations as possible.

DOI: 10.1038/NMAT3289



(a)

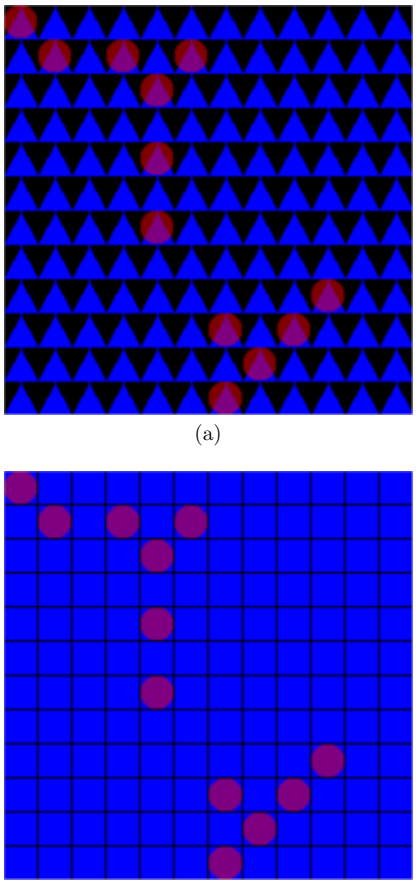
(b)

Figure 18: The sought signal (red) imposed on a sensor with circular pixels (blue). Note that in this case the best alignment (a) is automatically obtained as it results is a sparser reconstruction than a bad alignment (b).

4.1 The impact of the basis function

Let us now consider the situations where the basis function is chosen in a way that does not allow perfect reconstruction even without noise. Specifically, we consider basis functions in a shape of a square and a triangle, as shown in Figure 19.

DOI: 10.1038/NMAT3289

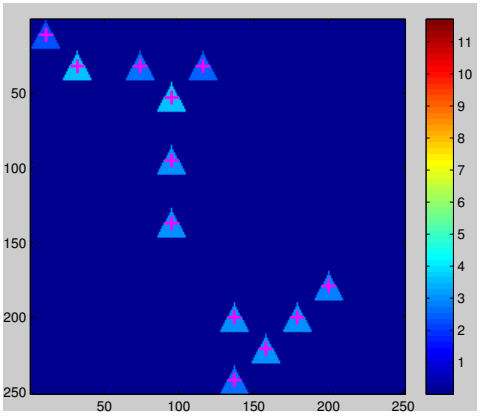


(a)

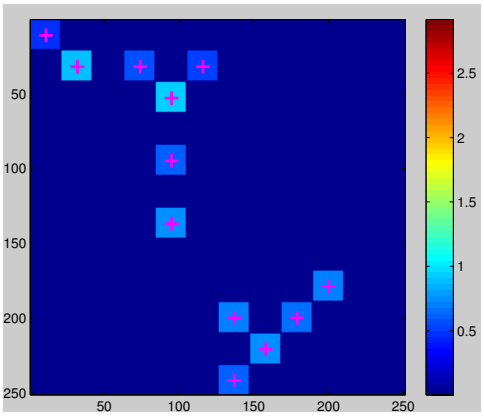
(b)

Figure 19: Basis functions that do not allow perfect reconstruction: triangles (a), and squares (b).

As is evident from Figure 20, the reconstruction in these cases matches our expectations: we obtain the correct "digitized" version of the sought signal that corresponds to the chosen basis function and the grid. We emphasize the fact that all experiments are done with the same actual data that contains noise (see page 8).



(a)



(b)

Figure 20: Reconstruction in the case of basis functions that do not match the sought signal.

Moreover, if we consider the progress of the reconstruction process (see Figure 21) we observe that even an incorrect choice of the basis function has no adverse effect on the reconstruction. This fact has a simple explanation: the difference between a circle and a square (or a triangle) of size 100nm is much smaller than 100nm. Hence, being able to distinguish be-

DOI: 10.1038/NMAT3289

tween these shapes would mean effective resolution that is much better than 100nm. Thus, we conclude that the shape of the basis function is not of great importance so long as its size matches the size of a typical feature in the sought signal. In what follows, we evaluate the possibility to discover the most appropriate grid pitch (basis function size) automatically, without any prior information.

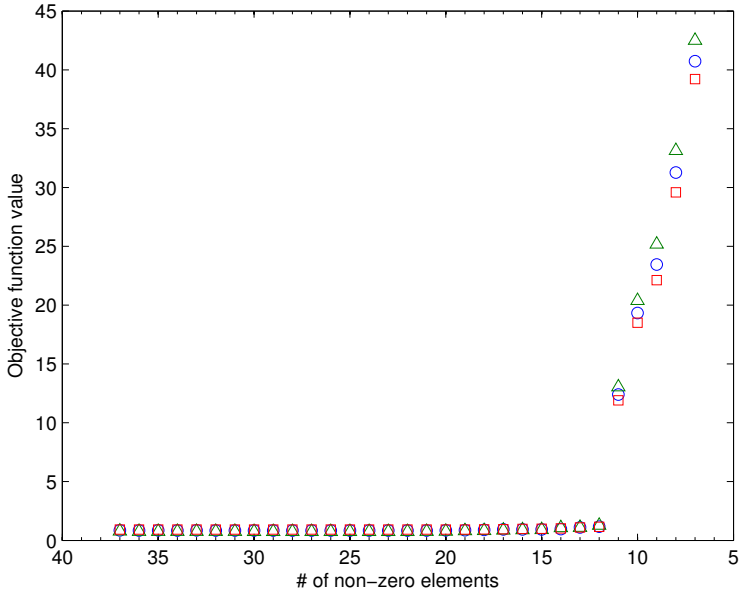


Figure 21: Reconstruction of the "random" image using different basis functions: objective function value (Fourier domain discrepancy) versus the number of circles/squares/triangles in the solution. Marker shape corresponds to the basis function shape.

4.2 A method for automatic grid determination

So far, we have seen that the shape of the basis function has no severe impact on the reconstruction process. Moreover, the best possible alignment is obtained automatically due to our requirement of maximal sparsity. These

two properties can be used for automatic determination of the optimal grid pitch. To this end we ran a series of experiments with different grids whose pitch varies from 10 to 32 pixels (corresponds to the range of 48–152nm) using the square basis function of the size that matches the grid cell. As was mentioned earlier, the results of Section 4.1 show that the particular choice of the basis function is not very important. Hence, we could choose any shape of the size equal to the grid pitch. The choice of the square basis function was stipulated by the fact that most digital images are comprised of square pixels. Hence, this basis function will, probably, be the first choice in the situation where nothing is known about the sought signal.

For each grid pitch we ran a few iterations of our method keeping the lowest discrepancy in the Fourier space as a numerical value that corresponds to the current grid pitch. There is no need to solve the problem completely, as our goal here is to see whether the sought signal can be represented well by the current grid. We expect that fine grids (small pitch) will allow good representations so long as the grid's pitch is smaller than or equal to the size of a typical feature in the signal. However, once the grid becomes too coarse, we expect a rapid growth of the objective function value. Hence, we expect the graph to have the distinctive "⌙"-shape, similar to the graphs in Figures 21, 10, and 8. As is evident from Figure 22, our expectations are confirmed by the experimental results.

DOI: 10.1038/NMAT3289

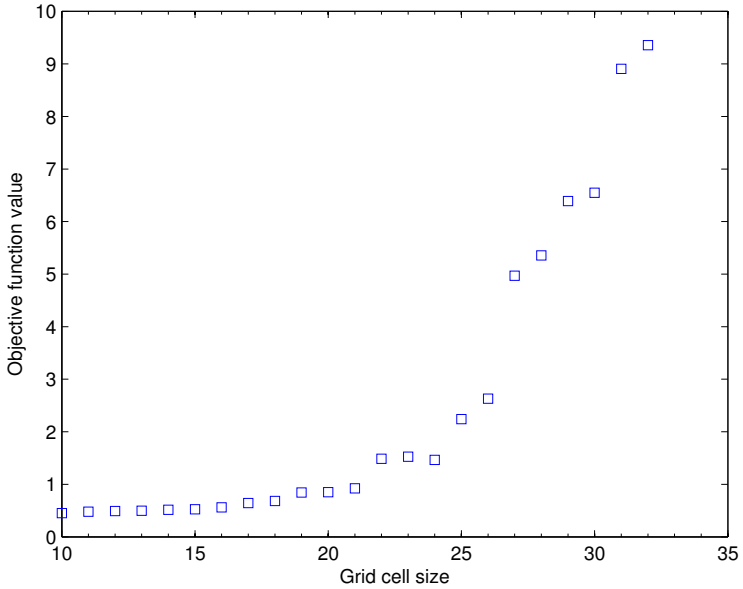


Figure 22: Objective function value (Fourier domain discrepancy) versus grid's pitch size.

Note that the first sharp jump in the objective function value happens during the transition from 21 pixels (the correct value) to 22 pixels. However, it may be argued that the transition is not sufficiently apparent and the true value may lie in some small interval around 21 pixels. Hence, we evaluate the behavior of our reconstruction method for a grid pitch lying in the interval of 18–24 pixels. As is evident from Figure 23, only the correct value of 21 pixels results in a clear and sharp jump after we dip below the correct value of squares (12). This property can be used for pinpointing the correct pitch size. Hence, an automatic subroutine for pitch determination is comprised of two steps: first, run a few iterations of our reconstruction method to obtain quantitative results indicating how well different gird sizes can represent the sought image; second, run a full reconstruction procedure for a limited range of pitches near the elbow in Figure 22 and check what pitch results in a clear evidence of existence of the sparsest solution (as in Figure 23).

Note that the obtained grid cell size is *optimal* in the sense that it satisfies

two important properties simultaneously: first, it allows good approximation of the sought signal; second, it leads to a highly evident sparse solution.

The suggested method is also based on the sparsity assumption: it works well when there are a few features in the sought signal that are of approximately the same size. This situation arises in many physical setups. However, we currently are working on extending the algorithm to the cases where the signal features may be of varying sizes.

DOI: 10.1038/NMAT3289

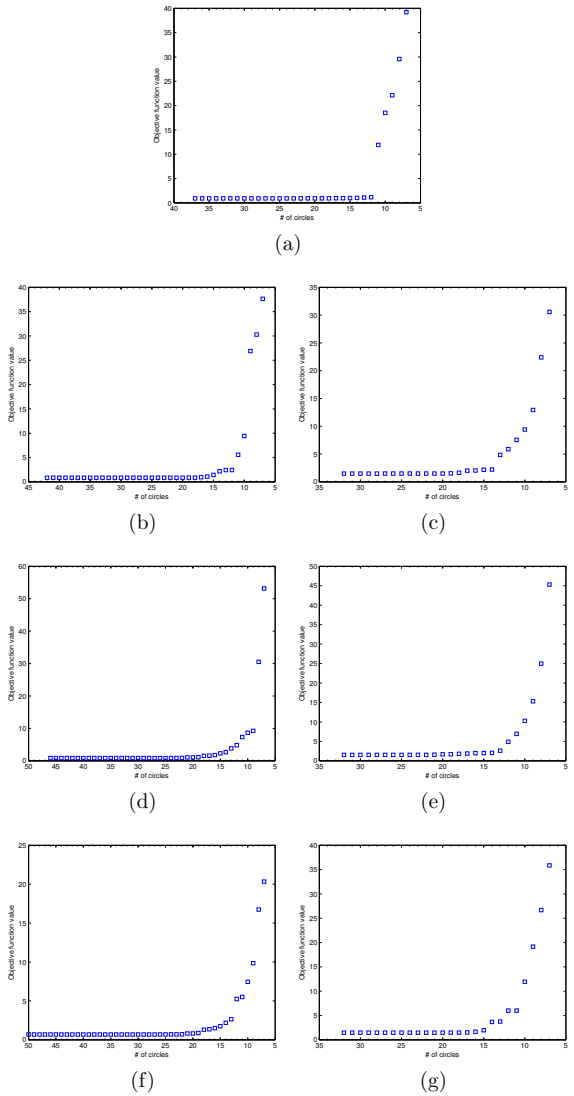


Figure 23: Objective function behavior versus the grid's pitch size (in pixels): (a) 21—the correct value, (b) 20, (d) 19, (f) 18, (c) 22, (e) 23, (g) 24.

5 Concluding remarks

We presented a simple greedy algorithm for bandwidth extrapolation of sparse signals. The method was demonstrated on two noisy measurements producing excellent reconstruction quality with resolution several times higher than the best possible theoretical value. The crucial role of sparsity was demonstrated by comparison with other algorithms that do not use this structural information. In addition, we presented an approach for automatic determination of alignment and grid size (resolution) that is also based on sparsity ideas. We certainly anticipate that sparsity-based techniques will be developed further in the near future, and will yield even better results.

We conclude by noting that our method here was designed for the specific problem of sub-wavelength CDI. However, conceptually, our approach is directly relevant to a broad class of problems where a sparse signal is to be reconstructed from a small number of nonlinear measurements. Examples range from super-resolution spectroscopy—where absolute-value (intensity) measurements are taken within a small temporal window, to the related problem of recovering the shape of ultrashort pulses with a detector of a smaller bandwidth. The underlying concept presented here is based on exploiting sparsity—which is intimately related to minimizing the number of degrees of freedom while recovering the sought signal. We expect that the next few years will witness a variety of related ideas in many areas of science, improving the resolution of many measurement devices and systems.

DOI: 10.1038/NMAT3289

References

[1] Krohn, V. E. & Ringo, G. R. Ion source of high brightness using liquid metal (1975).

[2] Prewett, P. D. & Jefferies, D. K. Characteristics of a gallium liquid metal field emission ion source. *Journal of Physics D: Applied Physics* **13**, 1747–1755 (1980).

[3] Gazit, S., Szameit, A., Eldar, Y. C. & Segev, M. Super-resolution and reconstruction of sparse sub-wavelength images. *Optics Express* **17**, 23920–23946 (2009).

[4] Eldar, Y. C. & Michaeli, T. Beyond bandlimited sampling. *IEEE Signal Processing Magazine* **26**, 48–68 (2009).

[5] Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* **45**, 503–528 (1989).

[6] Osherovich, E., Zibulevsky, M. & Yavneh, I. Approximate Fourier phase information in the phase retrieval problem: what it gives and how to use it. *Journal of the Optical Society of America A* **28**, 2124–2131 (2011).

[7] Fienup, J. R. Phase retrieval algorithms: a comparison. *Applied Optics* **21**, 2758–2769 (1982).

[8] Chen, S. S., Donoho, D. L. & Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM JOURNAL ON SCIENTIFIC COMPUTING* **20**, 33–61 (1999).

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

[10] Gerchberg, R. W. Super-resolution through error energy reduction. *Journal of Modern Optics* **21**, 709–720 (1974).

[11] Papoulis, A. A new algorithm in spectral analysis and band-limited extrapolation. *IEEE Transactions on Circuits and Systems* **22**, 735–742 (1975).

[12] Donoho, D. L. Compressed sensing. *IEEE Transactions on Information Theory* **52**, 1289–1306 (2006).

[13] Candes, E., Romberg, J. & Tao, T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on* **52**, 489–509 (2006).

[14] Candes, E. J. & Tao, T. Near-Optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory* **52**, 5406–5425 (2006).

[15] Mallat, S. G. & Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing* **41**, 3397–3415 (1993).

[16] Shechtman, Y., Eldar, Y. C., Szameit, A. & Segev, M. Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. *Optics Express* **19**, 14807–14822 (2011).

[17] Candes, E. J., Eldar, Y. C., Strohmer, T. & Voroninski, V. Phase retrieval via matrix completion. *arXiv:1109.0573* (2011).