Lithographic source optimization based on adaptive projection compressive sensing

Xu Ma,[1,*] Dongxiang Shi,[1] Zhiqiang Wang,[1] Yanqiu Li,[1] and Gonzalo R. Arce[2]

[1]*Key Laboratory of Photoelectronic Imaging Technology and System of Ministry of Education of China, School of Optoelectronics, Beijing Institute of Technology, Beijing 100081, China*
[2]*Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716, USA*
*maxu@bit.edu.cn

Abstract: This paper proposes to use the a-priori knowledge of the target layout patterns to design data-adaptive compressive sensing (CS) methods for efficient source optimization (SO) in lithography systems. A set of monitoring pixels are selected from the target layout based on blue noise random patterns. The SO is then formulated as an under-determined linear problem to improve image fidelity according to the monitoring pixels. Adaptive projections are then designed, based on the a-priori knowledge of the target layout, in order to further reduce the dimension of the optimization problem, while trying to retain the SO performance. Different from traditional CS methods, adaptive projections are constructed directly from the target layout data via a nonlinear thresholding operation. Adaptive projections are proved to achieve superior SO performance over the random projections. This paper also studies and compares the impact of different sparse representation bases on the SO performance. It is shown that the discrete cosine transform (DCT), spatial and Haar wavelet bases are good choices for source representation.

References and links

1. A. K. Wong, *Resolution Enhancement Techniques in Optical Lithography* (SPIE, 2001).
2. X. Ma and G. R. Arce, *Computational Lithography*, Wiley Series in Pure and Applied Optics, 1st ed. (John Wiley and Sons, 2010).
3. Y. Granik, "Source optimization for image fidelity and throughput," J. Microlith. Microfab. Microsyst. **3**(4), 509–522 (2004).
4. K. Tian, A. Krasnoperova, D. Melville, A. E. Rosenbluth, D. Gil, J. Tirapu-Azpiroz, K. Lai, S. Bagheri, C. C. Chen, and B. Morgenfeld, "Benefits and trade-offs of global source optimization in optical lithography," Proc. SPIE **7274**, 72740C (2009).
5. K. Iwase, P. D. Bisschop, B. Laenens, Z. Li, K. Gronlund, P. V. Adrichem, and S. Hsu, "A new source optimization approach for 2x node logic," Proc. SPIE **8166**, 81662A (2011).
6. J. C. Yu, P. Yu, and H. Y. Chao, "Fast source optimization involving quadratic line-contour objectives for the resist image," Opt. Express **20**(7), 8161–8174 (2012).
7. L. Wang, S. Li, X. Wang, G. Yan, and C. Yang, "Source optimization using particle swarm optimization algorithm in optical lithography," ACTA OPTICA SINICA **35**(4), 0422002 (2015).
8. H. Jiang and T. Xing, "A method of source optimization to maximize process window," Laser & Optoelectronics Progress **52**, 101101 (2015).
9. A. E. Rosenbluth, S. Bukofsky, C. Fonseca, M. Hibbs, K. Lai, A. Molless, R. N. Singh, and A. K. Wong, "Optimum mask and source patterns to print a given shape," J. Microlith. Microfab. Microsyst. **1**(1), 13–30 (2002).
10. A. Erdmann, T. Fühner, T. Schnattinger, and B. Tollkühn, "Towards automatic mask and source optimization for optical lithography," Proc. SPIE **5377**, 646–657 (2004).
11. X. Ma and G. R. Arce, "Pixel-based simultaneous source and mask optimization for resolution enhancement in optical lithography," Opt. Express **17**(7), 5783–5793 (2009).
12. J. Yu and P. Yu, "Gradient-based fast source mask optimization (SMO)," Proc. SPIE **7973**, 797320 (2011).
13. J. Li, Y. Shen, and E. Lam, "Hotspot-aware fast source and mask optimization," Opt. Express **20**(19), 21792–21804 (2012).
14. X. Ma, C. Han, Y. Li, L. Dong, and G. R. Arce, "Pixelated source and mask optimization for immersion lithography," J. Opt. Soc. Am. A **30**(1), 112–123 (2013).

15. J. Li, S. Liu, and E. Lam, "Efficient source and mask optimization with augmented Lagrangian methods in optical lithography," Opt. Express **21**(7), 8076–8090 (2013).
16. C. Han, Y. Li, X. Ma, and L. Liu, "Robust hybrid source and mask optimization to lithography source blur and flare," Appl. Opt. **54**(17), 5291–5302 (2015).
17. S. Hansen, "Source mask polarization optimization," J. Micro/Nanolith. MEMS MOEMS **10**(3), 033003 (2011).
18. X. Ma, L. Dong, C. Han, J. Gao, Y. Li, and G. R. Arce, "Gradient-based joint source polarization mask optimization for optical lithography," J. Micro/Nanolith. MEMS MOEMS **14**(2), 023504 (2015).
19. X. Guo, Y. Li, L. Dong, L. Liu, X. Ma, and C. Han, "Parametric source-mask-numerical aperture co-optimization for immersion lithography," J. Micro/Nanolith. MEMS MOEMS **13**(4), 043013 (2014).
20. C. Han, Y. Li, L. Dong, X. Ma, and X. Guo, "Inverse pupil wavefront optimization for immersion lithography," Appl. Opt. **53**(29), 6861–6871 (2014).
21. M. Mulder, A. Engelen, O. Noordman, G. Streutker, B. Drieenhuizen, C. Nuenen, W. Endendijk, J. Verbeeck, W. Bouman, A. Bouma, R. Kazinczi, R. Socha, D. Jürgens, J. Zimmermann, B. Trauter, J. Bekaert, B. Laenens, D. Corliss, and Greg McIntyre, "Performance of FlexRay, a fully programmable illumination system for generation of freeform sources on high NA immersion systems," Proc. SPIE **7640**, 76401P (2010).
22. L. Wei and Y. Li, "Hybrid approach for the design of mirror array to produce freeform illumination sources in immersion lithography," Optik **125**, 6166–6171 (2014).
23. Z. Song, X. Ma, J. Gao, J. Wang, Y. Li, and G. R. Arce, "Inverse lithography source optimization via compressive sensing," Opt. Express **22**(12), 14180–14198 (2014).
24. E. Candés, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inform. Theory **52**(2), 489–509 (2006).
25. D. Donoho, "Compressive sensing," IEEE Trans. Inform. Theory **52**(4), 1289–1306 (2006).
26. R. Ulichney, "Dithering with blue noise," Proc. IEEE **76**(1), 56–79 (1988).
27. L. Wei, "Multi-class blue noise sampling," ACM Transactions on Graphics **29**(4), 157-166 (2010).
28. D. L. Lau, R. Ulichney, and G. R. Arce, "Blue and green noise halftoning models ," IEEE Signal Processing Magazine **20**(4), 28–38 (2003).
29. T. Mitsa and K. J. Parker, "Digital halftoning technique using a blue-noise mask," J. Opt. Soc. Am. A **9**(11), 1920–1929 (1992).
30. D. L. Lau and G. R. Arce, *Modern Digital Halftoning*, 2st ed. (CRC, 2008).
31. G. R. Arce, D. J. Brady, L. Carin, H. Arguello, and D. S. Kittle, "Compressive coded aperture spectral imaging: an introduction," IEEE Signal Processing Magazine **31**(1), 105–115 (2014).
32. Z. Wang and G. R. Arce, "Variable density compressed image sampling," IEEE Trans. Image Process. **19**(1), 264–270 (2010).
33. A. P. Cuadros, C. Peitsch, H. Arguello, and G. R. Arce, "Coded aperture optimization for compressive X-ray tomosynthesis," Opt. Express **23**(25), 32788–32802 (2015).
34. H. Arguello and G. R. Arce, "Colored coded aperture design by concentration of measure in compressive spectral imaging," IEEE Trans. Image Process. **23**(4), 1896–1908 (2014).
35. D. L. Lau, G. R. Arce, and N. C. Gallagher, "Digital color halftoning with generalized error diffusion and multi-channel green-noise masks," IEEE Trans. Image Process. **9**(5), 923–935 (2000).
36. D. L. Lau, R. Ulichney, and G. R. Arce, "Blue and green-noise halftoning models," IEEE Signal Processing Magazine **20**(4), 28–38 (2003).
37. D. L. Lau, G. R. Arce, and Neal C. Gallagher, "Green-noise digital halftoning," Proc. IEEE **86**(12), 2424-2444 (1998).
38. S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation-based image restoration," Multiscale Model. Simul. **4**(2), 460–489 (2005).
39. J. F. Cai, S. Osher, and Z. Shen, "Linearized bregman iterations for compressed sensing," Mathematics of Computation **78**(267), 1515–1536 (2009).
40. http://www.itrs2.net/.

1. Introduction

Optical lithography is extensively used in very-large and ultra-large scale semiconductor fabrication. Figure 1 is a sketch of the optical lithography system. The lithographic source emits deep ultra-violet light, which is transmitted through the mask and projector, transferring the layout pattern from the mask onto the wafer. The immersion medium, such as deionized water, is often inserted between the projector and the wafer to further enhance the resolution. The projected mask pattern, however, is always distorted by multiple factors, such as the optical proximity effect, thick mask effect, defocus, exposure variation, and so on. Since the critical dimension (CD) of the integrated circuit shrinks into the deep sub-wavelength realm, resolution enhancement techniques (RET) become indispensable to guarantee imaging resolution and
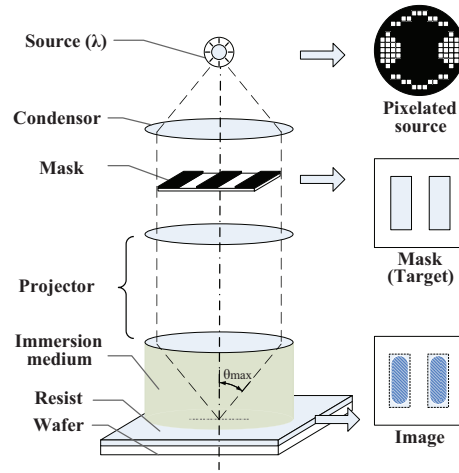
Fig. 1. The sketch of optical lithography system.

fidelity of the optical lithography systems [1, 2]. Pixelated source optimization (SO) is an important RET used to overcome the aforementioned distortion incurred in the lithography patterns. As depicted in Fig. 1, pixelated SO methods treat the source pattern as an array of equidistant pixels, and the intensities of all source pixels are designed to modify the directions and intensities of the incident light rays, so as to influence the diffraction orders captured by the projector and to compensate for the image distortion. Thus, we can achieve better imaging performance with an optimized source. In the past, several approaches to pixelated SO have been proposed in the literature [3–8]. SO methods can also be combined with other RETs to form joint optimization techniques among the source, mask, polarization, numerical aperture (NA), pupil wavefront, and so on [9–20].

Lithography systems employ Koehler illumination to generate uniform illumination on the mask plane [2]. Accordingly, the Abbe imaging model is adopted thus the aerial image on the wafer could be calculated as the superposition of the aerial images contributed by every point source. With the addition of programmable illumination using a micro mirror array like FlexRay, the freeform pixelated source with thousands of pixels is feasible [21]. For example, Wei, et al. recently realized freeform source consisting of 201×201 pixels [22]. The intensities of the source pixels can be tuned continuously. The size of source pixel is proportional to the product of diameter and focal length of micro lens, and inversely proportional to the focal length of condensor lens [22]. The pixelated SO methods benefit from the high degree of optimization freedom. However, the pixelated manipulation dramatically increases the number of the optimization variables, thus posing a challenge on the computational complexity. Recently, Song, et al. first developed a fast pixelated SO approach based on the compressive sensing (CS) [23–25]. Given a raster-scanned aerial image $\vec{I} \in \mathbb{R}^{N^2 \times 1}$ and source pattern $\vec{J} \in \mathbb{R}^{N_s^2 \times 1}$, and letting N and N_s be the lateral dimensions on the image and source planes, the optimal source can be solved from the linear equations of $\vec{I} = \mathbf{I}_{cc} \vec{J}$, where $\mathbf{I}_{cc} \in \mathbb{R}^{N^2 \times N_s^2}$ is the illumination cross coefficient (ICC) that represents the image transformation from source to wafer [23]. The ith column of $\mathbf{I}_{cc}$ indicates the coherent image generated by the ith source pixel, while the jth row of $\mathbf{I}_{cc}$ indicates the image of the jth pixel on the wafer contributed by different source pixels. Since $N_s \ll N$, the above equations belong to an over-determined problem, and its solution is computationally intensive. To reduce the number of the equations, the method in [23] randomly selected M ($M \ll N_s^2 \ll N^2$) monitoring pixels on the wafer. Assume that $\vec{I}_s \in \mathbb{R}^{M \times 1}$ and $\mathbf{I}_{cc}^s \in \mathbb{R}^{M \times N_s^2}$ are the aerial image and the ICC matrix corresponding to the

selected monitoring pixels. Then, the above equations reduce to an under-determined problem:

$$\vec{I}_s = \mathbf{I}_{cc}^s \, \vec{J}. \tag{1}$$

In general, the unique solution of the under-determined problem does not exist. The SO problem is an inverse reconstruction problem with the constraint that the intensities of all source pixels must be non-negative. In addition, the SO aims at a set of objectives so as to improve the process window (PW), image contrast, image fidelity and so on. Thus, SO is usually an ill-posed problem, since numerous source patterns can lead to the same print image. However, CS theory guarantees that the optimal source can be successfully reconstructed from Eq. (1) under the assumption that the source pattern can be sparsely represented on a certain basis [24, 25]. In [23], an l_1-norm reconstruction algorithm, namely linearized Bregman algorithm, was applied to solve for the inverse SO problem. This method can significantly improve the speed by reducing the equations' dimensionality. However, since only a portion of the equations were randomly chosen to reconstruct the source, the useful information borne in other equations was lost. Thus, the reduction of the dimensionality invariably degrades the SO performance and the image fidelity of the lithography system. The dimensionality reduction in CS typically assumes that no a-priori information of the signals is given, except for the sparsity assumption. In source optimization, however, the desired target layout is known and thus, this a-priori information should be used. This paper focuses on precisely exploiting this information.

In particular, a fast SO method based on the adaptive projections is proposed. Firstly, a set of monitoring pixels are selected on the target layout according to a blue noise pattern [26–28]. The blue noise pattern has two merits. First, blue noise is the high-frequency component of random white noise. Most of the monitoring pixels acquired by blue noise will distribute around the edges and corners, which correspond to the high-frequency components of the layout pattern. Thus, the monitoring pixels can extract the layout characteristics more effectively. Second, the blue noise pattern drives more uniform distribution of the monitoring pixels than the random sampling method. These two merits are beneficial to control the image fidelity of the entire layout. In the past, the blue noise sampling patterns have been shown to have clear advantages over random noise patterns in some compressive imaging systems, since the restricted isometry property (RIP) and the coherence of the associated sensing matrices exhibit better properties when blue noise is used [29–34]. Adaptive projections further reduce the dimension of the SO problem, but the compressive measurements still include the information from the original equations. To our best knowledge, we are the first to propose the principle of adaptive projections in CS based on the a-priori information of the target structure in the lithography realm. Traditional CS approaches assume that a-priori information about the original signal is unknown, and use the random projection matrix to compress the signal. Random projections typically apply to most of the sparse signals, but fall short to exploit the structure information of the underlying signal. In the SO problem, the target layout is known, which can be used as the prior knowledge to design the adaptive projection mechanisms. Adaptive projection relies on simple nonlinear transformations, namely random thresholding of the target layout to build the set of projection vectors. It will prove that the adaptive projection is an effective way to compress the underlying signal by capturing the signal structure characteristics. Under the same compression ratio, adaptive projections provide higher accuracy in source reconstruction than that obtained with random projections. The second contribution of this paper is to study and compare the impact of different sparse bases on the performance of the SO algorithm. The bases include the discrete cosine transform (DCT) basis, spatial basis, Haar wavelet basis, and the discrete Fourier transform (DFT) basis. The simulations show that the proposed adaptive projection CS approach attains superior imaging performance over the traditional CS method. Furthermore, the DCT, spatial and Haar wavelet bases outperform the DFT basis by enforcing the sparsity of the source pattern and achieve better imaging performance.

The remainder of this paper is organized as follows. The principle of the adaptive projection CS is provided and derived in Section 2. The SO framework based on the adaptive projection CS is proposed in Section 3. Simulations and comparisons using different CS methods and sparse bases are presented in Section 4. Conclusions are provided in Section 5.

2. Adaptive projection CS

A signal $\vec{X} \in \mathbb{R}^{N \times 1}$ is said to be sparse if it can be represented exactly, or at least accurately, as a weighted superposition of a small subset of basis functions from a fixed basis set. Let $\vec{X} = \boldsymbol{\Psi}\vec{\Theta}$, where $\boldsymbol{\Psi} = [\vec{\psi}_1, \vec{\psi}_2, \dots, \vec{\psi}_N] \in \mathbb{R}^{N \times N}$ is called the sparse basis, and $\vec{\Theta} \in \mathbb{R}^{N \times 1}$ is the corresponding coefficients. The signal is K-sparse if $\vec{\Theta}$ only has $K \ll N$ non-zero elements. Suppose we observe a small set of compressive measurements of $\vec{X}$:

$$\vec{Y} = \boldsymbol{\Phi}\vec{X} = \boldsymbol{\Phi}\boldsymbol{\Psi}\vec{\Theta}, \tag{2}$$

where $\boldsymbol{\Phi} = [\vec{\phi}_1, \vec{\phi}_2, \dots, \vec{\phi}_L]^T \in \mathbb{R}^{L \times N}$ is the projection matrix with $L \ll N$. CS theory enables the reconstruction of the K-sparse signal $\vec{X}$ from its measurements $\vec{Y}$. It is natural to ask what is the minimum number of measurements needed to successfully reconstruct $\vec{X}$.

Consider first the most general case where the number and locations of the non-zero basis components in $\vec{X}$ are not known a-priori. The signal reconstruction problem in this case is solved by the traditional CS [24, 25]. It has been proved that if $\boldsymbol{\Psi}$ is incoherent to $\boldsymbol{\Phi}$ and the rows of $\boldsymbol{\Phi}$ are chosen randomly, then $\vec{X}$ can be successfully recovered when $L = C \times K \times \log N \ll N$, where $C \geq 1$ is an oversampling factor. The random samples in $\boldsymbol{\Phi}$ can be Gaussian distributed, or taken from the binary random values. The mutual-coherence between $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ can be evaluated by

$$\mu = \max\left\{\left|\left\langle \vec{\phi}_i, \vec{\psi}_j \right\rangle\right|^2\right\}, \quad i = 1, 2, \dots, L \ \text{and} \ j = 1, 2, \dots, N. \tag{3}$$

Traditional CS is thus universal where no assumption is made on the original signal and the sparsity cardinality.

Consider next the most restricted case where the number and locations of the non-zero basis components in $\vec{X}$ are known. The set of K non-zero basis components are $\Upsilon = \{\vec{\psi}_{l(1)}, \vec{\psi}_{l(2)}, \dots, \vec{\psi}_{l(K)}\}$, where $l(i)$ refers to the location of the ith non-zero component. The reconstruction problem exploits the orthogonality principle and thus reduces to making the projection vector $\vec{\phi}_i$ equal to the basis vector $\vec{\psi}_{l(i)}$. Then, only K compressive measurements are sufficient to exactly recover $\vec{X}$. The mutual-coherence metric for this case is divided into two parts, such that

$$\mu_{\Upsilon} = \max_{\vec{\psi}_j \in \Upsilon}\left\{\left|\left\langle \vec{\phi}_i, \vec{\psi}_j \right\rangle\right|^2\right\} = 1, \quad \mu_{\overline{\Upsilon}} = \max_{\vec{\psi}_j \in \overline{\Upsilon}}\left\{\left|\left\langle \vec{\phi}_i, \vec{\psi}_j \right\rangle\right|^2\right\} = 0, \tag{4}$$

where $\overline{\Upsilon}$ is the complementary set of Υ. The vectors of $\vec{\phi}_i$ and $\vec{\psi}_j$ are normalized to have unit energy.

The mutual-coherence matrices in Eqs. (3) and (4) take on distinct goals. If a-priori knowledge of the signal is unknown, what we can only do is to minimize μ. On the other hand, if some side-information is available for reconstruction, we should try to design a sensing mechanism, by which μ is shaped in some fashion as Eqs. (4). This idea is given by the following design rule of the projection matrix. Given some side-information about the original signal, the CS projection matrix should maximize the difference between the μ_{Υ} and $\mu_{\overline{\Upsilon}}$ in Eq (4).

Now, consider a special case that we have the approximate (but not exact) observation of the original signal $\vec{X}$ in hand. Suppose the observation of $\vec{X}$ is represented as $\vec{S} = \vec{X} + \vec{N}$, where

$\vec{N} \in \mathbb{R}^{N \times 1}$ is a noise vector. The elements of $\vec{N}$ are the independent identical random variables obeying the distribution of $\mathcal{N}(0, \sigma_X^2)$. $\mathcal{N}(0, \sigma_X^2)$ indicates the Gaussian distribution with zero-mean and variance of σ_X^2. It is reasonable to ask if the signal reconstruction performance can be improved in virtue of the observation $\vec{S}$. To answer this question, we propose the adaptive projection CS method, where each projection vector $\vec{\phi}_i$ is generated by a thresholding operator. The jth element of $\vec{\phi}_i$ is

$$\vec{\phi}_{ij} = \frac{\mathrm{sgn}(\vec{S}_j - \Lambda_{ij})}{\sqrt{N}}, \tag{5}$$

where $\vec{S}_j$ is the jth element of $\vec{S}$, the threshold $\Lambda_{ij} \sim \mathcal{N}(0, \sigma_\Lambda^2)$ is a sample from the Gaussian random variable, and $\mathrm{sgn}(\cdot)$ is the sign operator. Note that $\vec{\phi}_{ij} \in [-1/\sqrt{N}, 1/\sqrt{N}]$ is a quantized, randomly thresholding version of $\vec{S}_j$. Random thresholding, also known as dithering, is extensively used in quantization theory and in particular the binary representations of signals [30]. It provides a mechanism to build incoherent "random" vectors that extract salient spectral characteristics of the underlying signal $\vec{X}$. In addition, thresholding is simple, thus the extra implementation cost is ignorable.

Next, the design rule is used to quantify the goodness of the adaptive projection method. In the Appendix, we prove that the adaptive projection makes the average values of μ_Υ and $\mu_{\overline{\Upsilon}}$ in Eq. (4) satisfy the following properties:

$$\bar{\mu}_\Upsilon \triangleq \max_{\vec{\psi}_j \in \Upsilon} E\left\{ \left| \langle \vec{\phi}_i, \vec{\psi}_j \rangle \right|^2 \right\} > \frac{\|\vec{X}\|^2}{N K^2 \theta_{max}^2}, \quad \bar{\mu}_{\overline{\Upsilon}} \triangleq \max_{\vec{\psi}_j \in \overline{\Upsilon}} E\left\{ \left| \langle \vec{\phi}_i, \vec{\psi}_j \rangle \right|^2 \right\} \approx 0. \tag{6}$$

where $E\{\bullet\}$ represents the mathematical expectation, and θ_{max} is the maximum element in the coefficient vector $\vec{\Theta}$. These properties mean that the adaptive projection method can separate μ_Υ and $\mu_{\overline{\Upsilon}}$ in a statistical sense. Thus, the adaptive projection method satisfies the design rule, and benefits in improving the performance of signal reconstruction.

3. SO framework based on adaptive projection CS

Given the target layout $\mathbf{Z}$, we first use the similar method in [6] to locate its critical regions $\mathbf{Z}_m$. As shown in Fig. 2, the critical regions include the inner (green) and outer (red) margins of the mask features, as well as a set of pixels in non-pattern regions (blue). The pixels selected in the non-pattern regions are helpful to suppress the side-lobe printings induced by sub-resolution assist features (SRAF) during the source optimization. The distance between the non-pattern critical regions and the layout features can be adjusted by the user-defined parameters, so as to remove the side-lobe printings at different locations. Afterwards, M monitoring pixels are selected in the critical regions according to the blue noise pattern. We compare $\mathbf{Z}_m$ with a blue noise pattern $\mathbf{B}$ to determine the positions of the monitoring pixels. If and only if $\alpha \mathbf{Z}_m(i, j) + \beta \geq \mathbf{B}(i, j)$, the pixel at (i, j) is chosen as the monitoring pixel, where α and β are the parameters to control the number of monitoring pixels. The parameter α is used to control the number of monitoring pixels covered by the features of $\mathbf{Z}_m$, while β is used to control the number of monitoring pixels surrounding the features of $\mathbf{Z}_m$. In the following simulations, all of the monitoring pixels are selected under the features of $\mathbf{Z}_m$. So, β is always set to be zero, and α is adjusted to obtain the desired number of monitoring pixels. The blue noise sampling is characterized by the uniform, random, aperiodic distribution. While random noise patterns contain frequency components across all of the spectra, blue noise patterns exclude low frequency components [35–37]. Figures 3(b) and 3(c) compare the monitoring pixels in blue noise selection mode and random selection mode used in [23]. The monitoring pixels are represented by the red circles. Note how the blue noise pattern prevents samples from

Vertical Line-Space   **Horizontal Block**

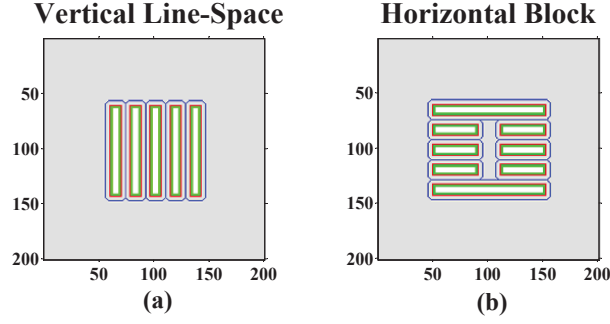(a)   (b)

Fig. 2. The critical regions for the (a) vertical line-space layout pattern and (b) horizontal block layout pattern.
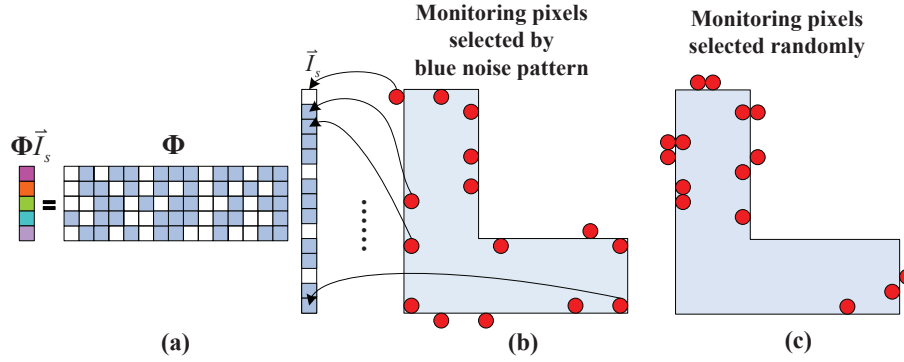

Fig. 3. The generation of monitoring pixels and compressive measurements.

being adjacent to each other. It is obvious that the blue noise samplings spread over the layout boundaries, and keep the layout features better than the random selection mode.

Based on the selected monitoring pixels, the SO problem can be formulated as an l_1-norm reconstruction problem:

$$\hat{\vec{\Theta}} = \min_{\vec{\Theta}} ||\vec{\Theta}||_1 \quad \text{subject to} \quad \mathbf{\Phi}\vec{Z}_s = \mathbf{\Phi}\vec{I}_s = \mathbf{\Phi}\mathbf{I}_{cc}^s\vec{J} = \mathbf{\Phi}\mathbf{I}_{cc}^s\mathbf{\Psi}\vec{\Theta}, \tag{7}$$

where $\vec{Z}_s \in \mathbb{R}^{M \times 1}$ and $\vec{I}_s \in \mathbb{R}^{M \times 1}$ are the vectors of target layout and aerial image corresponding to the monitoring pixels, respectively. $\mathbf{I}_{cc}^s \in \mathbb{R}^{M \times N_s^2}$ is the ICC matrix. The source pattern can be represented by $\vec{J} = \mathbf{\Psi}\vec{\Theta}$, where $\mathbf{\Psi} \in \mathbb{R}^{N_s^2 \times N_s^2}$ and $\vec{\Theta} \in \mathbb{R}^{N_s^2 \times 1}$ are the sparse basis and the corresponding coefficients, respectively. $\mathbf{\Phi} \in \mathbb{R}^{L \times M}$ $(L < M)$ is the adaptive projection matrix. The linear constraint of $\mathbf{\Phi}\vec{Z}_s = \mathbf{\Phi}\vec{I}_s$ in Eq. (7) enforces the aerial image equal to the target layout pattern on the monitoring pixels, therefore the contrast and normalized image log slope (NILS) of aerial image are enhanced during SO process. This is helpful to improve the PW. In addition, the pattern error (PE) and edge placement error (EPE) of the print image are reduced under the linear constraint.

Next, we will describe how to design $\mathbf{\Phi}$ in Eq. (7). Define $\tilde{\mathbf{\Psi}} = \mathbf{I}_{cc}^s\mathbf{\Psi}$ as a new sparse basis, and the optimal source is formulated as $\hat{\vec{J}} = \mathbf{\Psi}\hat{\vec{\Theta}}$. Then, the aerial image corresponding to $\hat{\vec{J}}$ should approximate the target layout, such that $\vec{Z}_s \approx \mathbf{I}_{cc}^s\hat{\vec{J}} = \tilde{\mathbf{\Psi}}\hat{\vec{\Theta}}$. According to Section 2, we can regard $\tilde{\mathbf{\Psi}}\hat{\vec{\Theta}}$ and $\vec{Z}_s$ as the signal $\vec{X}$ and its observation $\vec{S}$, respectively. In the SO problem,

$\vec{Z}_s$ is known a-priori, thus the (i, j)th element of $\mathbf{\Phi}$ is designed as

$$\vec{\phi}_{ij} = \frac{\text{sgn}(\vec{Z}_{s,j} - \Lambda_{ij})}{\sqrt{M}},\tag{8}$$

where $\vec{Z}_{s,j}$ is the jth element in vector $\vec{Z}_s$. As shown in Fig. 3(a), the adaptive projection further compresses the dimension of the equations, and reduces the computational complexity of the SO algorithm. In addition, the compressed equations in Eq. (7) hold the information from all of the original equations in Eq. (1). Thus, the adaptive projection method is able to improve the imaging performance compared to the method in [23].

To prove the benefit of the adaptive projection matrix in Eq. (8), we consider the mutual-coherence metrics in the following. It can be shown that $\bar{\mu}_\Upsilon > \|\vec{Z}\|^2/(NK^2\theta_{max}^2)$, where the sparse basis is defined as $\tilde{\mathbf{\Psi}} = \mathbf{I}_{cc}^s \mathbf{\Psi}$. The derivation of this inequality is the same as the first inequality of Eq. (6). The proof is provided in the Appendix. On the other hand, the second approximate equality of Eq. (6) does not hold rigorously. That is because the derivation of the approximate equality is based on the orthogonality of the sparse basis. However, $\tilde{\mathbf{\Psi}}$ is a non-orthogonal basis. Nevertheless, we can prove the property of adaptive projection matrix by numerical simulations. Table 1 compares the average value of mutual-coherence metrics between the random projection method and the proposed adaptive projection method. In this simulation, we implement the random and adaptive projection matrices for 100 times, and then calculate the average mutual-coherence metrics $\bar{\mu}_\Upsilon$ and $\bar{\mu}_{\overline{\Upsilon}}$. The average values of $\bar{\mu}_\Upsilon$ are much larger than 1, since $\tilde{\mathbf{\Psi}}$ is not a standard basis. It is obvious that the adaptive projection method further increases the difference between $\bar{\mu}_\Upsilon$ and $\bar{\mu}_{\overline{\Upsilon}}$ in contrast to the random projection method. Thus, the adaptive projection method outperforms random projection method according to the design rule described in Section 2.

Table 1. The comparison of average mutual-coherence metrics between the random and adaptive projection methods.

Pattern		Vertical Line-Space		Horizontal Block	
Criteria		Random Projections (M=300, L=25)	Adaptive Projections (M=300, L=25)	Random Projections (M=300, L=25)	Adaptive Projections (M=300, L=25)
$\bar{\mu}_\Upsilon$		384.560	528.070	285.345	320.565
$\bar{\mu}_{\overline{\Upsilon}}$		1.370	0.730	1.198	0.686

In the following, the linearized Bregman algorithm is used to solve for the SO problem, since it is computational efficient and can enhance the contrast of the acquired image [38, 39]. During the optimization procedure, the linearized Bregman algorithm relaxes the equality constraint in Eq. (7) to minimize the distance between the aerial image and target layout pattern. In this way, the proposed SO algorithm tries to suppress the intensity of side-lobe printing in the non-pattern regions, rather than retaining the rigorous equality constraint.

4. Simulation and analysis

This section provides a set of simulations of the proposed SO algorithm and compares it to the random projection CS method and Song's method in [23]. In addition, the impact of different sparse bases on the SO performance is also discussed.

4.1. Simulations of SO method based on adaptive projection CS

Figure 4 illustrates the simulations based on an immersion lithography system and a line-space layout pattern with CD=45nm. The wavelength of illumination is 193nm, the NA on the wafer side is 1.2, the demagnification factor of the lithography system is 4, and the refractive index of the immersion medium is 1.44. The DCT basis is used to sparsely represent the source pattern. Before the optimization, the initial source is a circular illumination with partial coherence factor of $\sigma = 1$. The initial source fills up the entire pupil and is normalized to have unit energy. The first row in Fig. 4 shows the optimized source patterns with unit energy. The second row shows the corresponding print images on the wafer. Based on the constant threshold resist model, the print image is calculated as $\Gamma\{\mathbf{I} - t_r\}$, where $\mathbf{I}$ is the aerial image, $\Gamma\{\cdot\}$ is a hard threshold function, and the threshold value is $t_r = 0.25$. The figure also presents the PEs and EPEs of the print images, as well as the NILSs and contrasts of the aerial images. In this paper, the PE is defined as the square of the Euclidean distance between the target layout and the print image. The EPE indicates the error of the actual printed edge position with respect to the target. In this paper, the EPE is defined as the average value of EPEs along the entire boundaries of layout including the corners. The contrast is calculated along the red dashed lines on the aerial images. The NILS measures the slope of the aerial image on a certain contour normalized by the CD, which is defined as

$$\text{NILS} = \frac{\text{CD}}{\mathbf{I}_{\text{con}}} \times \frac{d\mathbf{I}}{dx}\bigg|_{\mathbf{I}_{\text{con}}}, \tag{9}$$

where CD is the critical dimension, $\mathbf{I}_{\text{con}}$ is the aerial image intensity on the contour, and $d\mathbf{I}/dx$ is the differential of the intensity with respect to the length perpendicular to the contour. In this paper, the NILS is calculated along the contour of the target layout. In order to measure the NILS on all points on the contour, we modify the above definition as

$$\text{NILS} = \frac{1}{L_c} \left\{ \oint_c \frac{\mathbf{CD}}{\mathbf{I}_{\text{con}}} \times \frac{d\mathbf{I}}{dx}\bigg|_{\mathbf{I}_{\text{con}}} dc \right\}, \tag{10}$$

where $\oint_c \bullet dc$ is the integral along the contour, and L_c is the length of the contour.
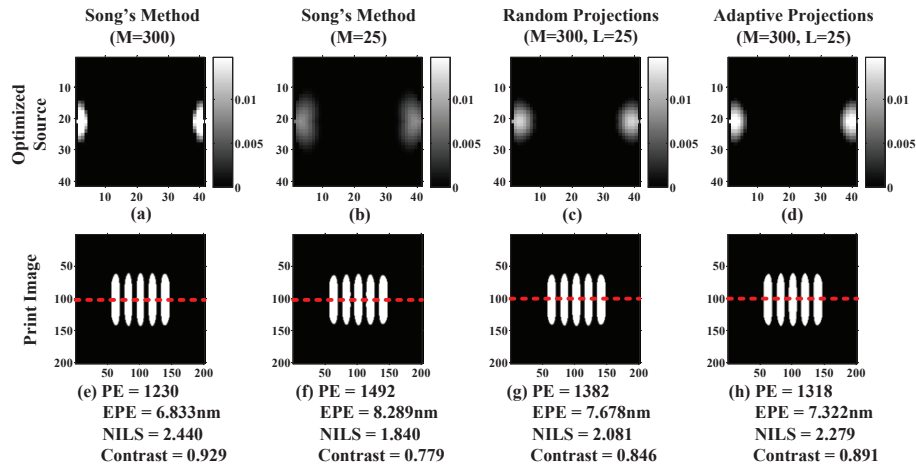

Fig. 4. The simulations of different SO methods based on the vertical line-space layout pattern.

In Fig. 4, the first and second columns illustrate the simulations using Song's method with 300 ($M = 300$) and 25 ($M = 25$) monitoring pixels, respectively. It is noticed that the reduction of monitoring pixel number will increase the pattern error. That is because Song's method skips

those equations in Eq. (1) that are not supported by the monitoring pixels. The third column is the simulation using random projection CS method. The formulation of the SO problem is the same as Eq. (7), but the elements in $\mathbf{\Phi}$ are independent and identical Bernoulli random variables. Each row of the projection matrix is normalized to have unit energy. In this simulation, we first chose 300 ($M = 300$) monitoring pixels on the layout, and then compress the dimensionality down to 25 ($L = 25$). The resulting image fidelity falls in between Song's methods with 300 and 25 monitoring pixels. The reason is that the randomly compressive measurements in Eq. (7) still include the information from the original 300 monitoring pixels, thus the SO performance is improved by only 25 equations. The fourth column shows the simulation using the proposed adaptive projection CS method, where the equations of 300 ($M = 300$) monitoring pixels are projected into 25 ($L = 25$) dimensional space. As analyzed in Section 2, the adaptive projection leads to higher efficiency of compression by extracting the spectral characteristics of the underlying signal. Thus, the adaptive projection method can further improve the imaging performance compared to the random projection method. Figure 5 gives an intuitive comparison between the (a) random and (b) adaptive projection matrices used in the above simulations. Both of the matrices in Figs. 5(a) and 5(b) are binary matrices with dimension of 25×300. The dimension in the x-axis is equal to $M = 300$, while the dimension in the y-axis is equal to $L = 25$. The white and black regions represent the elements with values of $1/\sqrt{M}$ and $-1/\sqrt{M}$ in the projection matrix $\mathbf{\Phi}$, respectively. Each column of the adaptive projection matrix is generated by randomly thresholding the corresponding monitoring pixel values. As shown is Figs. 5(b), the elements of adaptive projection matrix are more likely to be positive if the corresponding monitoring pixel value is equal to 1. Otherwise, the elements of adaptive projection matrix are more likely to be negative. Thus, the adaptive projection matrix appears to extract some structure information that is related to the underlying layout data.
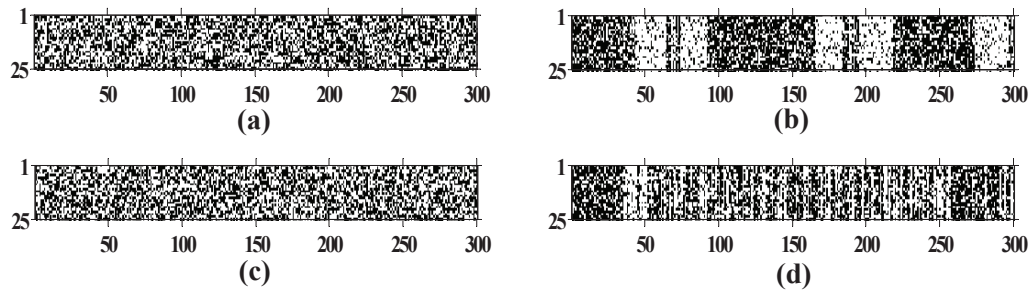

Fig. 5. Random and adaptive projection matrices ($M = 300$, $L = 25$) used for two different layout patterns.

Figures 6(a) and 6(b) are the convergence curves of the PE and contrast for different SO methods. It is observed that the convergence curves of Song's method changes a lot in the first few iterations. This can be explained as follows. As mentioned above, the initial source of the SO method is set to be a circular illumination filling up the entire pupil. In the first few iterations, Song's method shuts down most of the source pixels. In addition, some negative source pixels come out, which are automatically set to be zeros by the optimization algorithm. So, the source pattern is dramatically changed during the first few iterations, and the pattern errors are rapidly increased to form large jumps in the convergence curves. After another several iterations, some off-axis bright source pixels appear to form the dipole illumination, and the pattern errors rapidly drop. In contrast to Song's method, other SO methods provide much smoother convergence curves. However, we still observe a few kinks in the convergence curves due to the oscillation of the optimization algorithm. The oscillation may be induced by the operations to shut down the negative source pixels. Figure 7(a) compares the overlapped PWs obtained by different

methods, where the *x* and *y* axes represent the depth of focus (DOF) and exposure latitude (EL), respectively. The locations to measure the PWs are illustrated in Fig. 8. As shown in Fig. 8(a), we first calculate the PWs at locations ① and ②. Then, their overlapped PW is used to evaluate the process robustness of lithography system. It is observed that with the same compression ratio, the adaptive projection method results in larger PW than the random projection method. But, Song's method with $M = 300$ attains the largest PW, since it uses much more equations to recover the source pattern.
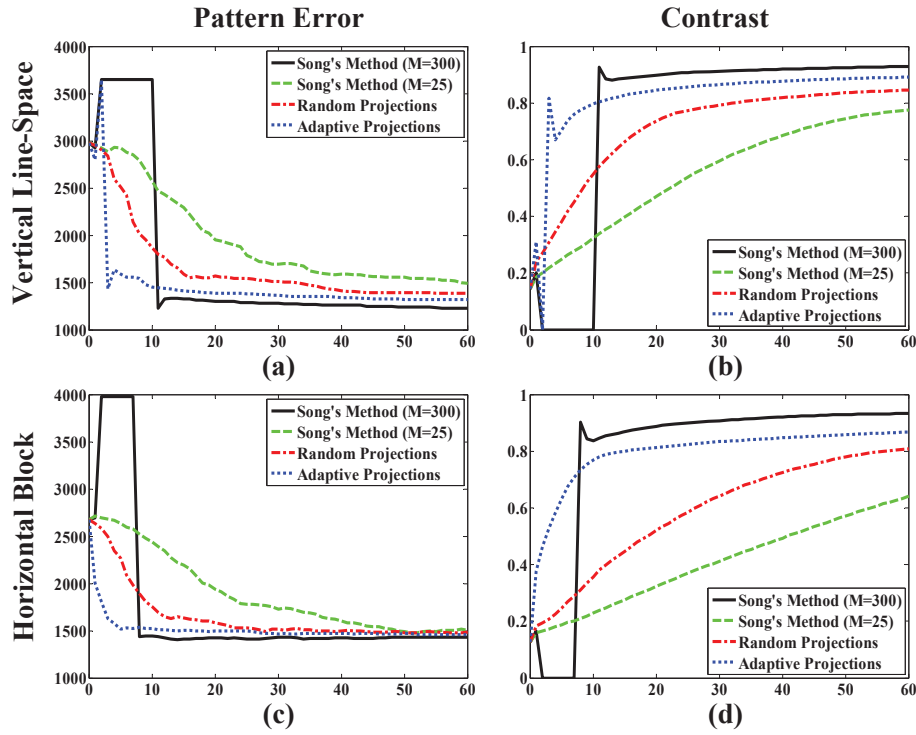
Fig. 6. The convergence of the PE and contrast for different SO methods, where (a) and (b) are the convergence curves for the vertical line-space layout pattern, while (c) and (d) are the convergence curves for the horizontal block layout pattern.
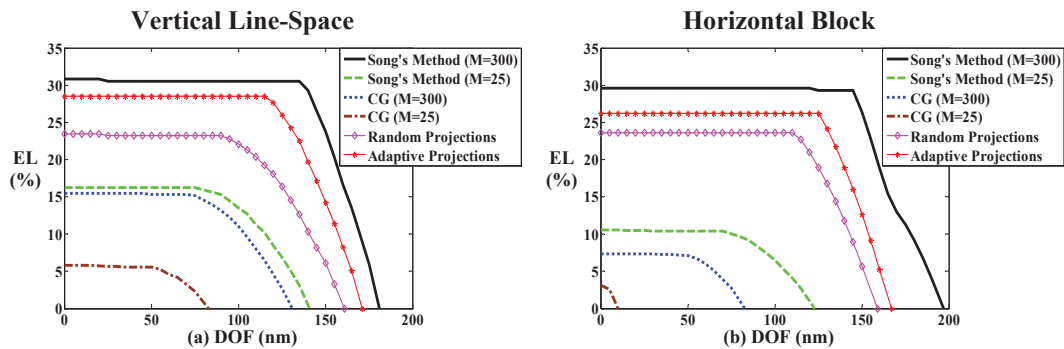

Fig. 7. The overlapped PWs obtained by different SO methods based on the (a) vertical line-space layout pattern and (b) horizontal block layout pattern.

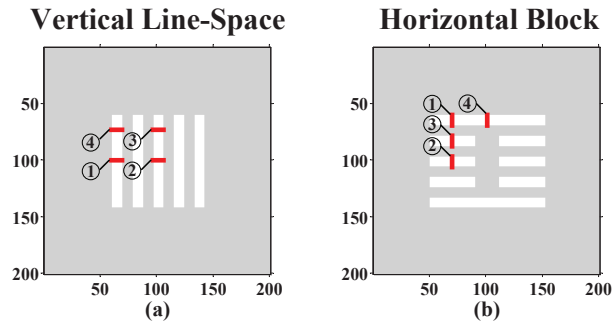**Vertical Line-Space**　　**Horizontal Block**

Fig. 8. The locations to measure the PWs for the (a) vertical line-space layout pattern and (b) horizontal block layout pattern.

The top half of Table 2 provides the average values of PEs, EPEs, NILSs, contrasts and runtimes of different SO methods by implementing the simulations for 100 times. All programming are implemented by Matlab and the computations are carried out on an Intel(R) Xeon(R) E7-4830 CPU(4 cores), 2.13GHz, 128GB of RAM. In contrast to Song's method with $M = 300$, the adaptive projection method achieves 14% speedup. The reduction of runtime is not proportional to the compression of dimensionality, since the algorithm codes include some other subsidiary functions, such as the symmetrization of the source pattern and so on. On the other hand, the runtime of the adaptive projection method is slightly longer than Song's method with $M = 25$ and the random projection method. The increment of runtime is attributed to the calculation of the adaptive projection matrix as described in Eq. (8). In order to further demonstrate the computational efficiency of the adaptive projection method, we compare it to Song's method, which uses significantly more monitoring pixels. In this simulation, Song's method selects 1000 ($M = 1000$) monitoring pixels on the layout, while the adaptive projection method compresses the dimensionality of the equations from 1000 to 25 ($L = 25$). For Song's method, PE=1173, EPE=6.518nm, NILS=2.528, contrast=0.959, and the runtime is 0.158s. For the adaptive projection method, PE=1262, EPE=7.010nm, NILS=2.418, contrast=0.917, and the runtime is 0.092s. Compared to Song's method, the adaptive projection method achieves 42% speedup. That means the improvement in speed of the adaptive projection method is more obvious when the number of monitoring pixel increases. It is known that the real layout pattern of integrated circuit is much larger than the test pattern in this paper, and we need to set a large number of monitoring pixels on the layout. So, the adaptive projection method is expected to be much faster than Song's method in practical applications. In summary, the proposed adaptive projection method outperforms the Song's method and random projection method in the imaging performance, while the runtimes of these methods are comparable under the same dimensionality.

Figure 9 shows the simulations of the SO methods based on another horizontal block pattern. Figures 5(c) and 5(d) illustrate the random and adaptive projection matrices used in these simulations. Figures 6(c) and 6(d) provide the convergence of the PE and contrast, respectively. The contrast is calculated along the red dashed lines on the aerial images. Figure 7(b) compares the overlapped PWs obtained by different SO methods. In particular, we first calculate the PWs at locations ① and ② in Fig. 8(b). Then, the overlap of these two PWs is presented in Fig. 7(b). The bottom half of Table 2 provides the average values of PEs, EPEs, NILSs, contrasts and runtimes of different SO methods by implementing the simulations for 100 times. The simulations of the horizontal blocks also prove the superiority of the adaptive projection method.

In order to further verify the proposed SO method, we compare it with the fast SO method

Table 2. The average PEs, EPEs, NILSs, contrasts and runtimes of different SO methods.

| Vertical Line-Space				
Criteria	Song' Method (M=300)	Song' Method (M=25)	Random Projections (M=300, L=25)	Adaptive Projections (M=300, L=25)
PE	1238	1628	1514	1339
EPE (nm)	6.811	9.057	8.453	7.376
NILS	2.438	1.702	1.771	2.217
Contrast	0.930	0.707	0.743	0.871
Runtime (s)	0.105	0.084	0.089	0.090
Horizontal Block				
Criteria	Song' Method (M=300)	Song' Method (M=25)	Random Projections (M=300, L=25)	Adaptive Projections (M=300, L=25)
PE	1431	1625	1598	1488
EPE (nm)	6.970	7.898	7.750	7.247
NILS	2.255	1.522	1.623	2.033
Contrast	0.932	0.652	0.673	0.859
Runtime (s)	0.104	0.086	0.088	0.089
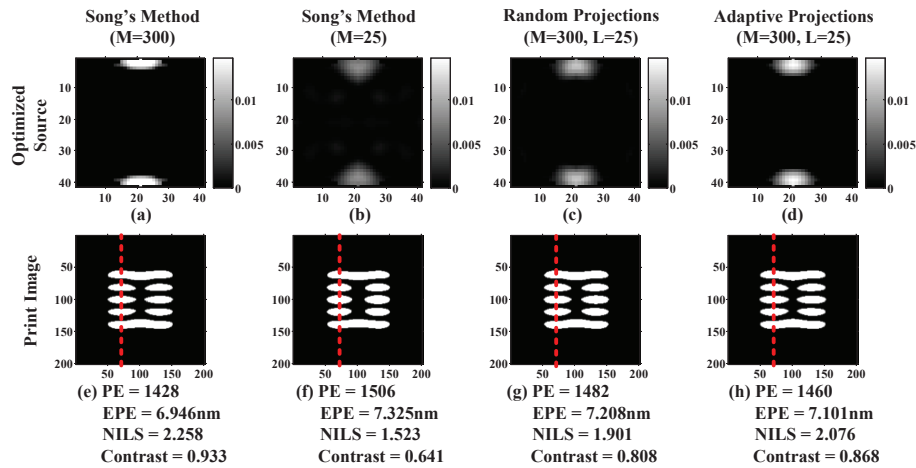
Fig. 9. The simulations of different SO methods based on the horizontal block layout pattern.

based on conjugate-gradient (CG) algorithm developed in [6]. Figure 10 shows the SO results using the CG algorithm. Figures 10(a) and 10(b) show the optimized source patterns for vertical line-space layout pattern with M=300 and 25, respectively. Figures 10(c) and 10(d) show the optimized source patterns for horizontal block layout pattern with M=300 and 25, respectively. The bottom row shows the print images corresponding to the optimized source patterns in the top row. The overlapped PWs obtained by the CG method for both layout patterns are illustrated in Figs. 7(a) and 7(b), respectively. Table 3 provides the average values of PEs, EPEs, NILSs, contrasts and runtimes of CG method by implementing the simulations for 100 times. It is observed that the proposed SO method achieves superior imaging performance over the CG algorithm.
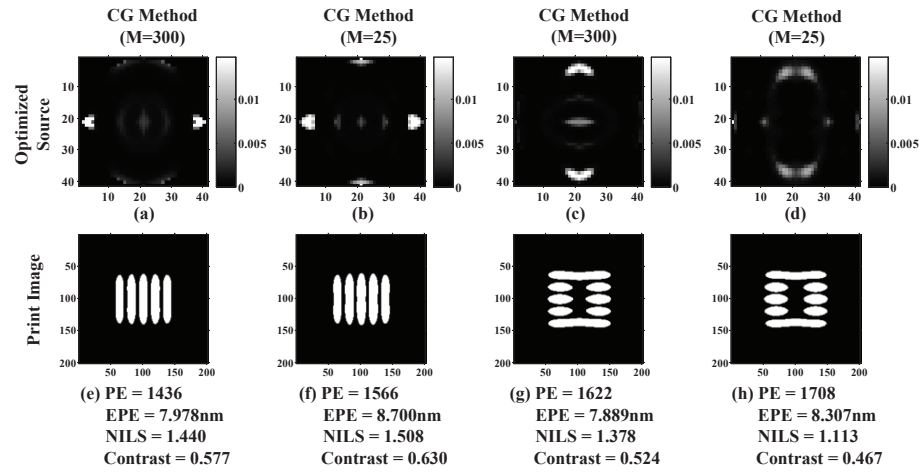
Fig. 10. The simulations of CG method based on the vertical line-space layout pattern and horizontal block layout pattern.

Table 3. The average PEs, EPEs, NILSs, contrasts and runtimes of CG methods.

Pattern	Vertical Line-Space		Horizontal Block	
Criteria	(M=300)	(M=25)	(M=300)	(M=25)
PE	1416	1678	1548	1829
EPE (nm)	7.866	9.325	7.530	8.898
NILS	1.411	1.308	1.288	1.232
Contrast	0.567	0.547	0.486	0.466
Runtime (s)	1.635	1.603	1.625	1.618

The aforementioned simulations show that the proposed SO method can effectively improve the imaging performance. However, the use of the SO method alone may not improve the PWs at the line ends, which are critical to retain the image fidelity. Next, we demonstrate that the PWs at the line ends can be effectively improved by jointly utilizing the proposed SO method in concert with mask optimization, which is also called source mask optimization (SMO). SMO exploits the synergy in the joint optimization of source and mask, and can results in superior imaging performance over the individual source or mask optimization methods [14]. Figure 11 shows the simulations of SMO method. From left to right, Fig. 11 shows the optimized sources, masks and the corresponding print images. The first and second rows illustrate the simulations of the proposed SMO method for the line-space layout pattern and horizontal block layout pattern, respectively. In these simulations, we first optimize the source patterns using the proposed method in this paper, then optimize the mask patterns using the method described in [14]. Figure 12 compares the overlapped PWs obtained by the proposed SO method and the SMO method. In particular, we first calculate the PWs at locations ①, ②, ③ and ④ in Fig. 8. Then, the overlap of these four PWs is presented in Fig. 12. The overlapped PWs takes into account the imaging performance at different locations including the line ends. The overlapped PW of SO method disappears in Fig. 12(a), while the proposed SMO method can effectively extend the overlapped PWs. It is also noted that the NILS and contrast obtained by the proposed SMO method are worse than those of the SO method. That is because the mask optimization method in [14] aims at preserving the image fidelity of print image on the focal plane, rather than improving the slope of the aerial image on the layout boundary. In future work, we will improve the SMO method to further enhance the NILS and contrast.

The third and fourth rows of Fig. 11 illustrate the simulations of the traditional SMO method, where both of the source and mask are optimized using the methods in [14]. The PEs, EPEs, NILSs, contrasts and runtimes are all presented in Fig. 11. It is shown that the proposed SMO method results in better NILS and contrast compared to the traditional SMO method. What is more important, the proposed SMO method is much faster than the traditional SMO method. The reason is that the computational complexity of the proposed SO algorithm is much lower than the traditional SO algorithm. In addition, the proposed SO algorithm results in less bright source pixels than the traditional algorithm, which makes the following mask optimization faster. Figure 12 also shows the overlapped PWs of the traditional SMO method. In general, the proposed SMO method is more beneficial to extend the PWs, since the optimized source pattern is sparser in the spatial domain compared to the traditional SMO method. In the future, we will study the fast mask optimization method based on the adaptive projection CS to further speed up the SMO algorithm.
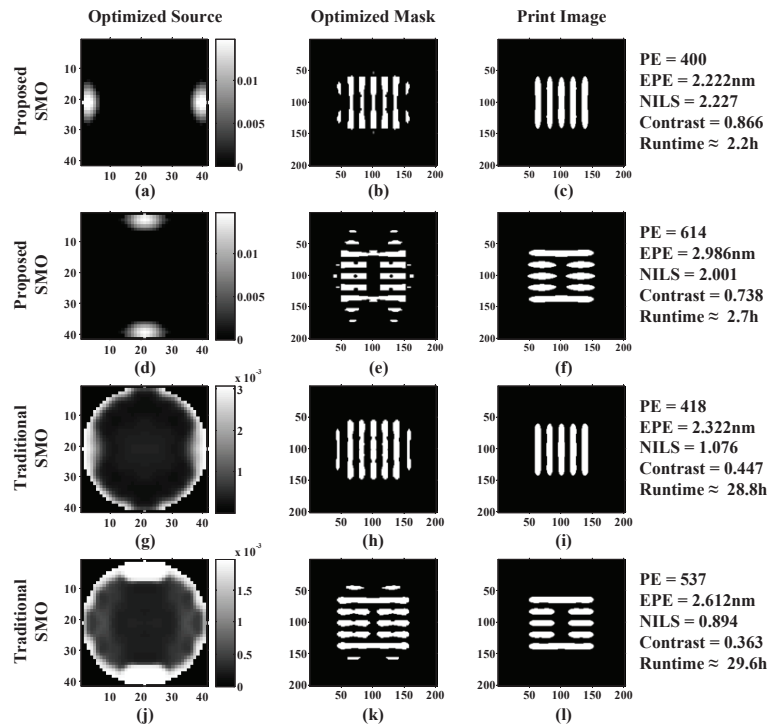

Fig. 11. The simulations of SMO methods.

At the end of this section, we present the simulations of SO and SMO methods based on line-space layout pattern at 14nm technology node. The half-pitch at 14nm technology node is 28nm [40]. Assume the triple patterning technique is used to print the dense line-space features. The single mask for one patterning process is illustrated in Fig. 13(b), where the CD of the lines is 28nm, and the duty ratio is 1:5. The top row of Fig. 13 shows the simulations of the proposed SO method. The bottom row shows the simulations of SMO method. The SMO method first optimizes the source, then mask, then source and mask again. It is obvious that the SMO method achieves superior image fidelity of print image over the SO method due to the higher degrees of optimization freedom. However, the NILS and contrast are degraded a little, while the reason has been explained in the last paragraph.
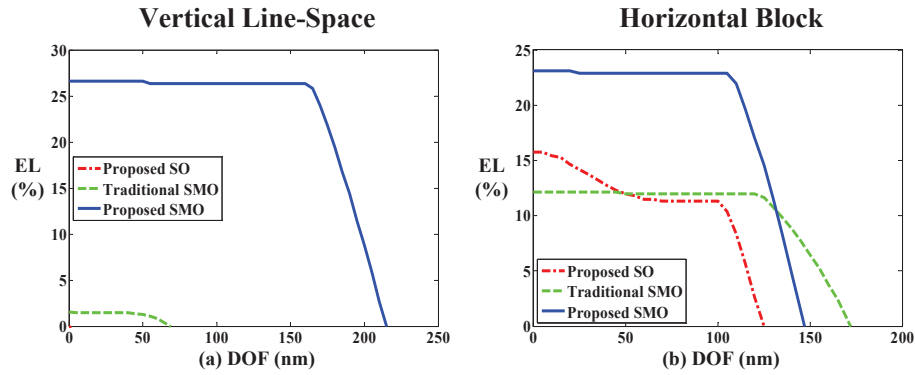
Vertical Line-Space

Horizontal Block

Fig. 12. The comparison of overlapped PWs obtained by SO and SMO methods based on the (a) vertical line-space layout pattern and (b) horizontal block layout pattern.
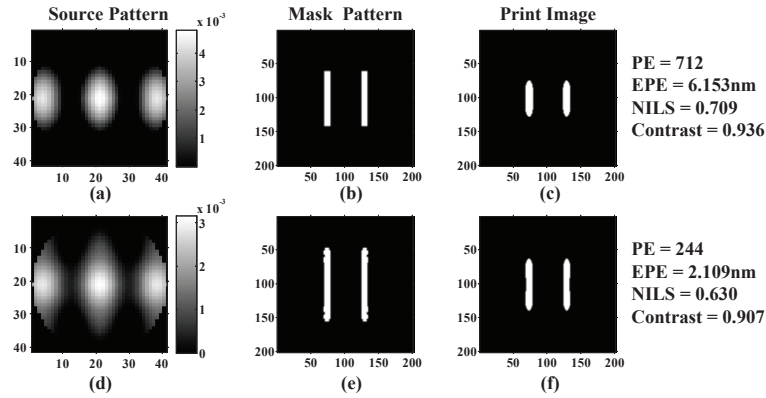

Fig. 13. The simulations of SO and SMO methods using line-space layout pattern at 14nm technology node.

4.2. Simulations using different sparse bases

In the following, we compare the impact of four different sparse bases on the SO performance. The bases under consideration includes the DCT basis, spatial basis, Haar wavelet basis, and DFT basis. Table 4 summarizes the average values of PEs, EPEs, NILSs and contrasts for different sparse bases by implementing the simulations for 100 times. The adaptive projection method is used to optimize the sources. It is noted that the spatial basis matrix Ψ is just an identity matrix. It is shown that the imaging performances obtained by the DCT, spatial and Haar wavelet bases are similar to each other, but better than the DFT basis. The reason is that the optimal source usually consists of multi-pole pattern, which can be represented more sparsely on the DCT, spatial and Haar wavelet bases. Figure 14 illustrates the absolute values of the source coefficients on different bases in the logarithmic scale. In Fig. 14, all of the coefficients are plotted in descending order. In addition, the coefficients smaller than 10^{-20} are ignored and not shown in the figure. According to Fig. 14, the spatial basis has very sparse coefficient distribution. The DCT and Haar wavelet bases also have good energy compaction properties. However, the DFT coefficients descend very slowly. Therefore, the DCT, spatial and Haar wavelet bases are good choices for the source representation in our proposed SO method.

It is noted that the freeform pixelated source patterns in the above simulations can be implemented by current programmable illumination systems. In addition, the sparsity of source

Table 4. The average PEs, EPEs, NILSs and contrasts of the adaptive projection method using different sparse bases.

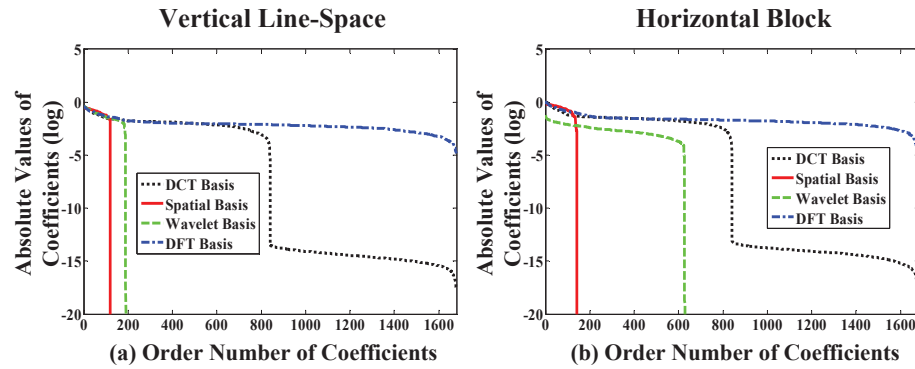| Pattern	Vertical Line-Space				Horizontal Block			
Basis	DCT	Spatial	Haar	DFT	DCT	Spatial	Haar	DFT
PE	1339	1339	1342	1366	1488	1488	1488	1509
EPE (nm)	7.376	7.376	7.403	7.543	7.247	7.246	7.261	7.322
NILS	2.217	2.217	2.214	2.162	2.033	2.033	2.027	2.032
Contrast	0.871	0.871	0.870	0.858	0.859	0.859	0.859	0.858

Fig. 14. The coefficients of source patterns on different sparse bases.

pattern on the DCT and Haar wavelet bases will help to avoid isolated bright source pixels or a very low pupil fill percentage that may lead to the aberrations due to the lens heating, or even lens-damage effect. Reference [23] also proved that the SO methods using spatial and DCT bases will result in similar optimized source patterns. Thus, the proposed SO method based on DCT, spatial and Haar wavelet bases will achieve manufacturable optimized source patterns.

5. Conclusion

This paper introduced a novel adaptive projection CS method and applied it to develop a fast and robust SO algorithm. The source optimization was formulated as an l_1-norm inverse problem based on a system of linear equations. The adaptive projection was used to reduce the dimensionality of the SO problem, while try to keep the information from all linear equations. The target layout was involved into the projection matrix through the random thresholding to effectively enhance the compression efficiency. To prove its advantage, mathematical expectation bounds for the adaptive projection method were proved. Simulations illustrated that the adaptive projection method outperformed the traditional CS methods. In addition, the impacts of four different sparse bases on the SO performance were also compared and discussed.

A. Appendix

The proof of the first inequality in Eq. (6) is as follows:

$$
\begin{aligned}
\bar{\mu}_\Upsilon &= \max_{\vec{\psi}_j \in \Upsilon} E\left\{ \left| \langle \vec{\phi}_i, \vec{\psi}_j \rangle \right|^2 \right\} > \frac{1}{\theta_{max}^2} \max_{\vec{\psi}_j \in \Upsilon} E\left\{ \left| \langle \vec{\phi}_i, \vec{\psi}_j \theta_j \rangle \right|^2 \right\} \\
&> \frac{1}{K^2 \theta_{max}^2} \max_{\vec{\psi}_j \in \Upsilon} E\left\{ \left| \langle \vec{\phi}_i, \sum_{j=1}^{K} \vec{\psi}_j \theta_j \rangle \right|^2 \right\} > \frac{1}{MK^2 \theta_{max}^2} E\left\{ \sum_{i=1}^{M} \left| \langle \vec{\phi}_i, \vec{X} \rangle \right|^2 \right\} \\
&= \frac{1}{MK^2 \theta_{max}^2} E\{ \|\mathbf{\Phi}\vec{X}\|_2^2 \},
\end{aligned}
\tag{11}
$$

where θ_j is the jth element in the coefficient vector $\vec{\Theta}$. In the above equation,

$$
\|\mathbf{\Phi}\vec{X}\|_2^2 = \frac{1}{N} \sum_{i=1}^{M} \left[\sum_{j=1}^{N} \mathrm{sgn}(\vec{S}_j - \Lambda_{i,j}) \vec{X}_j \right]^2.
\tag{12}
$$

For each i, denote that $\Delta_i = \left[\sum_{j=1}^{N} \mathrm{sgn}(\vec{S}_j - \Lambda_{i,j}) \vec{X}_j \right]^2$. Then we can calculate the mathematical expectation of Δ_i as:

$$
E(\Delta_i) = \|\vec{X}\|_2^2 + \sum_{r=1}^{N} \sum_{j=1, j \neq r}^{N} [T_1 - T_2],
\tag{13}
$$

where

$$
T_1 = \vec{X}_r \vec{X}_j \left[P_r(\Lambda'_{i,r} < \vec{X}_r) P_r(\Lambda'_{i,j} < \vec{X}_j) + P_r(\Lambda'_{i,r} > \vec{X}_r) P_r(\Lambda'_{i,j} > \vec{X}_j) \right],
$$

$$
T_2 = \vec{X}_r \vec{X}_j \left[P_r(\Lambda'_{i,r} > \vec{X}_r) P_r(\Lambda'_{i,j} < \vec{X}_j) + P_r(\Lambda'_{i,r} < \vec{X}_r) P_r(\Lambda'_{i,j} > \vec{X}_j) \right],
$$

where $\Lambda'_{i,j} = (\Lambda_{i,j} - \vec{N}_j) \sim \mathcal{N}(0, \sigma_\Lambda^2 + \sigma_X^2)$, and $P_r\{\bullet\}$ means the probability of the argument. Thus, Eq. (13) can be written as:

$$
\begin{aligned}
E(\Delta_i) &= \|\vec{X}\|_2^2 + \sum_{r=1}^{N} \sum_{j=1, j \neq r}^{N} \vec{X}_r \vec{X}_j \left[1 - 2Q\left(\vec{X}_r / \sqrt{\sigma_\Lambda^2 + \sigma_X^2} \right) - 2Q\left(\vec{X}_j / \sqrt{\sigma_\Lambda^2 + \sigma_X^2} \right) \right. \\
&\quad \left. + 4Q\left(\vec{X}_r / \sqrt{\sigma_\Lambda^2 + \sigma_X^2} \right) Q\left(\vec{X}_j / \sqrt{\sigma_\Lambda^2 + \sigma_X^2} \right) \right],
\end{aligned}
\tag{14}
$$

where the Q-function is defined as $Q(x) = \int_x^{+\infty} (1/\sqrt{2\pi}) \times \exp\left(-t^2/2 \right) dt$. Let $\tau_r = Q(\vec{X}_r / \sqrt{\sigma_\Lambda^2 + \sigma_X^2})$. If $\vec{X}_r \geq 0$, then $\tau_r \in (0, 0.5]$, otherwise $\tau_r \in (0.5, 1)$. Then, the term inside bracket in Eq. (14) can be abbreviated as $1 - 2\tau_r - 2\tau_j + 4\tau_r \tau_j = (1 - 2\tau_r)(1 - 2\tau_j)$. It is easy to prove that

$$
1 - 2\tau_r - 2\tau_j + 4\tau_r \tau_j \begin{cases} \leq 0 &: \quad \vec{X}_r \vec{X}_j < 0 \\ \geq 0 &: \quad \vec{X}_r \vec{X}_j > 0 \end{cases}.
\tag{15}
$$

Substituting Eq. (15) into Eq. (14), we find that $E(\Delta_i)$ is always larger than $\|\vec{X}\|_2^2$. According to Eq. (12), we have $E\{\|\mathbf{\Phi}\vec{X}\|_2^2\} > \frac{M}{N} \|\vec{X}\|_2^2$. Substituting this inequality into Eq. (11), we have

$$
\bar{\mu}_\Upsilon = \max_{\vec{\psi}_j \in \Upsilon} E\left\{ \left| \langle \vec{\phi}_i, \vec{\psi}_j \rangle \right|^2 \right\} > \frac{\|\vec{X}\|^2}{NK^2 \theta_{max}^2}.
\tag{16}
$$

The proof of the second approximate equality in Eq. (6) is as follows. Let $\hat{\vec{\phi}}$ and $\hat{\vec{\psi}}$ be the vectors that maximize the mathematical expectation, and $\hat{\Lambda}$ is the corresponding threshold. Thus,

$$
\begin{aligned}
\bar{\mu}_{\overline{\Upsilon}} &= \max_{\vec{\psi}_j \in \overline{\Upsilon}} E\left\{ \left| \left\langle \vec{\phi}_i, \vec{\psi}_j \right\rangle \right|^2 \right\} = E\left\{ \left| \left\langle \hat{\vec{\phi}}, \hat{\vec{\psi}} \right\rangle \right|^2 \right\} = \frac{1}{N} E\left\{ \left| \left\langle \mathrm{sgn}(\vec{X} - \hat{\Lambda}), \hat{\vec{\psi}} \right\rangle \right|^2 \right\} \\
&= \frac{1}{N} E\left\{ \sum_{p=1}^{N} \mathrm{sgn}(\vec{X}_p - \hat{\Lambda}_p)\hat{\vec{\psi}}_p \right\}^2 = \frac{1}{N} \sum_{m=1}^{N} \sum_{n=1}^{N} \hat{\vec{\psi}}_m \hat{\vec{\psi}}_n E\left\{ \mathrm{sgn}(\vec{X}_m - \hat{\Lambda}_m)\mathrm{sgn}(\vec{X}_n - \hat{\Lambda}_n) \right\} \\
&= \frac{1}{N} \sum_{m=1}^{N} \sum_{n=1}^{N} \hat{\vec{\psi}}_m \hat{\vec{\psi}}_n \left[1 - 2Q\left(\vec{X}_m / \sqrt{\sigma_\Lambda^2 + \sigma_X^2} \right) - 2Q\left(\vec{X}_n / \sqrt{\sigma_\Lambda^2 + \sigma_X^2} \right) \right. \\
&\quad \left. + 4Q\left(\vec{X}_m / \sqrt{\sigma_\Lambda^2 + \sigma_X^2} \right) Q\left(\vec{X}_n / \sqrt{\sigma_\Lambda^2 + \sigma_X^2} \right) \right].
\end{aligned}
\tag{17}
$$

When the argument of the Q-function is much smaller than 1, we have

$$
Q(x) \approx \frac{1}{2} - \frac{1}{\sqrt{2\pi}} x.
\tag{18}
$$

Based on Eq. (18) and the assumptions of $\left| \vec{X}_m / \sqrt{\sigma_\Lambda^2 + \sigma_X^2} \right| \ll 1$ and $\left| \vec{X}_n / \sqrt{\sigma_\Lambda^2 + \sigma_X^2} \right| \ll 1$, Eq. (17) can be transformed as

$$
\bar{\mu}_{\overline{\Upsilon}} = \max_{\vec{\psi}_j \in \overline{\Upsilon}} E\left\{ \left| \left\langle \vec{\phi}_i, \vec{\psi}_j \right\rangle \right|^2 \right\} \approx \frac{1}{N} \sum_{m=1}^{N} \sum_{n=1}^{N} \hat{\vec{\psi}}_m \hat{\vec{\psi}}_n \frac{2\vec{X}_m \vec{X}_n}{\pi(\sigma_\Lambda^2 + \sigma_X^2)} = \frac{2\langle \vec{X}, \hat{\vec{\psi}} \rangle^2}{N\pi(\sigma_\Lambda^2 + \sigma_X^2)} = 0,
\tag{19}
$$

where the last equality comes from $\vec{\psi}_j \in \overline{\Upsilon}$, which means $\vec{\psi}_j$ is orthogonal to $\vec{X}$.

Funding