USC–SIPI REPORT #314

Diffractive Optical Elements for Space-Variant Interconnections in Three-Dimensional Computation Structures

by

Ching-Chu Huang

August 1997

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Room 400
Los Angeles, CA 90089-2564 U.S.A.

Dedication

To my parents: Chin-Feng Yang and Chin-Cheng Huang

To my sisters: Shui-lien Huang, Vicky Huang, Li-Mei Huang, Chao-Huan Huang, and Chiu-Yuen Huang

To my parents-in-law: Chu-Shi Kuo and Chu-Shung Chen

and

To my wife: Chao-Pei Chen

Acknowledgements

I am deeply grateful to my research advisor, Dr. B. Keith Jenkins, for providing me a rich research environment. I have benefited greatly from his professional and personal guidance, as well as from his broad knowledge in many fields. Over my years of graduate study, all the discussion with him has been an inspiring and precious experience. He has taught me how to pay special attention to some critical detail in conducting research, making presentations, and writing papers. His thoroughness has helped me become more cautious in the details that I used to overlook. I have enjoyed years of friendship and research meeting with him. I am also grateful for his financial support.

Special thanks to Professor Alexander A. Sawchuk and Dr. Charles Kuznia for their kind help in both research and other matters over the years. I also thank my committee: Professor Charles Lanski, Professor Richard Leahy, and Professor Mark Pinkston. Their teaching and guidance have been of great value.

I like to thank a very special friend, Daniel J. Hsia, for his many years of warm friendship and help. Thanks also goes to Sabino Piazzolla, Nainjeet Ramlagan, Adam Goldstein, Greg Petrisor, Clare Waterson, Yunsong Huang, Bogdan Hoanca, Jeng-Feng Lin, Wei-Feng Hsu, Jen-Ming Wu, Chih-Hao Chen, Kuang-Yu Li, Ching-Han Hsu and Chih-Yung Cheng for their technical help and friendship during my graduate career.

I also like to thank Gloria Halfacre and Linda Varilla for their tremendous administration help. Also thanks to Dr. Allan Weber for his patience in solving computer problems.

Lastly, I thank my parents, Chin-Feng Yang and Chin-Cheng Huang, my parents-in-law, Chu-Shi Kuo and Chu-Shung Chen for their unconditional love and support in the pursuit of this dream. I thank my sisters, Shui-Lien Huang, Vicky Huang, Li-Mei Huang, Chao-Huan Huang, and Chiu-Yuen Huang for their love and support over the years. Most of all, I thank my lovely wife, Chao-Pei Chen for her enthusiastic love and endless support. She has made my life as a graduate student much more enjoyable and less stressful. She came into my life in the unique way. She has brought me the long missing part of my life. I thank God for bringing her to me.

Contents

List of Figures

List of Tables

Abstract

Optical architectures for fully connected and limited-fanout space-variant weighted interconnections based on diffractive optical elements for fixed-connection multilayer neural networks are investigated and compared in terms of propagation lengths, system volumes, connection densities, and interconnection crosstalk. The architectures are designed to be physically compact, and to be usable for optical interconnections between optoelectronic smart-pixel arrays for 3-D optically interconnected multichip modules. For a small overall system volume the limited-fanout architecture can accommodate a much larger number of input and output nodes. However, the interconnection crosstalk of the limited-fanout space-variant architecture is relatively high due to noise from the diffractive optical element reconstruction. Therefore, a crosstalk reduction technique based on a modified design for diffractive optical elements is proposed. It rearranges the reconstruction pattern of the diffractive optical elements such that less noise lands on each detector region. This technique is verified by simulating one layer of an interconnection system with 128×128 input nodes, 128×128 output nodes, and 5×5 nearest neighbor weighted connections from each input node to the output nodes.

We have also analyzed and simulated the effect of incorrect illumination wavelength and etch depth error on DOE reconstruction. The DOE-independent properties found in the simulations for binary-phase-level DOEs are theoretically verified. A way of cancelling the effect of etch depth error on DOE reconstruction by adjusting the illumination wavelength is described. The effect of incorrect illumination wavelength and

etch depth error on DOE reconstruction with more than two phase levels has founded to be DOE-dependent.

Two design algorithms, Gerchberg-Saxton and simulated annealing, are studied and modified for better DOE performance. A dual-cost-function simulated annealing algorithm that produces better DOE performance than the original simulated annealing algorithm has been proposed. Finally, a new design approach that increases design freedom without increasing fabrication complexity is developed. It allows the design algorithm to find the best non-uniform quantization of phase levels and can generally obtain DOEs with substantially better performance.

Chapter 1

Introduction

1.1 Motivation and Objectives

Neural networks have important applications in many fields such as image processing, vision, speech recognition, pattern recognition, and associative memory [1]. The key components of neural networks are neuron units and weighted interconnections (synapses) between neuron units. The number of weighted interconnections will increase as the square of the number of neuron units in a fully connected network [6]. Both electronic and optical implementations of artificial neural networks have been intensively studied [1, 2, 3, 4, 5]. Recently, there has been a great deal of interest in optoelectronic implementations of neural networks [7, 8, 9, 10, 11, 12, 13, 14]. Usually, an optoelectronic neural network uses optics in the third dimension to implement interconnections (either weighted or unweighted), while it employs a two-dimensional electronic chip to implement neural functionalities such as summing the inputs, nonlinear thresholding, or other more complex processing. There are different ways to implement weighted interconnections in optics [3, 4, 5]. In the case of fixed-connection neural networks, one possible way of implementing weighted interconnections is to use diffractive optical elements (DOEs). Diffractive optical elements can provide high diffraction efficiencies and they can be physically very small because they

are manufacturable using lithographic techniques in conjunction with single or multiple etch or deposition steps standard in semiconductor processing [25, 26]. The other components of the system (e. g., optoelectronic devices for neuron unit array) can also be made small and the collection of the components can be assembled into a mutually compatible, small, integrated structure. Therefore, optoelectronic implementations and employment of diffractive optical elements have the potential for building large scale neural networks with small size [10, 11, 15, 16, 17].

We are interested in optoelectronic implementations of large scale space-variant (S-V) fixed-connection neural networks with physically compact multilayer structures [11, 12, 13]. Such structures can process a two-dimensional array of interconnection channels in parallel with the potential for high bandwidth (100 MHz per channel) operation. Therefore, such a system can be applied to many fields that need to process large amounts of data in parallel. One example is the early vision processing of mammals. In these biological vision systems, the optical signal detected by the retinal photoreceptors is processed by a number of layers of neuron units in the retina for low level signal processing before being communicated to the visual cortex [36, 37, 38]. In adult mammals the interconnections among layers of the retina are generally believed to be nonadaptive, mostly local, and highly parallel [39]. Thus, neural-inspired vision preprocessing operations may make an interesting application for a parallel, multilayer optoelectronic system. If the system is also physically compact, it may have uses in a variety of intelligent-sensing applications.

In this effort, we focus on the interconnection portion of these networks by investigating possible free-space optical architectures. The desired interconnection system should have the following characteristics: it should be capable of analog weighted fan-out, cascadable to form multilayer feedforward structures, and capable of short propagation length between layers. Moreover, the system should be designed so that it can be fabricated using semiconductor processing techniques in order to have the potential for large-volume and low-cost manufacturing. Based on the use of diffractive optical elements, the purpose of the first part of this thesis is then to study what interconnection architectures are suitable for multilayer feedforward structures for neural network implementations in terms of system size, space-bandwidth product requirement, and interconnection crosstalk. Since diffractive optical elements are the key components of the interconnection systems, the second portion of the thesis focuses on the analysis of the effect of incorrect illumination wavelength and etch depth error on DOE reconstruction. Lastly, because the performance of diffractive optical elements is critically dependent on the design algorithm used, we have investigated two design algorithms, Gerchberg-Saxton and simulated annealing, and, based on these two original algorithms, developed new design approaches that are substantially better than the original algorithms.

1.2 Organization of the Thesis

Chapter 2 describes previous related works by other researchers. Because diffractive optical elements are the key components in the systems of interest, we investigate the properties of DOEs and establish a reconstruction model of DOEs in Chapter 3. In this

thesis, we investigate architectures that have different connection patterns for different neuron units. These types of architectures are referred as space-variant (S-V) interconnection systems. Of the space-variant systems, we have studied two different architectures: the fully connected system in which each neuron unit connects to *all* other neuron units; the limited-fanout system in which each neuron unit only connects to its nearest neighbor. Chapter 4 studies the characteristics of the fully connected space-variant interconnection architecture such as propagation length, lateral dimension, system aspect ratio, interconnection density, space-bandwidth product requirement, and interconnection crosstalk. Chapter 5 analyzes the characteristics of the limited-fanout space-variant interconnection architecture and compares them with the fully connected space-variant interconnection architecture. Although the limited-fanout space-variant architecture can provide a compact system size for large scale neural networks, the interconnection crosstalk is unacceptably high due to noise from the reconstructions of the surrounding diffractive optical elements. Therefore, we propose a crosstalk reduction technique by modified DOE design for a limited-fanout S-V system in Chapter 5, verify its validity by additional simulations, and discuss the other potential advantages and limitations of this crosstalk reduction technique.

In order to have good performance from a designed diffractive optical element, the DOE needs to be etched correctly in fabrication and illuminated by laser source with the correct wavelength. In the second portion of the thesis, we will investigate the effect of an incorrect illumination wavelength and/or incorrect etch depth on the reconstruction of the DOE. In Chapter 6, we will study the effect on DOE reconstruction when the illumination

wavelength is different from the designed optical wavelength. In Chapter 7, we will investigate the effect of etch depth error in fabrication on the reconstruction of a diffractive optical element. In the last part of the thesis (Chapter 8), we will develop new design approaches that will increase the performance of DOEs and compare the designed DOE results of the new approaches with DOEs designed using two commonly used algorithms (Gerchberg-Saxton and simulated annealing). Finally, Chapter 9 summarizes the key results of this thesis and describes future research directions for related work.

1.3 Contributions

The original contributions of this dissertation are summarized as follows:

(1) Properties and Reconstruction Model of Diffractive Optical Elements: Characterized the properties of diffractive optical elements and categorized the reconstructed diffraction orders. Derived the relations among the diffraction orders and developed a reconstruction model that takes the non-ideal reconstruction lens into consideration.

(2) Fully Connected Space-Variant Interconnection Architecture: Analyzed the fully connected space-variant interconnection architectures for DOE-based multi-layer feed-forward 3-D computational structures and their important characteristics such as propagation length, lateral dimension, system aspect ratio, interconnection density, and space-bandwidth product requirement. The interconnection crosstalk for the fully-connected interconnection system was also analyzed and simulated.

(3) Limited-Fanout Space-Variant Interconnection Architecture: Analyzed the limited-fanout space-variant interconnection architectures for DOE-based multi-layer feed-forward 3-D computational structures and their important characteristics such as propagation length, lateral dimension, system aspect ratio, interconnection density, and space-bandwidth product requirement. The characteristics of the limited-fanout space-variant interconnection architecture were then compared with the characteristics of the fully connected space-variant interconnection architecture. The interconnection crosstalk for the limited-fanout interconnection system was analyzed and simulated. The crosstalk was likely too high for most of the applications. A crosstalk reduction technique based on a modified design for diffractive optical elements that rearranges the reconstruction pattern such that less noise lands on each detector region is then developed. Simulations showed that the technique can effectively reduce the crosstalk in the system.

(4) Effect of Incorrect Illumination Wavelength on DOE Reconstruction: Analyzed and simulated the effect of incorrect illumination wavelength on the reconstruction of multiple-phase-level DOEs. Theoretically verified the special properties found in the simulations for binary-phase-level DOEs. Similar expression also derived for unconventional binary-phase-level DOEs.

(5) Effect of Etch Depth Error in Fabrication on DOE Reconstruction: Analyzed and simulated the effect of two different kinds of etch depth error in fabrication on the reconstruction of multiple-phase-level DOEs. For both conventional and

unconventional binary-phase-level DOEs, the combined effects of incorrect illumination wavelength and etch depth error on reconstruction were theoretically derived. A way of cancelling the effect by adjusting the illumination wavelength was described.

(6) New Design Approaches for Diffractive Optical Element: Modified two commonly used design algorithms (Gerchberg-Saxton and simulated annealing algorithms) for better DOE performance. Developed a new design approach that relieves unnecessary constraints and increases the number of degrees of freedom in the DOE design. The new approach is based on non-uniform phase quantization in the DOE design and keep the number of etch steps unchanged. This new design approach can substantially increase the performance of diffractive optical elements without increasing the fabrication complexity of diffractive optical elements.

Chapter 2

Previous Related Work

2.1 Space-Variant Interconnection System Using a Sub-Hologram Array

In this thesis, we consider optical architectures that can connect a two-dimensional array of light transmitters (modulators or emitters) to a two-dimensional array of detectors. In a multilayer feedforward neural network, the light transmitters represent the outputs from neuron units in the previous layer and the detectors represent the neural receptors of next layer. The transmitters can be treated as the input nodes of the interconnection system and the detectors can be treated as the output nodes of the interconnection system. The interconnection patterns are assumed to be space-variant i.e., the fan-out pattern for every input node is different. The space-variant interconnections can be implemented by a two-dimensional array of sub-holograms, one for each input node. Here, each sub-hologram can be an optically recorded hologram, a computer generated hologram, or a diffractive optical element. For a fully connected system, each input node connects to all of the output node and the interconnections are effectively four-dimensional. Each sub-hologram must encode the connections from one input node to all the output nodes. This means that the space-bandwidth product of each sub-hologram must be at least N^2 in a system with $N \times N$

neuron units. In each layer, yielding a space-bandwidth product of at least N^4 for the entire sub-hologram array.

This type of optical architecture that employs a sub-hologram array for space-variant interconnections was first described by Jenkins *et al.* for digital optical computing applications [61]. They also analyzed the interconnection crosstalk in such an architecture and the effect on the maximum number of fan-in for each logic gate. Caulfield later applied this sub-hologram concept to implement weighted interconnections of neural networks [6].

Keller and Gmitro developed a similar space-variant architecture for weighted interconnections of neural networks but using diffractive optical elements (DOEs) instead of computer generated holograms [42, 62]. They also estimated that current DOE fabrication technology can support weighted interconnections for 5000 neuron units to 5000 neuron units. Conventional computer generated holograms use amplitude modulation and detour phase (or other kinds of techniques) to represent the complex wave fronts. Because of the amplitude modulation, conventional computer generated holograms have low diffraction efficiencies (~ 10%) [42]. On the other hand, diffractive optical elements use only phases to encode the complex waveform and usually have much higher diffraction efficiencies (60% ~ 90%) [28, 29]. However, because of the phase-only representation of complex waveforms data, a diffractive optical element will reconstruct undesired noise as well as desired signals. When many diffractive optical elements diffract light at the same time, the interconnection noise in the system could be serious and need to be carefully analyzed.

While the approaches above use optics to implement fan-out and fan-in, Yayla *et al.* recently proposed a different approach for space-variant interconnections [9]. They used optics for fan-out from neuron units but used electronics for fan-in to each neuron unit.

2.2 Compact Multilayer Structures Using Micro-Optic Components

Veldkamp described a multilayer electro-optic architecture that blend layers of analog circuitry, microoptics (diffractive optical elements), microlaser arrays and detector arrays for neural retinal designs [10]. Each neuron unit only connects to some nearest neighbor neuron units in the next layer. These localized interconnections are achieved by using two-dimensional arrays of diffractive optical elements, one for each neuron unit. All the micro-components in the system can be fabricated by a sequence of compatible production processes. This common fabrication technology not only eliminates the need for external components and alignment in the system, but also reduces size and weight, improves stability, and creates possibility of large-volume low-cost mass production.

Tanguay *et al.* suggested another dense 3-dimensional integrated electronic/photonic computing structures enabled by diffractive optical elements [11, 12, 18, 19]. Figure 2-1 shows the multilayer MCM structure. In this multilayer structure, each layer consists of an optoelectronic module, which is integrated by filp-chip bonding a silicon-based signal processing electronic chip with a multiple quantum well light modulator array chip, a waveguide array as an optical power bus to provide highly compact readout of reflection

mode optical modulator array in the third dimension, and an array of diffractive optical elements to provide localized weighted interconnections to the next layer (or a volume holographic optical element to provide the global interconnection to the next layer). The incoming information at each neuron unit is processed by the electronic circuitry next to the neuron unit, fanned out to the next layer through the light modulator and diffractive optical element. The three-dimensional integration of silicon modules with parallel optical interconnections will have the potential of high capacity computational ability for analog and digital applications in a compact, lower power module.

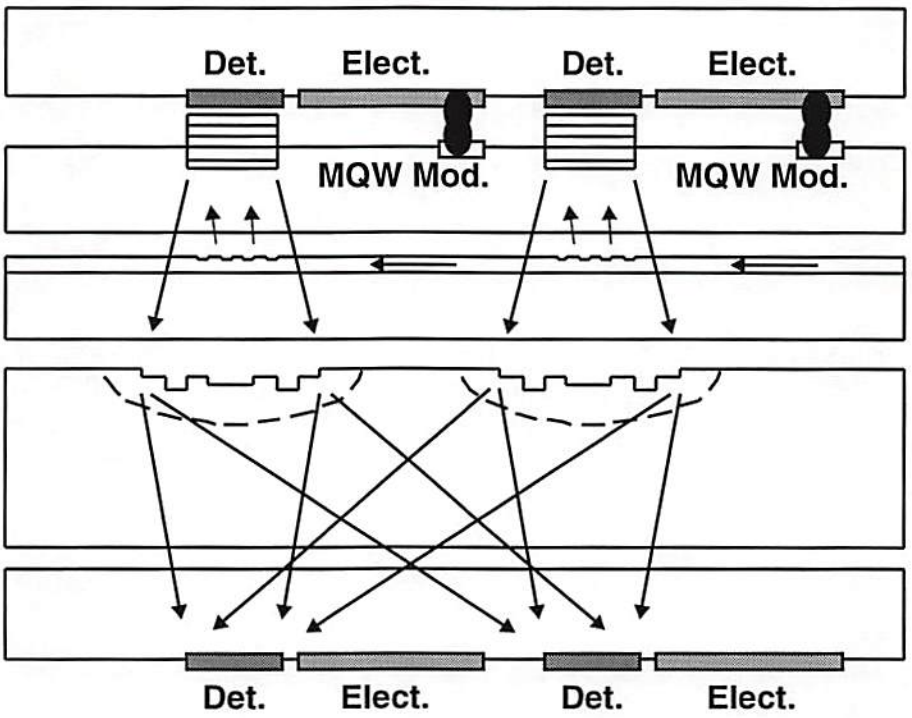

Figure 2-1: Integrated 3-D electronic/photonic multi-chip module (MCM) structure.

2.3 Diffractive Optical Elements

Diffractive Optical Elements have applications in many fields [27, 28, 29] such as optical interconnection systems [20, 21], optical head for optical disks [22], photonic circuit packaging [23], fiber communication [24], laser beam shaping [31], spot array generation [32, 33, 34]and many others [30, 35]. In this thesis, we assume DOEs are two-dimensional, consisting of $C \times C$ replications of two-dimensional DOE periods of size $T \times T$. Each DOE period is made up of $J \times J$ phase elements, each of size $\Delta \times \Delta$. The phase element is the smallest physical feature in the DOE. Each phase element takes on a constant discrete phase value over its $\Delta \times \Delta$ aperture that corresponds to the phase delay it imparts on the light striking it. Figure 2-2 shows the layout and phase distribution of a 4×4 period DOE containing 16×16 phase elements in each period, in which each phase element can be one of four values: 0, $\pi/2$, π, or $3\pi/2$.

Usually a lens is placed immediately after a DOE to obtain the reconstructed intensity pattern at the focal plane of the lens. The reconstructed intensity and the DOE amplitude transmittance are therefore related by a squared-magnitude of Fourier transform. Figure 2-3 shows the center portion of the reconstructed intensity pattern of the DOE in Figure 2-2 with $C = 4$.

In order to have good performance from a well designed diffractive optical element, the DOE needs to be fabricated appropriately. In the fabrication process for DOEs, an incorrect etch depth for each phase element, an incorrect phase element size, and a non-vertical sidewall are the commonly seen fabrication error. Jahn *et al.* [31] described the

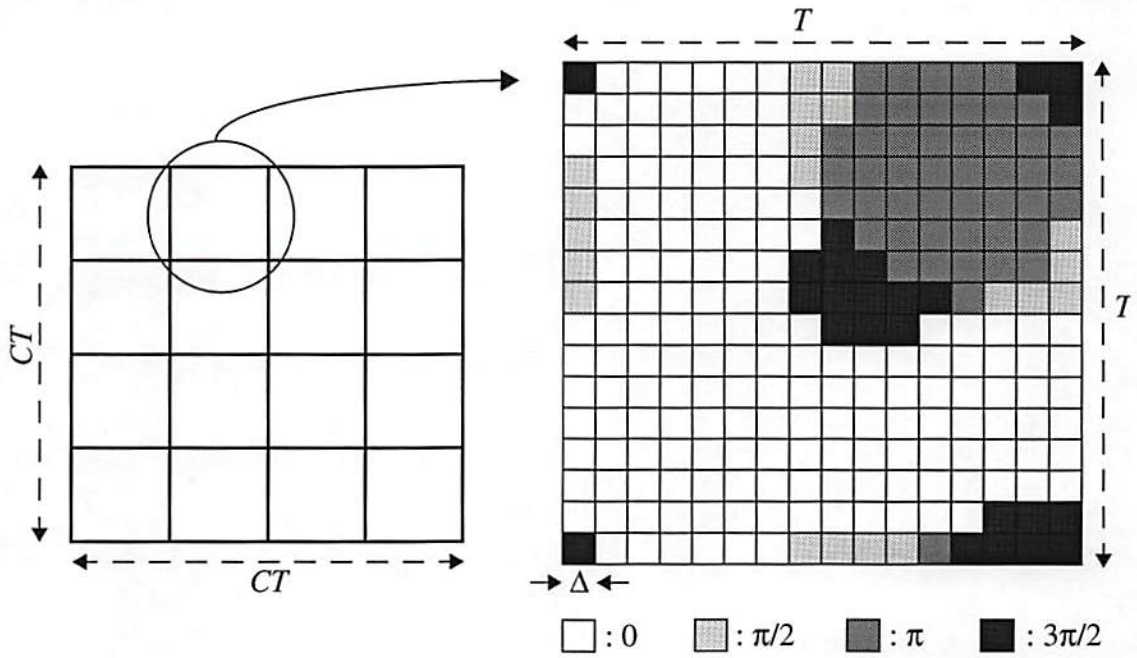

Figure 2-2: The layout and phase distribution of a four-phase-level DOE. T represents the period length and C represents the compression ratio. In this design, $T = 16$ phase elements.

fabrication process for DOEs and characterized the error sources for an incorrect etch depth and an incorrect phase element size from the fabrication processing for binary phase level DOEs.

In order to obtain the desired reconstructed intensity pattern from a diffractive optical element, the phase distribution in a DOE needs to be designed carefully. In general, because a DOE is a phase-only element, a phase-only distribution that gives the exact desired intensity pattern does not exist. Therefore, the goal of the design process is to optimize the phase distribution of the DOE to obtain a reasonable approximation to the desired reconstruction. Many different DOE design algorithms have been proposed and

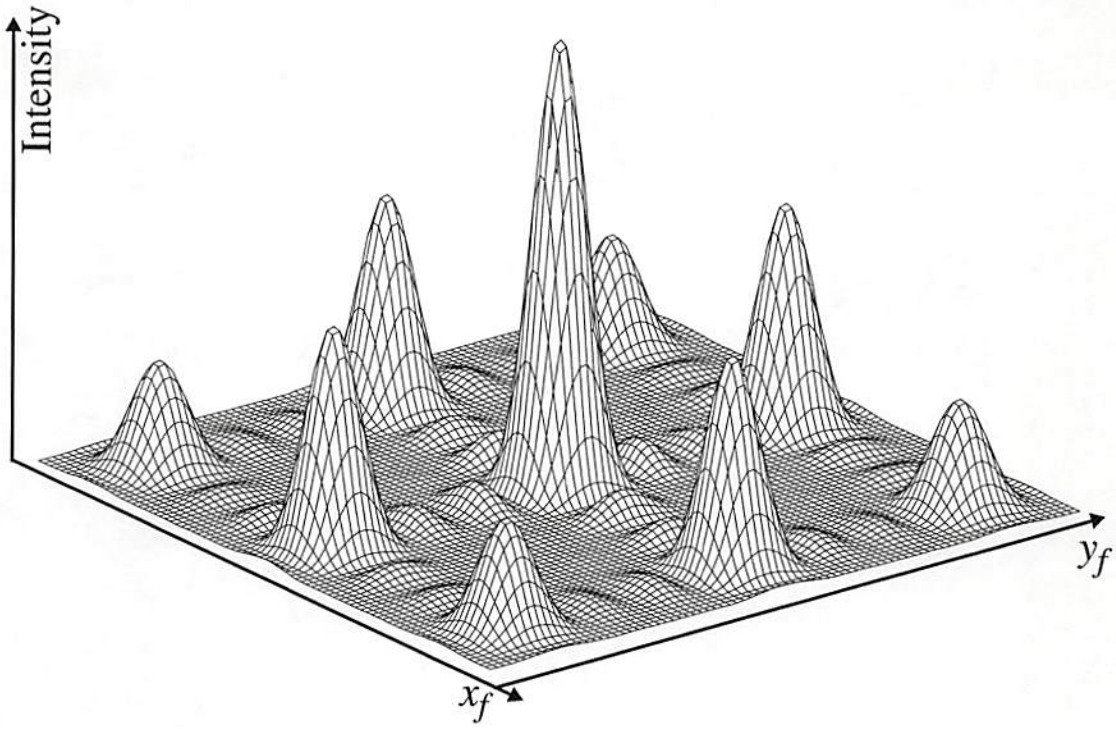

Figure 2-3: The center portion of the reconstructed intensity pattern of the DOE in Figure 2-2 with a compression ratio (C) of 4.

investigated [40, 42, 43, 44, 56, 57, 64, 65, 66, 67, 69, 70]; the two most commonly used algorithms are Gerchberg-Saxton [42, 45, 46, 47, 48, 49, 58, 59, 60, 63] and simulated annealing [50, 51, 52, 53, 54, 55, 52, 52, 68]. The Gerchberg-Saxton algorithm is an iterative Fourier transform procedure that attempts to iteratively satisfy the required constraints on both the DOE plane and reconstruction plane. These constraints include a discrete phase-only distribution (amplitude is equal to one) in the DOE plane, and the desired intensity (magnitude) connection pattern in the reconstruction plane. The phases in the reconstruction plane are generally arbitrary and can be used as available degrees of

freedom in the design process. Usually, the Gerchberg-Saxton algorithm converges to a reasonable solution in a short amount of time.

Simulated annealing, on the other hand, is a very time-consuming process that can theoretically obtain an optimum solution by using statistical searching procedures with an appropriate annealing schedule and unlimited computation time. The algorithm begins with a randomly selected phase distribution and a initial annealing temperature. In each iteration, the algorithm randomly select a phase element from one period of the DOE and randomly change its phase to one of the pre-defined discrete levels. An appropriate cost, which is usually some figure of merit of DOE performance, is then calculated for this trial DOE phase distribution. If the cost of the trial is less than the original cost, the change is unconditionally accepted. Otherwise, the change is accepted only if the probability of accept, which is usually some function of the annealing temperature and the difference between the trial cost and the original cost, is greater than some randomly selected minimum probability. This mechanism of accepting a bad change is very important because it is used to move the algorithm out of a local minimum. After some time, the annealing temperature is decreased to reduce the probability of accepting a bad change. This process is then repeated till it converges to the final solution. Because, in each iteration, the simulated annealing algorithm operates on only one of the phase elements in one period of the DOE rather than on all the phase elements in one period of the DOE simultaneously as in the case of the Gerchberg-Saxton algorithm. Therefore, the simulated annealing algorithm will, in general, take much more iterations to converge than the Gerchberg-Saxton

algorithm. Further, the total number of iterations will also increase as the number of phase levels increases.

Conventional diffractive optical elements usually use discrete phase levels that are equally distributed between 0 and π. diffractive optical elements of this kind are easier to design and fabricate. Feldman *et al.* [50] suggested a non-conventional phase level assignment for binary-phase-level DOEs. In stead of using 0 and π for the two phase level assignment, they have used simulated annealing algorithm to search for the optimum phase level assignment and the optimum phase distribution at the same time for the DOE. Using their proposed scheme, they have designed the DOEs for spot array generators and have found the optimum phase level assignment other than 0 and π.

Chapter 3

Properties and Reconstruction Model of Diffractive Optical Elements

In this chapter, we will derive the relations between a diffractive optical element and its reconstructed diffraction orders. These relations are necessary for the system characteristics study in Chapters 4 and 5. We will also derive the relations among reconstructed diffraction orders. Based on the relations among diffraction orders, we then establish the reconstruction model necessary for analyzing the interconnection crosstalk in Chapters 4 and 5.

3.1 Characteristics and Properties of Diffractive Optical Elements

In this section, we will derive the relations between a diffractive optical element and its reconstructed diffraction orders as well as the relations among reconstructed diffraction orders. These relations are necessary for analyzing the crosstalk in Chapter 4 and 5. In this thesis, we assume DOEs are two-dimensional, consisting of $C \times C$ replications of a two-dimensional DOE period of physical size $T \times T$. Each DOE period is made up of $J \times J$ phase elements, each of physical size $\Delta \times \Delta$. The phase element is the smallest physical feature in the DOE. Each phase element takes on a constant discrete phase value over its

$\Delta \times \Delta$ aperture that corresponds to the phase delay it imparts on the light striking it. Figure 1(a) shows the layout and phase distribution of a 4×4 period DOE containing 16×16 phase elements in each period, in which each phase element can be one of four values: 0, $\frac{\pi}{2}$, π, or $\frac{3\pi}{2}$.

Because of its periodicity, the DOE amplitude transmittance function, $\underline{g}(x, y)$, can be expressed as a linear combination of two-dimensional complex exponential functions [40] as

$$\underline{g}(x, y) = \text{rect}\left(\frac{x}{CT}\right) \bullet \text{rect}\left(\frac{y}{CT}\right) \bullet \underline{g}_I(x, y), \tag{3-1}$$

where

$$\underline{g}_I(x, y) = \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} \underline{A}(r, s) \exp\left(j2\pi \frac{(rx + sy)}{T}\right), \tag{3-2}$$

with

$$\underline{A}(r, s) = \frac{1}{T^2} \int\int_{\Sigma} \underline{g}_I(x, y) \exp\left(-j2\pi \frac{(rx + sy)}{T}\right) dxdy. \tag{3-3}$$

where Σ is any one complete period of $\underline{g}_I(x, y)$. The term $\text{rect}[x/(CT)] \cdot \text{rect}[y/(CT)]$ in Eq. (3-3) imposes the finite aperture size of the DOE with the rect function defined as

$$\text{rect}(x) = \begin{cases} 1 & |x| \le \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}. \tag{3-4}$$

If $\underline{h}(p, q)$ represents the complex amplitude transmittance of the (p, q)'th phase element in one period of $\underline{g}(x, y)$, then a single DOE period can be expressed as

$$\underline{g}_T(x, y) = \sum_{p=0}^{J-1} \sum_{q=0}^{J-1} \underline{h}(p, q) \cdot \text{rect}\left(\frac{x - p\Delta}{\Delta}\right) \cdot \text{rect}\left(\frac{y - q\Delta}{\Delta}\right), \tag{3-5}$$

where J is the number of phase elements along each dimension of a single DOE period. Equation (3-5) describes the continuous period of $\underline{g}(x, y)$ as a discrete set of $J \times J$ phase elements, each of physical size $\Delta \times \Delta$.

Because the integral portion of Eq. (3-3) integrates over one period of $\underline{g}_I(x, y)$, we can substitute the expression for one DOE period $\underline{g}_T(x, y)$ of Eq. (3-5) in for $\underline{g}_I(x, y)$ in Eq. (3-3), and take the integration region Σ to be coincident with the DOE period defined in Eq. (3-5) to obtain

$$\underline{A}(r,s) = J^{-2}\operatorname{sinc}\left(\frac{r}{J}\right)\cdot\operatorname{sinc}\left(\frac{s}{J}\right)\sum_{p=0}^{J-1}\sum_{q=0}^{J-1}\underline{h}(p,q)\exp\left(\frac{-j2\pi}{J}(rp+sq)\right), \qquad (3\text{-}6)$$

where the sinc function is defined by

$$\operatorname{sinc}(x) = \frac{\sin(\pi x)}{\pi x}. \qquad (3\text{-}7)$$

Note that the double summation portion of Eq. (3-6) can be calculated by the two-dimensional discrete Fourier transform of $\underline{h}(p,q)$. The function $\underline{A}(r,s)$ is the product of a periodic function (with a period of J in each dimension) and an envelope given by $\operatorname{sinc}\left(\frac{r}{J}\right)\cdot\operatorname{sinc}\left(\frac{s}{J}\right)$.

If we assume a transmissive DOE illuminated by a normally-incident unit-amplitude plane wave, then under the conditions that allow the DOE diffraction to be described by scalar diffraction theory [40, 41], the complex amplitude distribution immediately behind the DOE is

$$\underline{g}(x,y) = \operatorname{rect}\left(\frac{x}{CT}\right)\cdot\operatorname{rect}\left(\frac{y}{CT}\right)\sum_{r=-\infty}^{\infty}\sum_{s=-\infty}^{\infty}\underline{A}(r,s)\exp\left(j\frac{2\pi}{T}(rx+sy)\right). \qquad (3\text{-}8)$$

In Eq. (3-8), $\underline{g}(x, y)$ can be treated as a linear combination of many apertured plane waves with different directions [42]. We refer to each of the plane waves as one diffraction order. If a lens is placed immediately behind the DOE, the complex amplitude distribution, $\underline{G}(x_f, y_f)$, at the focal plane behind the lens is the Fourier transform of $\underline{g}(x, y)$, given by

$$\begin{aligned}&\underline{G}(x_f, y_f)\\ &= \underline{Q}(x_f, y_f) \bullet \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} \underline{A}(r, s)\, \mathrm{sinc}\left[\frac{CT}{\lambda f}\left(x_f - \frac{r\lambda f}{T}\right)\right] \mathrm{sinc}\left[\frac{CT}{\lambda f}\left(y_f - \frac{s\lambda f}{T}\right)\right],\end{aligned} \tag{3-9}$$

with

$$\underline{Q}(x_f, y_f) \equiv \frac{C^2 T^2 \exp\left[\frac{j\pi}{\lambda f}(x_f^2 + y_f^2)\right]}{j\lambda f}, \tag{3-10}$$

where f is the focal length of the lens and λ is the optical wavelength. The corresponding intensity distribution, $I(x_f, y_f)$, is given by

$$I(x_f, y_f) = \left|\underline{G}(x_f, y_f)\right|^2. \tag{3-11}$$

Equation (3-9) describes a diffraction pattern with infinitely many spots. The (r, s)'th spot is the Fourier transform of the corresponding diffraction order and has an amplitude profile

$$\tilde{\underline{A}}(x_f, y_f) = \underline{\underline{Q}}(x_f, y_f) \bullet \underline{A}(r, s)\,\mathrm{sinc}\left[\frac{CT}{\lambda f}\left(x_f - \frac{r\lambda f}{T}\right)\right]\mathrm{sinc}\left[\frac{CT}{\lambda f}\left(y_f - \frac{s\lambda f}{T}\right)\right]. \quad (3\text{-}12)$$

Without loss of generality, the phase and constant factor, $\frac{C^2T^2\exp\left[\frac{j\pi}{\lambda f}(x_f^2 + y_f^2)\right]}{j\lambda f}$, will be dropped in the following sections for simplicity. From Eq. (3-12), it can be seen that each diffraction order spot is centered at $\left(\frac{r\lambda f}{T}, \frac{s\lambda f}{T}\right)$ with a spot width (the width between the first zeros of the sinc function) of

$$S_w = \frac{2\lambda f}{CT}, \quad (3\text{-}13)$$

and a center-to-center spot spacing of

$$S_S = \frac{\lambda f}{T}. \quad (3\text{-}14)$$

The diffraction angle of each diffraction order relative to the optical axis of the lens is

$$(\theta_r, \theta_s) = \left(\tan^{-1}\left(\frac{r\lambda}{T}\right), \tan^{-1}\left(\frac{s\lambda}{T}\right) \right). \quad (3\text{-}15)$$

There are two important properties that can be observed from above equations. If we define the resolution at the reconstruction plane as the ratio of spot spacing, S_s, to the spot width, S_w, then the resolution is equal to $\frac{C}{2}$ and is independent of wavelength (λ), focal length (f), and period length (T). Hence, the number of periods in a DOE in each dimension, C, is also referred to as the compression ratio [40]. The compression ratio (C) is usually greater than two in order to ensure a minimum separation for the reconstructed spots. It is also convenient to consider the compression ratio in terms of the resulting diffraction pattern. Thus $C = \frac{2S_S}{S_w}$, or twice the center-to-center spot spacing normalized by the spot width.

3.2 Reconstruction Model of Diffractive Optical Elements

Equation (3-6) shows that the complex amplitude of each diffraction order, $\underline{A}(r, s)$, is a sinc enveloped periodic function with a period of J in each dimension. The number of independent diffraction orders in the reconstruction is equal to the number of phase elements in one period of a DOE. Therefore, to create a DOE whose reconstruction has $J_0 \times J_0$ independently weighted fan-outs, each period of a DOE needs to have at least $J_0 \times J_0$ phase elements. Additional phase elements are usually used to improve the DOE performance (such as diffraction efficiency, uniformity between diffraction orders, and signal-to-noise ratio). We define the ratio of J over J_0 as the oversampling ratio, B. Thus, $J = BJ_0$ and $T = BJ_0\Delta$. The oversampling ratio is usually greater than one for performance reasons. Accordingly, there are $BJ_0 \times BJ_0$ independent diffraction orders in the reconstructed pattern. We refer to the $BJ_0 \times BJ_0$ diffraction orders centered at the optical axis of the lens as the *mainlobe* of the reconstruction and each adjacent set of $BJ_0 \times BJ_0$ diffraction orders as one *sidelobe*. The reconstruction consists of one mainlobe and many surrounding sidelobes. The amplitude of each diffraction order in any sidelobe is dependent (through the sinc envelope) on the amplitude of the corresponding order in the mainlobe. We refer to the $J_0 \times J_0$ desired diffraction orders within the mainlobe as the *signal diffraction orders* or simply *signals*. Each sidelobe contains an array of $J_0 \times J_0$ diffraction orders centered within the sidelobe that corresponds to the signals weighted by the envelope sinc function. We refer to these diffraction orders as *signal sidelobes* (SS's). The remaining of the diffraction orders in both the mainlobe and sidelobes are *spurious*

diffraction orders (SDO's). Hence, a reconstructed pattern will consist of signals, signal sidelobes, and spurious diffraction orders, as shown in Figure 3-1. Signal sidelobes and

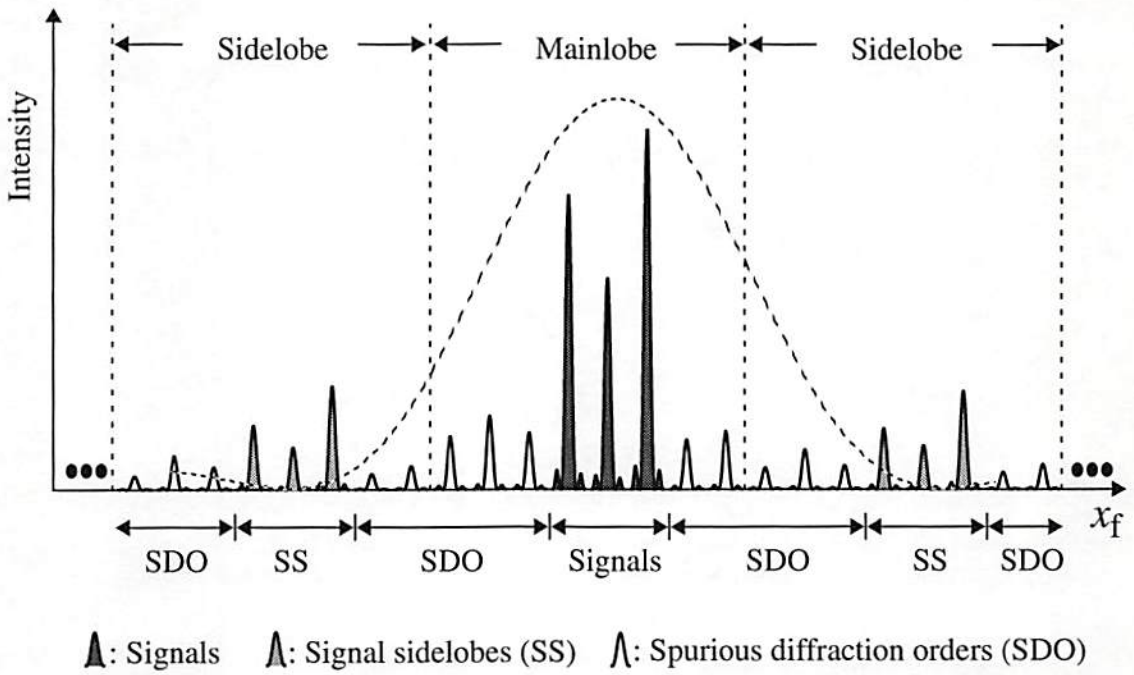

Figure 3-1: The reconstruction pattern (shown in one dimension) of a typical DOE consists of signals, signal sidelobes (SS), and spurious diffraction orders (SDO). The overall envelope (due to the finite size of the phase elements) is shown by the dashed curve. The example shown has 8×8 phase elements in one period of a DOE. The mainlobe and two complete sidelobes of the reconstruction are plotted.

spurious diffraction orders are undesired noise that causes interconnection crosstalk, as will be shown later, in a limited-fanout space-variant interconnection system. We will use (u, v) to represent the indices of $J_0 \times J_0$ signals $(-\frac{(J_0-1)}{2} \le u \le \frac{(J_0-1)}{2}$ and $-\frac{(J_0-1)}{2} \le v \le \frac{(J_0-1)}{2}$, assuming J_0 is odd) and (k_1, k_2) to represent the indices of

sidelobes ($k_1, k_2 \in \{0, \pm 1, \pm 2, ...\}$ and $(k_1, k_2) \neq (0, 0)$). For the (u, v)'th signal, the corresponding signal sidelobe in the (k_1, k_2)'th sidelobe is in the $(k_1 BJ_0 + u, k_2 BJ_0 + v)$'th diffraction orders of the DOE. By replacing J with BJ_0, we can calculate the intensity ratio of the signal sidelobe over the corresponding connection order from Eq. (3-6) as

$$\frac{I(k_1 BJ_0 + u, k_2 BJ_0 + v)}{I(u, v)} = \left| \frac{\left(\frac{u}{BJ_0}\right)}{k_1 + \left(\frac{u}{BJ_0}\right)} \right|^2 \times \left| \frac{\left(\frac{v}{BJ_0}\right)}{k_2 + \left(\frac{v}{BJ_0}\right)} \right|^2, \tag{3-16}$$

Note that the relative strength of a given signal sidelobe diffraction order (u, v, k_1 and k_2 fixed) is only dependent on the oversampling ratio (B) for a given J_0. This relative strength becomes lower if the oversampling ratio gets larger.

While the reconstruction above assumed a perfect Fourier transform lens, in reality the actual reconstructed pattern will depend on the type (e.g., refractive or diffractive) and the aberration properties of the lens. In general, the reconstructed intensity pattern from a DOE will be a relatively accurate rendition of its power spectrum in a local region near the lens optical axis, but will degrade due to aberrations and non-paraxial effects farther away from the axis. To approximate these effects, we model the paraxial-region reconstruction as the exact power spectrum of a DOE, and the reconstruction everywhere outside this paraxial region as an uniform-intensity blur [denoted as non-paraxial region noise (NRN)]

as shown in Figure 3-2. Because of the sinc squared intensity envelope, if the paraxial region includes at least the region out to the second zero of the sinc function (in each dimension), most (>90%) of the light energy is in the paraxial reconstruction region near the optical axis of the lens, and relatively little light energy is in the NRN region.

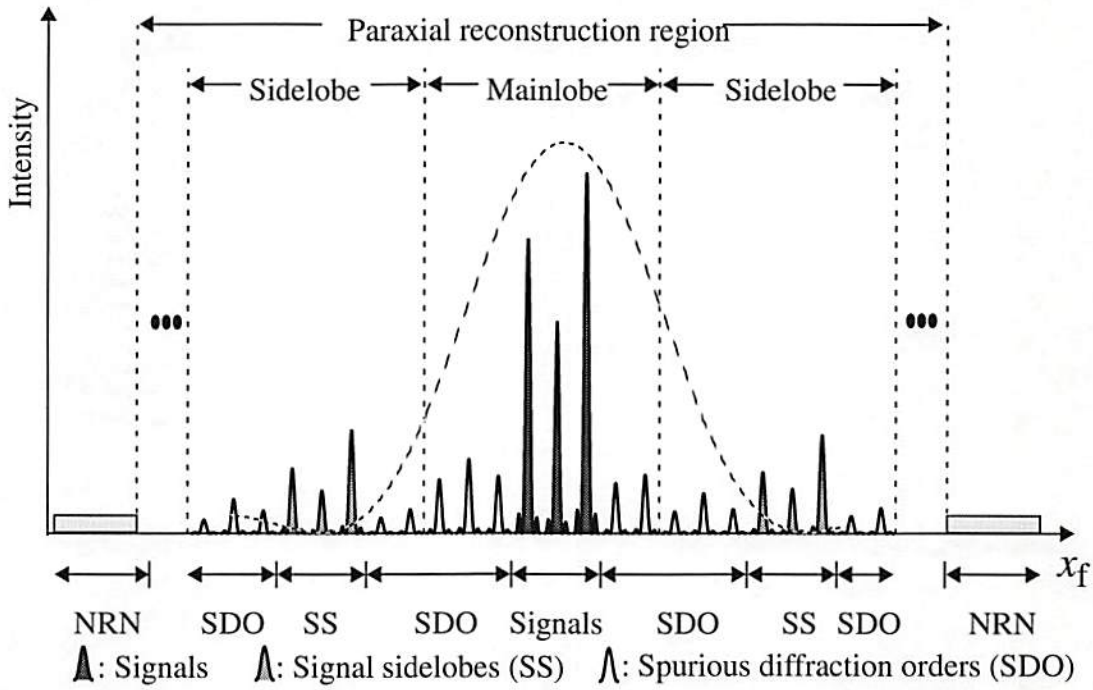

Figure 3-2: Taking non-idealities of a lens into consideration, the reconstruction pattern of a DOE consists of signals, signal sidelobes (SS), spurious diffraction orders (SDO), and non-paraxial region noise (NRN). The overall envelope (due to the finite size of the phase elements) is shown by the dashed curve. The example shown has 8×8 phase elements in one period of a DOE.

Chapter 4

DOE-Based Multi-Layer Feed-Forward Fully Connected Space-Variant Interconnection Architecture and Crosstalk Analysis

In this chapter, we investigate the interconnection architectures for optoelectronic implementations of fully connected neural networks. Section 4.1 describes the generic DOE-based interconnection architecture for optoelectronic implementation of multi-layer feed-forward neural networks. Section 4.2 focuses on the fully connected space-variant interconnection architectures and their important characteristics such as propagation length, lateral dimension, system aspect ratio, interconnection density, and space-bandwidth product requirement. Section 4.3 describes the origin and definition of the crosstalk in such DOE-based space-variant interconnection systems. Section 4.4 then presents the analysis and simulation of the interconnection crosstalk for the fully-connected interconnection systems.

4.1 Optoelectronic Implementation of Multi-Layer Feed-Forward Space-Variant Interconnection Architecture for Neural Networks

Shown in Figure 4-1 is the DOE-based multilayer feed-forward space-variant interconnection architecture considered in this thesis. Each layer of the architecture

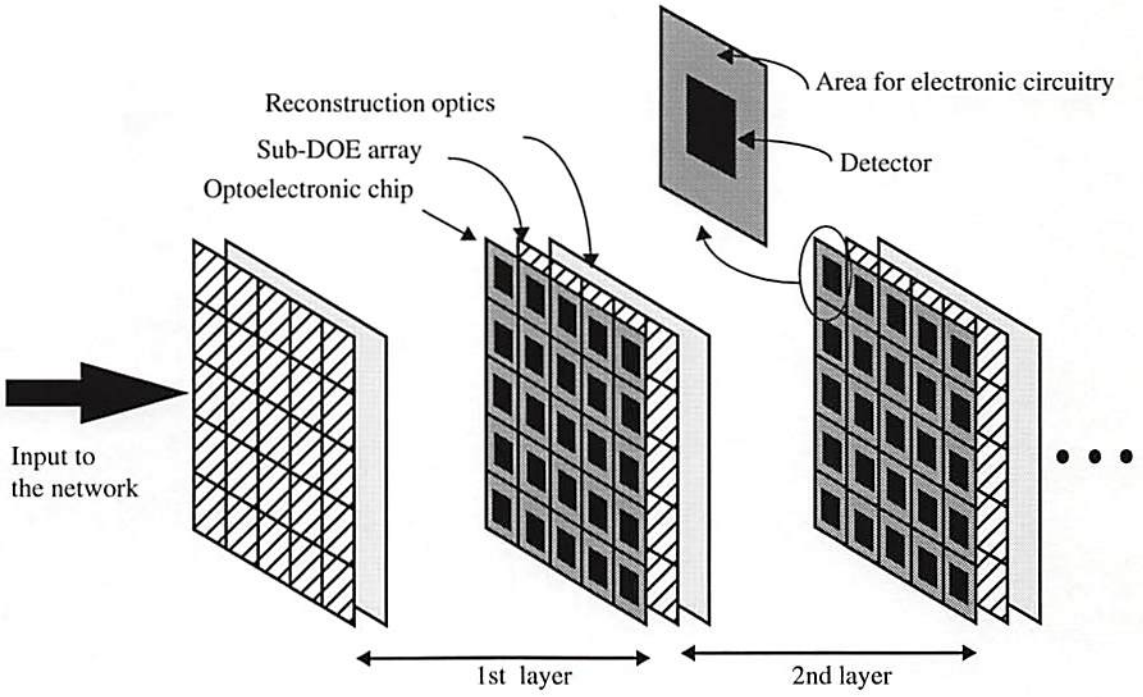

Figure 4-1: Schematic diagram for optoelectronic implementation of multilayer feed-forward neural networks. Note that the light emitter array (or modulator array and readout-illumination optics) on the back side of each optoelectronic chip is not shown. Note that the figure is not to scale.

contains four different functional elements. The first functional element (from left to right in Figure 4-1) is an optoelectronic chip comprising a two-dimensional array of contiguous

nodes. Each node consists of one detector with some electronic circuitry for post-detection processing on the input side of the chip, and one active (emitter such as a vertical cavity surface emitting laser) or passive (modulator) optical source on the output side of the chip. Following the optoelectronic chip are some collimation optics to collimate the output light from the optical sources, in turn illuminating the next functional element, a two-dimensional array of contiguous sub-DOEs. The sub-DOE array together with reconstruction optics provides the space-variant weighted interconnections to the next layer. in this thesis, we assume that the lateral physical size of a node on the optoelectronic chip is equal to the lateral physical size of a sub-DOE, and their shapes are both square. We further assume that the optoelectronic chip, the collimation optics, the sub-DOE array, and the reconstruction optics all have the same lateral physical sizes to maintain cascadability over layers. Therefore, the center-to-center spacing, or pitch, of sub-DOEs is equal to the pitch of detectors (or nodes) on the optoelectronic chip. Depending on the kind of neuron units (unipolar or bipolar) implemented in the system, each neuron unit can consist of one or more nodes on the optoelectronic chip. Without loss of generality, we assume that each neuron unit consists of one node in this thesis. In order to have space-variant interconnections, each node must have a dedicated sub-DOE to provide the weighted fan-out.

The operations of the multilayer feed-forward space-variant interconnection system can be described as follows. The input scene to the system is detected by the two-dimensional array of detectors on the optoelectronic chip in the first layer. The detected information is processed (e. g., thresholded) by the electronic circuitry of each node.

After the post-detection processing, the result is encoded onto the output light beam of the corresponding optical source on the back side of the optoelectronic chip of the first layer. The light beams of the source array are then collimated to illuminate the sub-DOE array of the first layer. The sub-DOE array and reconstruction optics then provide the weighted interconnection fanout to the second layer. Therefore, each detector in the second layer will receive a fan-in of weighted signals (weighted by the corresponding interconnection weight) from the first layer. We assume that the degree of fanout from each sub-DOE is the same as the degree of fan-in for each detector. These operations of detection, processing, and weighted fanout are carried out through each of the following layers in the system. Note that except for the first layer of the system, each detector acts as a neuron unit receptor and collects a weighted sum of the inputs to that detector.

in this thesis, we focus on the interconnection portion of the system (i.e. the sub-DOE array and reconstruction optics). In our analysis, there are many interrelated parameters. For the work described in this thesis, we choose a set of independent system parameters as follows (some of the definitions given below are repeated from those given in Chapter 3 for the reader's convenience):

N: number of nodes in one dimension, assumed equal in the input and output planes, also assumed even

λ: optical wavelength

Δ: lateral width of each phase element, assumed square

P: lateral width of each detector, assumed square

B: oversampling ratio (number of phase elements per sub-DOE period in each dimension divided by the fan-out in each dimension)

C: compression ratio (number of periods in each dimension of a sub-DOE)

M: fanout from each node in the input plane in one dimension for limited-fanout architectures, assumed odd (explained below in Chapter 5)

Y: crosstalk reduction parameter in each dimension for limited-fanout architectures (explained below in Chapter 5)

and consider all other parameters as dependent on these. Throughout this and next chapters when we discuss the effects of changing one of the independent parameters, it is implied (even though not always explicitly stated) that the other independent parameters are being held constant. We also assume that the independent parameters Δ, B, C and M for every sub-DOE is the same for a given architecture. Therefore, all the sub-DOEs have the same period length, T, for a given architecture.

4.2 Fully Connected Space-Variant Interconnection Architecture

In a fully connected space-variant interconnection system, each neuron unit in any layer must have the weighted interconnections to all the neuron units in the next layer. The weighted interconnection patterns will be different for some (or all) of the neuron units in the same layer in order to be space-variant. A possible realization of a layer for a fully connected space-variant interconnection system is shown in Figure 4-2 [42, 61, 62]. Note

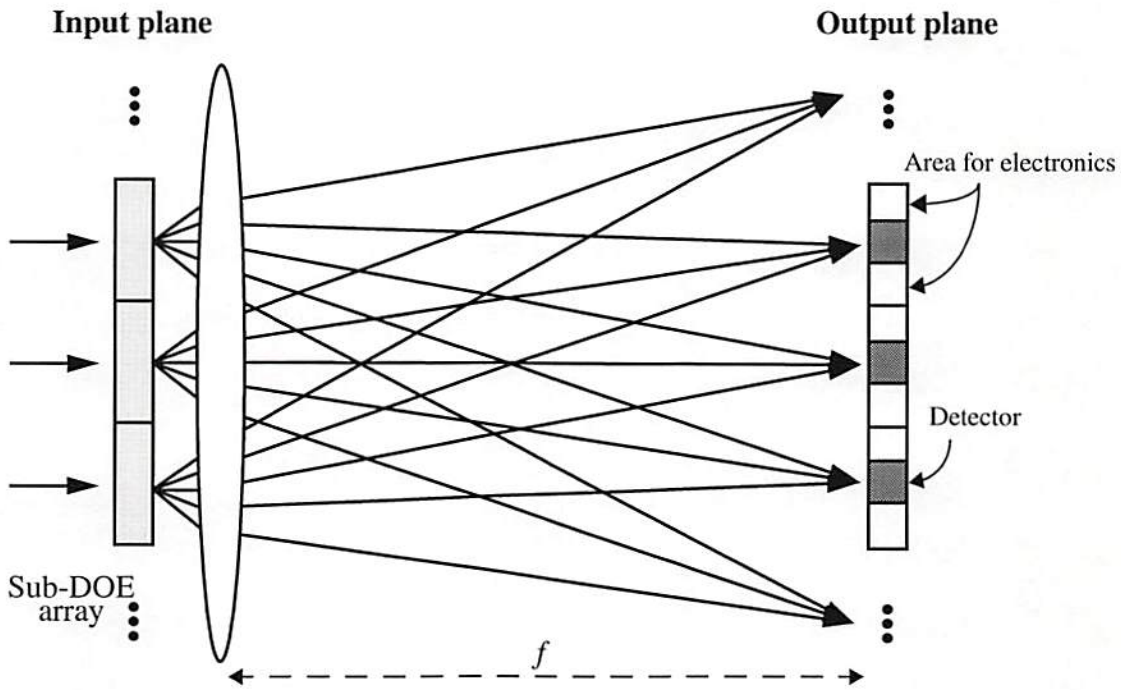

Figure 4-2: Optical architecture for fully connected space-variant interconnection systems (shown in one dimension, not drawn to scale). In the input plane, there is an array of sub-DOEs. In the output plane, there is an array of detectors. The solid lines represent the principal rays of the diffraction orders. The space between detectors is for electronic circuitry.

that only the sub-DOE array, the reconstruction optics, and the detectors of the optoelectronic chip are shown. As shown in Figure 4-2, we refer to the sub-DOE plane as the input plane of the interconnection system and the detector plane as the output plane of the interconnection system (or the reconstruction plane of the DOE). Here, the reconstruction optics is a Fourier transform lens, whose aperture accommodates light from all sub-DOE elements. Each sub-DOE, together with the Fourier transform lens, reconstructs a set of $N \times N$ weighted interconnections (synapses) to all of the detectors in the output plane. Because of the shift property of the Fourier transform, the reconstruction

patterns from all the sub-DOEs will coincide on the detector array. Therefore, each detector in the output plane has a fan-in of N^2. The fully connected architecture has the advantage of global interconnections. However, the space-bandwidth product ($SBWP$) requirement of each sub-DOE in the system will be proportional to N^2 and the space-bandwidth product of the entire sub-DOE array will be proportional to N^4 [14]. This large space-bandwidth product requirement may limit the number of neuron units that can be implemented in the system.

Using the independent parameters given above, the one-dimensional period length of each sub-DOE, T, is equal to $BN\Delta$, the size of each node (or sub-DOE) is $CT \times CT$, and the area available for electronic circuitry in each node is equal to the node area minus the detector area, $(CT)^2 - P^2$. An optoelectronic chip with more intelligence will require more area for electronic processing, hence, for a fixed detector size, a larger C or T is needed. A smaller detector size is usually preferred since a smaller detector has a smaller capacitance, hence can operate at higher speed. Moreover, the choice of the detector size affects the interconnection crosstalk level in the system as will be shown later. The required space-bandwidth product (numbers of phase elements) of each sub-DOE is $(CBN)^2$ and the space-bandwidth product of the entire sub-DOE array is $N^2(CBN)^2$. The width of the sub-DOE array (or the optoelectronic chip) is $N(CBN\Delta)$. The propagation length (Z or f) between layers can be calculated by forcing the spot spacing (S), equal to the detector spacing (CT), i. e.,

$$\frac{\lambda f}{T} = CT. \tag{4-1}$$

Solving Eq. (4-1) for f and substituting T with $BN\Delta$, we have

$$Z = f = \frac{C}{\lambda}(BN\Delta)^2. \tag{4-2}$$

Table 4-1 shows the required space-bandwidth product (*SBWP*) of the entire sub-DOE array in each layer, lateral dimension (W) of the sub-DOE array, propagation length (Z) between layers, system aspect ratio (W/Z), *f-number* of the lens, number of interconnections (NI) in each layer, volume (V) of interconnection system between layers, and interconnection density per unit volume. Note that the *f-number* of the lens is the reciprocal of the aspect ratio. The *f-number* of the lens can be small for some choices of independent parameters.

Table 4-1: Characteristics of fully connected space-variant interconnections

SBWP of sub-DOE array	Width of sub-DOE array (W)	Propagation length (Z)	Aspect ratio (W/Z)	*f-number* of lens	NI	System volume (ZW^2)	Connection density (NI/ZW^2)
$C^2B^2N^4$	$CB\Delta N^2$	$\frac{C}{\lambda}(BN\Delta)^2$	$\left(\frac{1}{B}\right)\left(\frac{\lambda}{\Delta}\right)$	$B\left(\frac{\Delta}{\lambda}\right)$	N^4	$\frac{C^3B^4N^6\Delta^4}{\lambda}$	$\left(\frac{\lambda}{C^3B^4\Delta^4}\right)\left(\frac{1}{N^2}\right)$

4.3 Definition and Origin of Interconnection Crosstalk

In order to study the interconnection crosstalk of the fully connected and limited-fanout systems, an appropriate measure of interconnection crosstalk is needed. In digital systems, maximum noise over minimum signal provides a worst-case measure of crosstalk [61]. This crosstalk measure is appropriate when essentially error-free performance of each bit operation is required. For the neural case, such a measure isn't so appropriate. Instead, a crosstalk measure that scales more reasonably with the size of the network and with the degree of fan-in to each neuron unit is more appropriate [38]. In this case, we wouldn't require a neuron unit with high fan-in to give error-free performance for all possible input states, but instead use a measure that is more indicative of typical crosstalk than worst-case crosstalk. Let I_0 represent the maximum possible signal power and I_1 represent the maximum possible noise power in an output-plane detector, then we can define crosstalk, β, as:

$$\beta \equiv \frac{I_1}{I_0}. \tag{4-3}$$

Here, noise is referring to any light other than the signals that enter a detector region. Note that, in both fully connected and limited-fanout space-variant systems, maximum noise and maximum signals occur when all the input nodes are fully on. in this thesis, we consider the interconnection crosstalk caused by sub-DOEs only and analyze its characteristics. Sub-DOEs can provide the desired signal connections for the systems, but the development below shows that they can also cause significant interconnection crosstalk. As mentioned

in Chapter 3 and shown in Figure 3-2, besides the signals, the reconstructed intensity pattern also contains signal sidelobes (SS), spurious diffraction orders (SDO), and non-paraxial region noise (NRN). All these three components are possible sources of interconnection crosstalk. There is a fourth crosstalk component that results from the fact that not all of the reconstructed spot energy lies within the detector area. Instead, a small amount of the energy extends outside the intended detector and falls on other detectors. This crosstalk component arises from the tails of reconstructed spots.

For the purpose of crosstalk analysis, we assume the input light beams are mutually incoherent. Consequently, the total light intensity on a detector is equal to the sum of the intensities of all the diffraction orders on that detector. If the input light beams are coherent, they will create interference fringes on the aperture of each detector. Because a detector takes a spatial average over its active area, the result is approximately equivalent to an incoherent sum under conditions of interest in this thesis [61].

4.4 Interconnection Crosstalk in a Fully Connected Space-Variant System

In a fully connected space-variant interconnection system, the shift property of Fourier transforms will cause the reconstruction patterns from all the sub-DOEs to coincide on the detector array. Each detector receives N^2 weighted signals (weighted by the corresponding interconnection weights). Signal sidelobes, spurious diffraction orders, and non-paraxial region noise will all fall outside the detector array area. The only crosstalk component due

to the theoretical DOE reconstruction is the tails of the reconstructed spots. Hence, the maximum signal power should be much higher than the maximum noise power at any detector. This implies a low level of interconnection crosstalk.

In general, the crosstalk level will be different at different detector locations for a given system. In this thesis, we focus on the crosstalk level for the detector located at the center of the output plane since, on average, it receives the largest number of the tails of the other reconstructed spots. The maximum signal power, l_0, and maximum noise power, l_1, can be calculated as follows. From Eq. (3-12) and the fact that the reconstructed spot pitch is equal to the detector pitch, it is clear that each input node in the input plane will generate N^2 weighted signals, each centered at $\left(\frac{m\lambda f}{T}, \frac{n\lambda f}{T}\right)$ in the output plane, where (m, n) represents the indices of the detector locations in the output plane $(-\frac{N}{2} \le m \le \frac{N}{2} - 1,\ -\frac{N}{2} \le n \le \frac{N}{2} - 1)$. Each weighted signal will have an intensity profile of

$$I_{(m,n)\,(i,j)} = V_{(i,j)} W_{(m,n)(i,j)} \mathrm{sinc}^2\left[\frac{CT}{\lambda f}\left(x_f - \frac{m\lambda f}{T}\right)\right] \mathrm{sinc}^2\left[\frac{CT}{\lambda f}\left(y_f - \frac{n\lambda f}{T}\right)\right], \qquad (4\text{-}4)$$

where $V_{(i,j)}$ represents the signal intensity illuminating the (i, j)'th sub-DOE and (i, j) is the indices of the sub-DOE (or node location) in the input plane $(-\frac{N}{2} \le i \le \frac{N}{2} - 1,\ -\frac{N}{2} \le j \le \frac{N}{2} - 1)$. The $W_{(m,n)(i,j)}$ term in Eq. (4-4) represents the weighted interconnection from the (i, j)'th sub-DOE in the input plane to the (m, n)'th detector in the output plane. At the center detector location ($m = 0$ and $n = 0$), the intensity profile of each weighted signal is $V_{(i,j)} W_{(0,0)(i,j)} \mathrm{sinc}^2 [CTx_f / (\lambda f)] \mathrm{sinc}^2 [CTy_f / (\lambda f)]$. Since the

system is fully connected, the intensity profile of the total weighted signals at the center detector is

$$I_{(0,0)}(x_f, y_f) = \sum_{i=\frac{-N}{2}}^{\frac{N}{2}-1} \sum_{j=\frac{-N}{2}}^{\frac{N}{2}-1} V_{(i,j)} W_{(0,0)(i,j)} \mathrm{sinc}^2\left(\frac{CTx_f}{\lambda f}\right) \mathrm{sinc}^2\left(\frac{CTy_f}{\lambda f}\right), \tag{4-5}$$

and the total signal power on the center detector will be integral of Eq. (4-5) over the $P \times P$ region of the center detector,

$$\tilde{I}_0 = \int_{-\frac{P}{2}}^{\frac{P}{2}} \int_{-\frac{P}{2}}^{\frac{P}{2}} \sum_{i=\frac{-N}{2}}^{\frac{N}{2}-1} \sum_{j=\frac{-N}{2}}^{\frac{N}{2}-1} V_{(i,j)} W_{(0,0)(i,j)} \mathrm{sinc}^2\left(\frac{CTx_f}{\lambda f}\right) \mathrm{sinc}^2\left(\frac{CTy_f}{\lambda f}\right) dx_f dy_f. \tag{4-6}$$

Since the maximum signal power and maximum noise power occur when all the input nodes are fully on, we will use the largest value of each input signal to calculate the maximum signal power from Eq. (4-6) with the assumption of $0 \le V_{(i,j)} \le V_{max}$. Physically, V_{max} is the situation where the light intensity illuminating each sub-DOE is identical and is equal to the maximum that could occur in system operation. Further, we will use the average connection weight of the layer, $\langle W \rangle$ to calculate the average crosstalk in the system. The average maximum signal, l_0, at the center detector location can be obtained directly from Eq. (4-6) as

$$l_0 = N^2 V_{max} \langle W \rangle \int_{-\frac{P}{2}}^{\frac{P}{2}} \int_{-\frac{P}{2}}^{\frac{P}{2}} \mathrm{sinc}^2\left(\frac{CTx_f}{\lambda f}\right) \mathrm{sinc}^2\left(\frac{CTy_f}{\lambda f}\right) dx_f dy_f. \tag{4-7}$$

At an off-center detector location (m, n fixed and $(m, n) \neq (0, 0)$), each weighted signal will have an intensity profile as shown in Eq. (4-5) and the total intensity profile of N^2 such weighted signals at that detector location will be

$$I_{(m,n)}(x_f, y_f) = \sum_{i=\frac{-N}{2}}^{\frac{N}{2}-1} \sum_{j=\frac{-N}{2}}^{\frac{N}{2}-1} V_{(i,j)} W_{(m,n)(i,j)} \mathrm{sinc}^2\left[\frac{CT}{\lambda f}\left(x_f - \frac{m\lambda f}{T}\right)\right] \mathrm{sinc}^2\left[\frac{CT}{\lambda f}\left(y_f - \frac{n\lambda f}{T}\right)\right] . \tag{4-8}$$

The center detector will receive some portion of the tail of the intensity profile in Eq. (4-8) and the power it collects can be calculated by the integral of Eq. (4-8) over the $P \times P$ region of the center detector,

$$\tilde{l}_1' = \int_{-\frac{P}{2}}^{\frac{P}{2}} \int_{-\frac{P}{2}}^{\frac{P}{2}} I_{(m,n)}(x_f, y_f)\, dx_f dy_f . \tag{4-9}$$

The total tail power collected by the center detector can be calculated by summing Eq. (4-9) over all the off-center detector locations as

$$\tilde{I}_1 = \sum_{\substack{m=\frac{-N}{2} \\ (m,n)\neq(0,0)}}^{\frac{N}{2}-1} \sum_{n=\frac{-N}{2}}^{\frac{N}{2}-1} \int_{-\frac{P}{2}}^{\frac{P}{2}} \int_{-\frac{P}{2}}^{\frac{P}{2}} I_{(m,n)}(x_f, y_f)\, dx_f dy_f \quad . \tag{4-10}$$

Using V_{max} and $\langle W \rangle$, the maximum noise power, I_1, at the center detector location can be readily obtained from Eq. (4-10) as

$$I_1 = \sum_{\substack{m=\frac{-N}{2} \\ (m,n)\neq(0,0)}}^{\frac{N}{2}-1} \sum_{n=\frac{-N}{2}}^{\frac{N}{2}-1} N^2 V_{max} \langle W \rangle \int_{-\frac{P}{2}}^{\frac{P}{2}} \int_{-\frac{P}{2}}^{\frac{P}{2}} \operatorname{sinc}^2\left[\frac{CT}{\lambda f}\left(x_f - \frac{m\lambda f}{T}\right)\right] \operatorname{sinc}^2\left[\frac{CT}{\lambda f}\left(y_f - \frac{n\lambda f}{T}\right)\right] dx_f dy_f \quad . \tag{4-11}$$

In general, the near-center reconstructed spots contribute much more than the far-from-center reconstructed spots to the maximum noise power, I_1, at the center detector location. Therefore, the crosstalk in a fully connected system should only increase slightly as the number of the nodes in each dimension, N, increases for a sufficient large value of N. The crosstalk in the system should decrease as the separation of the adjacent reconstructed spots (relative to the reconstructed spot size) increases. As mentioned in Chapter 3, this separation is a function of the compression ratio, C, (the number of periods in each dimension of a sub-DOE) only. Furthermore, the crosstalk should also depend on the detector size (relative to reconstructed spot size) in the system. For the purpose of

crosstalk analysis, we define the normalized detector width, K, as the ratio of the detector width, P, to the reconstructed spot width, $\frac{2\lambda f}{CT}$. The value $K = 1$ indicates that the width of the reconstructed spot is equal to the width of the detector and $K > 1$ indicates that the detector width is greater than the reconstructed spot width. With change of variables and $K = \frac{PCT}{2\lambda f}$, Equation (4-7) and (4-11) can be simplified to:

$$l_0 = DN^2 V_{max} \langle W \rangle \int_{-K}^{K} \int_{-K}^{K} \operatorname{sinc}^2(x) \operatorname{sinc}^2(y)\, dxdy, \tag{4-12}$$

and

$$l_1 = D \sum_{\substack{m = \frac{-N}{2} \\ (m,n) \neq (0,0)}}^{\frac{N}{2}-1} \sum_{n = \frac{-N}{2}}^{\frac{N}{2}-1} N^2 V_{max} \langle W \rangle \int_{-K}^{K} \int_{-K}^{K} \operatorname{sinc}^2(x - mC) \operatorname{sinc}^2(y - nC)\, dxdy, \tag{4-13}$$

where D ($= \frac{\lambda^2 f^2}{C^2 T^2}$) is a constant.

To study the effect of the number of the nodes in the system on the crosstalk level, we calculated the maximum signal power, l_0, and maximum noise power, l_1, from Eq. (4-12) and (4-13) as a function of N with compression ratio as a parameter ($C = 2, 3, 4$), and fixed normalized detector width ($K = 1$). As shown in Figure 4-3, the crosstalk level in the system (l_0/l_1) only increases slightly for $N \geq 30$ with fixed values of C and K. To study the effects of compression ratio and normalized detector width, $N = 100$ is assumed

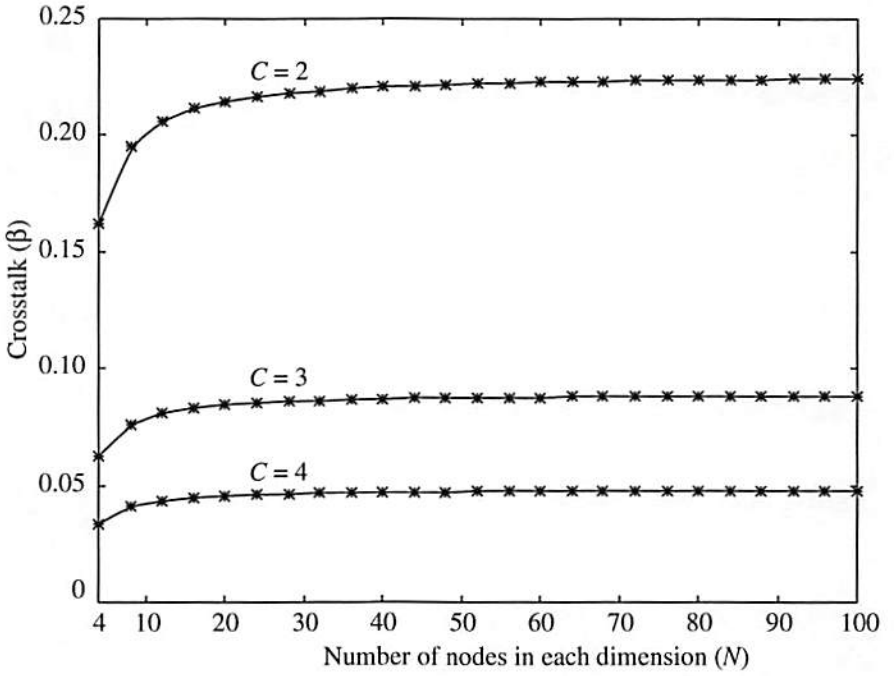

Figure 4-3: Crosstalk of a fully connected space-variant interconnection system as a function of N (number of nodes in each dimension) for K (normalized detector width) = 1 and C (compression ratio) = 2, 3, 4.

for calculating the asymptotic crosstalk levels. Shown in Figure 8 is the crosstalk level calculated from Eq. (4-12) and (4-13) as a function of compression ratio, C, and normalized detector width, K, with $N = 100$. Because the size of a sub-DOE is assumed to be equal to the size of a node, the size of a detector must be smaller than the size of a sub-DOE. This constrains the allowable values of C and K. This constraint can be derived as follows. Using Eq. (4-2) for the propagation length, the reconstructed spot width, $\frac{2\lambda f}{CT}$, can be found as

$$S_S = 2(BN\Delta) = 2T, \tag{4-14}$$

where $T\,(=\,BN\Delta)$ is the period length of each sub-DOE in each dimension. The width of the detector, P, can then be expressed as $2KT$. Because the with of the detector (P) can not be greater than the width of the sub-DOE (CT), C and K are constrained by the relation $C \geq 2K$. In addition, when $C = 2K$, the detector size is equal to the sub-DOE size and there is no space for electronic processing circuitry. Therefore, $C > 2K$ is required in a practical system. Accordingly, a minimum C value of $2K+1$ is assumed for each different K value in Figure 4-4. As shown, for a constant normalized detector width, K, the crosstalk decreases as the compression ratio increases because the reconstructed spots are better separated. On the other hand, the crosstalk level increases as normalized detector width, K, increases for a fixed value of compression ratio. In this case, the detector width (normalized by the reconstructed spot width) is increasing, while the reconstructed spot spacing (also normalized by the reconstructed spot width) is held constant. This change results in an increase in crosstalk because, when the detectors are enlarged, they collect more signal light as well as more tail noise, but the fraction of increase for the noise is larger than that for the signals. Therefore, reducing the normalized detector width, K, will reduce the crosstalk in the system. This is true even for a detector size that is smaller than the reconstructed spot size. As shown in Figure 4-4, the $K = 0.5$ curve has the lowest crosstalk level although some portion of the signal light is wasted. In general, a lower crosstalk level can be obtained either by increasing the compression ratio, C, (hence, a larger system size, since system lateral width W and propagation length Z are each proportional to C) or reducing the detector size.

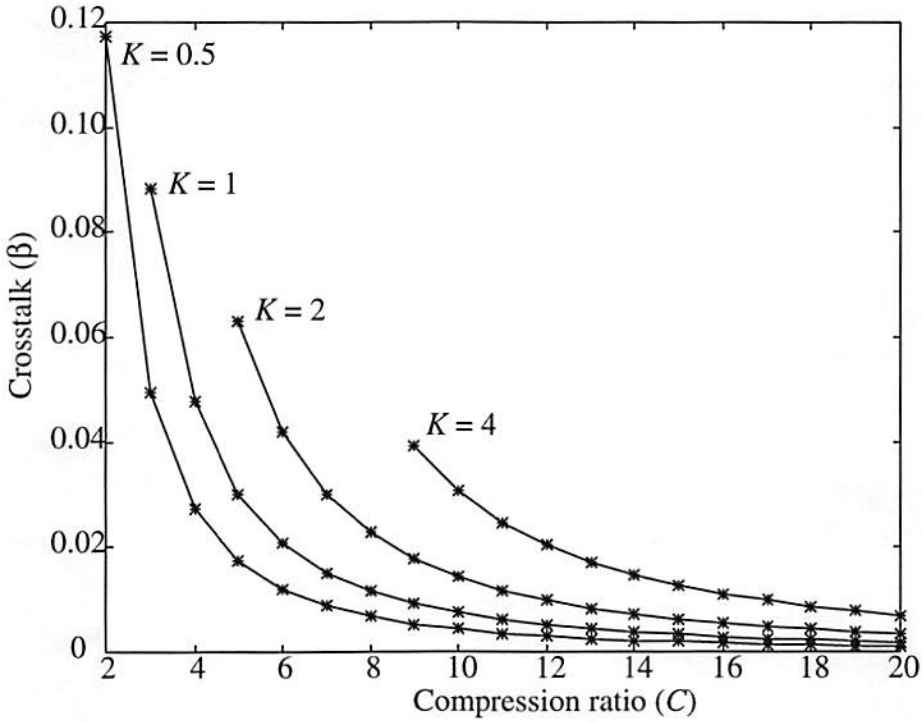

Figure 4-4: Crosstalk of a fully connected space-variant interconnection system as a function of C (compression ratio) for K (normalized detector width) = 0.5, 1, 2, 4, assuming the number of nodes in each dimension (N) = 100.

Chapter 5

DOE-Based Multi-Layer Feed-Forward Limited-Fanout Space-Variant Interconnection Architecture and Crosstalk Analysis

In this chapter, we investigate the interconnection architectures for optoelectronic implementations of limited-fanout neural networks. In Section 5.1, the important characteristics such as propagation length, lateral dimension, system aspect ratio, interconnection density, and space-bandwidth product requirement will be derived and compared with the characteristics of a fully connected system. The interconnection crosstalk in such a system will then be analyzed and simulated in Section 5.2. The limited-fanout architecture is more suited for physically compact but computationally large-scale multilayer neural networks because the required propagation length is much smaller than the propagation length of the fully connected architecture. However, the interconnection crosstalk of the limited-fanout space-variant architecture is likely too high for many applications, due to noise from the reconstructions of the surrounding diffractive optical elements. Therefore, a crosstalk reduction technique based on a modified design for diffractive optical elements is proposed in Section 5.3. The key is to rearrange the

reconstruction pattern of each diffractive optical element such that less noise lands on each detector region (and more noise lands between detector region.) This technique is verified by simulating one layer of an interconnection system with 128×128 input nodes, 128×128 output nodes, and 5×5 nearest neighbor weighted connections from each input node to the output nodes. In Section 5.4, the additional advantages and possible limitations of the crosstalk reduction technique are discussed. In addition to the crosstalk reduction, it will be shown that the technique can also be used to reduce the propagation length and system volume.

5.1 Limited-Fanout Space-Variant Interconnection Architecture

If we replace the bulk lens in a fully connected system with an array of microlenses and restrict each sub-DOE fan-out pattern to a neighborhood region in the next layer, then the system becomes a limited-fanout space-variant interconnection system (Figure 5-1) [10, 12]. For this limited-fanout neural network interconnection, each sub-DOE stores one weighted fanout pattern that connects to the nearest $M \times M$ neighbors in the output plane. Each detector in the output plane has a fan-in of M^2. Each period length of a sub-DOE (T) now changes from $BN\Delta$ to $BM\Delta$. The space-bandwidth product requirement of each sub-DOE becomes $(CBM)^2$ and the space-bandwidth product of the entire sub-DOE array is $N^2(CBM)^2$. The propagation length (Z) can be calculated by forcing the spot spacing equal to the detector spacing and substituting T with $BM\Delta$. The propagation length is

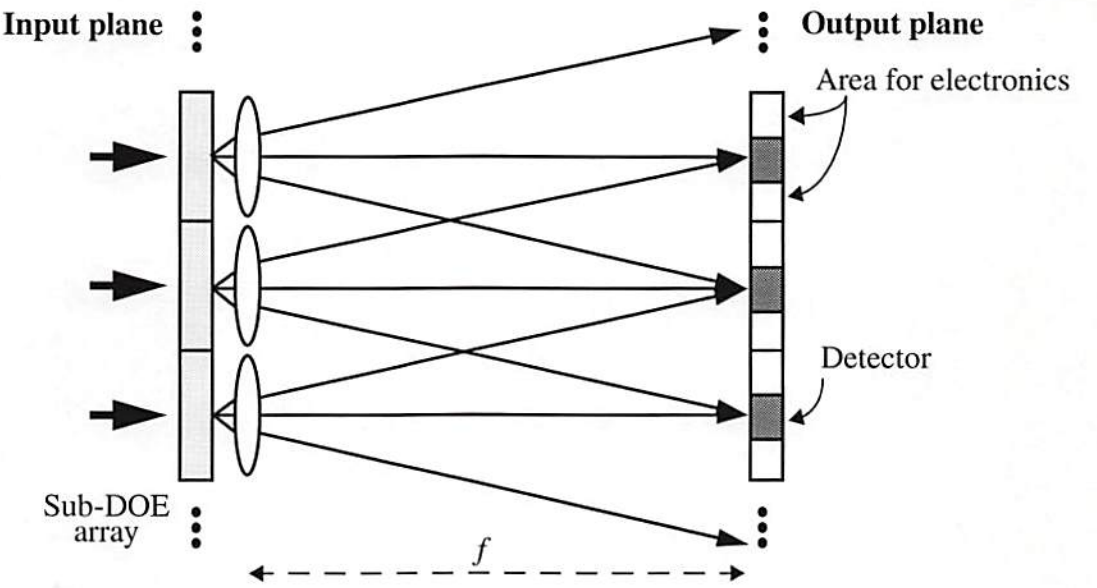

Figure 5-1: Optical architecture for limited-fanout space-variant interconnection systems (shown in one dimension, not drawn to scale). In the input plane, there is an array of sub-DOEs. In the output plane, there is an array of detectors. The solid lines represent the principal rays of the diffraction orders. The space between detectors is for electronic circuitry.

$$Z = f = \frac{C}{\lambda}(BM\Delta)^2. \tag{5-1}$$

If N is much larger than M, then because of the localized interconnections, the space-bandwidth product requirements will be much less and the propagation length will be much shorter than those of the fully connected system. In addition, the low *f-number* requirement on the lens is now relieved since the limited-fanout space-variant system has one microlens

for each sub-DOE. The characteristics of the fully connected and limited-fanout systems are shown in Table 5-1.

Table 5-1: Characteristics comparison between fully connected systems (the first row) and limited-fanout systems (the second row)

SBWP of sub-DOE array	W	Z	Aspect ratio *(W/Z)*	*f-number* of lens	NI	System volume (ZW^2)	Connection density (NI/ZW^2)
$C^2B^2N^4$	$CB\Delta N^2$	$\frac{C}{\lambda}(BN\Delta)^2$	$\left(\frac{1}{B}\right)\left(\frac{\lambda}{\Delta}\right)$	$B\left(\frac{\Delta}{\lambda}\right)$	N^4	$\frac{C^3B^4N^6\Delta^4}{\lambda}$	$\left(\frac{\lambda}{C^3B^4\Delta^4}\right)\left(\frac{1}{N^2}\right)$
$C^2B^2N^2M^2$	$CB\Delta MN$	$\frac{C}{\lambda}(BM\Delta)^2$	$\left(\frac{1}{B}\right)\left(\frac{\lambda}{\Delta}\right)\left(\frac{N}{M}\right)$	$MB\left(\frac{\Delta}{\lambda}\right)$	M^2N^2	$\frac{C^3B^4N^2M^4\Delta^4}{\lambda}$	$\left(\frac{\lambda}{C^3B^4\Delta^4}\right)\left(\frac{1}{M^2}\right)$

From Table 5-1, it can be seen that the propagation length, system volume, and connection density in one layer are proportional to M^2, N^2M^4, and $1/M^2$ for the limited-fanout space-variant system, and are proportional to N^2, N^6, and $1/N^2$ for the fully connected space-variant system, respectively. Therefore, for constant values of the other independent parameters, the limited-fanout space-variant system will have a much shorter propagation length as well as a much smaller system volume if N is much larger than M. It is also interesting to notice that although the fully connected space-variant system has more interconnections, the limited-fanout system has higher connection density. As an example, Table 5-2 shows the resulting interconnect characteristics for the two systems calculated using the following values: $N = 100$, $\lambda = 850\text{nm}$, $\Delta = 2\mu\text{m}$, $B = 2$, $C = 2$, and $M = 10$. It is clear that the fully connected architecture is limited to a much smaller number of input and output nodes although it can provide global interconnections. On the other hand, the limited-fanout system can be used in a much smaller volume and is

therefore better suited for compact multilayer structures. However, the crosstalk levels of the limited-fanout system can be high due to reconstruction noise (e.g., signal sidelobes and spurious diffraction orders) of each sub-DOE. In the next section, the crosstalk levels of the limited-fanout system will be examined.

Table 5-2: An example for comparing fully connected S-V systems and limited-fanout S-V systems with parameters: $N = 100$, $\lambda = 850$ nm, $\Delta = 2$ μm, $P = 20$ μm, $B = 2$, $C = 2$, and $M = 10$.

Type	*SBWP* of sub-DOE array	Width of sub-DOE array (*W*) (*mm*)	Propagation length (*Z*) (*mm*)	Aspect ratio (*W/Z*)	*f-number* of lens	*NI*	System volume (mm^3)	Connection density (connections/mm^3)
Fully connected	1.6×10^9	80	376	0.21	4.7	10^8	2.4×10^6	4.2×10^1
Limited-fanout	1.6×10^7	8	3.76	2.1	47	10^6	2.4×10^2	4.2×10^3

5.2 Interconnection Crosstalk in a Limited-Fanout Space-Variant System

In a limited-fanout S-V interconnection system, each sub-DOE has its own Fourier transform lens. Therefore, the reconstructed pattern of a sub-DOE will be centered at the optical axis of its own reconstruction lens. Under the reconstruction model of a sub-DOE from Chapter 3, each sub-DOE will reconstruct in the output plane M^2 desired interconnections and other undesired high diffraction orders terms (signal sidelobes and spurious diffraction orders) within its paraxial reconstruction region centered at the optical axis of its own lens as shown in Figure 5-2). The reconstruction outside the paraxial region of each sub-DOE will depend on the aberration properties of the Fourier transform lens, and is modeled as an uniform-intensity blur (denoted as non-paraxial region noise NRN). Non-paraxial region noise is normally much weaker than the signal sidelobes and spurious diffraction orders from the same sub-DOE. Each output plane detector will receive M^2 desired weighted signals and many undesired diffraction order terms from the many paraxial reconstruction regions of input plane sub-DOEs. In addition, each detector will also receive NRN noise from the many different reconstruction regions of input plane sub-DOEs and many reconstruction tails from all other detector locations. The strength of the signal sidelobes, spurious diffraction orders, and non-paraxial region noise from one well-designed sub-DOE are usually much lower than the strength of the signal diffraction orders of the sub-DOE. However, as an increasing number of reconstruction patterns are superimposed (with relative shifting) on the detector array, the accumulated noise level at

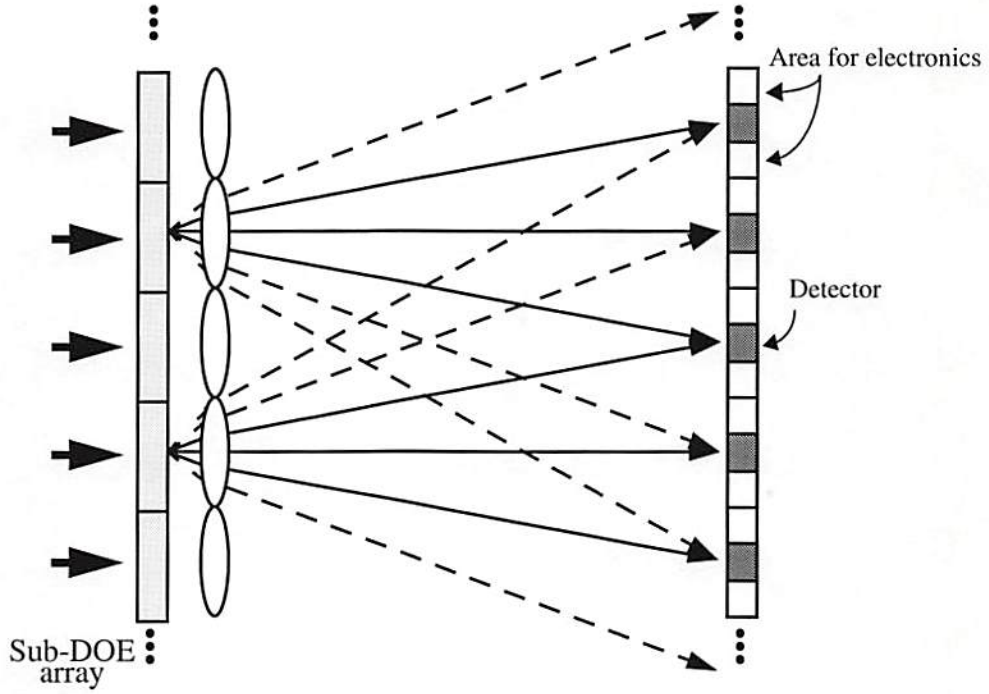

Figure 5-2: Sources of interconnection crosstalk for limited-fanout space-variant interconnection systems. Solid lines represent the desired connections and dashed lines represent the noise connections (signal sidelobes or spurious diffraction orders). Only two sets of sub-DOE reconstructions are shown.

an output plane detector can become comparable to the sum of the M^2 weighted signals at that detector. Consequently, the interconnection crosstalk can be much higher than the crosstalk in a fully connected system.

The expressions for the maximum signal power and four different components of maximum noise power are developed as follows: We still focus the crosstalk analysis for the center detector in the output plane. Because each sub-DOE generates M^2 interconnections to the output plane, the center detector will receive one desired interconnection from each of the center $M \times M$ sub-DOEs. We refer to this set of sub-

DOEs as the paraxial signal reception sub-DOEs, X_1, of the center detector. If the paraxial reconstruction region of each sub-DOE is of physical size $L \times L$, then any sub-DOE located within the center $L \times L$ region of the input plane but outside the region containing X_1 will have an undesired interconnection (either signal sidelobes or spurious diffraction orders) to the center detector. We refer to the sub-DOEs in this region that generate signal sidelobes as the paraxial signal sidelobe reception sub-DOEs, X_2, of the center detector. We also refer to the sub-DOEs in this region that generate spurious diffraction orders as the paraxial spurious diffraction order reception sub-DOEs, X_3, of the center detector. The center detector is in the non-paraxial reconstruction region of any sub-DOE not in X_1, X_2 and X_3. We refer to this set of sub-DOEs as the non-paraxial reception sub-DOEs, X_4, of the center detector. The center detector receives M^2 desired interconnections (each denoted as $W_{(0,0)(i,j)}$) from the sub-DOEs in X_1 group, many signal-sidelobes (each denoted as $W^{SS}_{(0,0)(i,j)}$) from the sub-DOEs in X_2 group, and spurious-diffraction-order interconnections (each denoted as $W^{SDO}_{(0,0)(i,j)}$) from the sub-DOEs in X_3 group. The intensity profiles at the center detector from these three different sources can be written as

$$I^{Signal}_{(0,0)}(x_f, y_f) = \sum_{i,j \in X_1} \sum V_{max} W_{(0,0)(i,j)} \mathrm{sinc}^2\left(\frac{CTx_f}{\lambda f}\right) \mathrm{sinc}^2\left(\frac{CTy_f}{\lambda f}\right), \qquad (5\text{-}2)$$

$$I^{SS}_{(0,0)}(x_f, y_f) = \sum_{i,j \in X_2} \sum V_{max} W^{SS}_{(0,0)(i,j)} \mathrm{sinc}^2\left(\frac{CTx_f}{\lambda f}\right) \mathrm{sinc}^2\left(\frac{CTy_f}{\lambda f}\right), \qquad (5\text{-}3)$$

and

$$I^{SDO}_{(0,0)}(x_f, y_f) = \sum_{i,j \in X_3} \sum V_{max} W^{SDO}_{(0,0)(i,j)} \mathrm{sinc}^2\left(\frac{CTx_f}{\lambda f}\right) \mathrm{sinc}^2\left(\frac{CTy_f}{\lambda f}\right). \quad (5\text{-}4)$$

Note that we have used V_{max} for the signal level of each input node since the maximum noise and maximum signals occur when all the input nodes are fully on. The intensity profile at the center detector due to these components is:

$$I_{(0,0)}(x_f, y_f) = I^{Signal}_{(0,0)}(x_f, y_f) + I^{SS}_{(0,0)}(x_f, y_f) + I^{SDO}_{(0,0)}(x_f, y_f). \quad (5\text{-}5)$$

The maximum signal power at the center detector is the integral of the intensity profiles of M^2 weighted signals in Eq. (5-2) over the $P \times P$ region of the center detector as follows

$$I_0 = \int_{-\frac{P}{2}}^{\frac{P}{2}} \int_{-\frac{P}{2}}^{\frac{P}{2}} \sum_{i,j \in X_1} \sum V_{max} W_{(0,0)(i,j)} \mathrm{sinc}^2\left(\frac{CTx_f}{\lambda f}\right) \mathrm{sinc}^2\left(\frac{CTy_f}{\lambda f}\right) dx_f dy_f. \quad (5\text{-}6)$$

Using the normalized detector width, K, and some variable substitutions, the maximum signal power at the center detector can be simplified to

$$l_0 = D \int_{-K}^{K} \int_{-K}^{K} \sum_{i,j \in X_1} \sum V_{max} W_{(0,0)(i,j)} \mathrm{sinc}^2(x)\, \mathrm{sinc}^2(y) dxdy. \tag{5-7}$$

The calculation of the maximum noise power at the center detector is much more complicated because it consists of signal sidelobes, spurious diffraction orders, reconstructed spot tails, and non-paraxial region noise. The maximum noise power due to signal sidelobes collected at the center detector location, l_1^{SS}, is the integral of the intensity profile in Eq. (5-3) over the $P \times P$ region of the center detector,

$$l_1^{SS} = \int_{-\frac{P}{2}}^{\frac{P}{2}} \int_{-\frac{P}{2}}^{\frac{P}{2}} \sum_{i,j \in X_2} \sum V_{max} W_{(0,0)(i,j)}^{SS} \mathrm{sinc}^2\left(\frac{CTx_f}{\lambda f}\right) \mathrm{sinc}^2\left(\frac{CTy_f}{\lambda f}\right) dx_f dy_f, \tag{5-8}$$

and can be simplified to

$$l_1^{SS} = D \int_{-K}^{K} \int_{-K}^{K} \sum_{i,j \in X_2} \sum V_{max} W_{(0,0)(i,j)}^{SS} \mathrm{sinc}^2(x)\, \mathrm{sinc}^2(y)\, dxdy. \tag{5-9}$$

Similarly, the maximum noise power due to spurious diffraction orders collected at the center detector location, l_1^{SDO}, is the integral of the intensity profile in Eq. (5-4) over the $P \times P$ region of the center detector,

$$I_1^{SDO} = \int_{-\frac{P}{2}}^{\frac{P}{2}} \int_{-\frac{P}{2}}^{\frac{P}{2}} \sum_{i,j \in X_3} \sum V_{max} W_{(0,0)(i,j)}^{SDO} \operatorname{sinc}^2\left(\frac{CTx_f}{\lambda f}\right) \operatorname{sinc}^2\left(\frac{CTy_f}{\lambda f}\right) \, dx_f dy_f, \qquad (5\text{-}10)$$

and can be simplified to

$$I_1^{SDO} = D \int_{-K}^{K} \int_{-K}^{K} \sum_{i,j \in X_3} \sum V_{max} W_{(0,0)(i,j)}^{SDO} \operatorname{sinc}^2(x) \operatorname{sinc}^2(y) \, dxdy. \qquad (5\text{-}11)$$

At any off-center detector location, the intensity profile, $I_{(m,n)}(x_f, y_f)$ (m and n are the indices of the detector and $(m,n) \neq (0,0)$), is centered at the corresponding detector location and has similar form as Eq. (5-5) except that the values of interconnection weights (both desired ones and undesired ones) will be different. The reconstructed spot tail power from each off-center detector location is the integral of $I_{(m,n)}(x_f, y_f)$ over the $P \times P$ region of the center detector. The total reconstructed tail power, I_1^{tail}, collected by the center detector can be calculated by summing over all the off-detector locations as

$$I_1^{tail} = \sum_{\substack{m = \frac{-N}{2} \\ (m,n) \neq (0,0)}}^{\frac{N}{2}-1} \sum_{n = \frac{-N}{2}}^{\frac{N}{2}-1} \int_{-\frac{P}{2}}^{\frac{P}{2}} \int_{-\frac{P}{2}}^{\frac{P}{2}} I_{(m,n)}(x_f, y_f) \, dx_f dy_f, \qquad (5\text{-}12)$$

and can be simplified to

$$l_1^{tail} = D \sum_{\substack{m=\frac{-N}{2} \\ (m,n)\neq(0,0)}}^{\frac{N}{2}-1} \sum_{n=\frac{-N}{2}}^{\frac{N}{2}-1} \int_{-K}^{K} \int_{-K}^{K} I_{(m,n)}(x,y)\,dxdy \tag{5-13}$$

We calculate the total non-paraxial region noise l_1^{NRN}, as follows. Let the paraxial reconstruction region of each sub-DOE be of physical size $L \times L$, and the nonparaxial reconstruction region that includes the detector array region be called the nonparaxial region of interest. Also let the integrated light power in the paraxial reconstruction region of the sub-DOE, expressed as a percentage of the light power incident on the sub-DOE, be η_L. Then the percentage of light power in the nonparaxial reconstruction region of that sub-DOE is $(1-\eta_L)$, and we assume that this light power is uniformly distributed over the nonparaxial region of interest. Therefore, the center detector will collect some portion of the non-paraxial region noise from all the sub-DOEs in the non-paraxial reception region, X_4, and then l_1^{NRN} is the sum of all of these contributions. An accurate model for the non-paraxial region noise is unimportant for the crosstalk analysis since, as mentioned in Chapter 3, the light energy in this region is small.

The maximum noise power, l_1, is the sum of l_1^{SS}, l_1^{SDO}, l_1^{tail}, and l_1^{NRN}. in this thesis, we denote the four crosstalk components at the center detector as β_{SS}, due to signal sidelobes, β_{SDO}, due to spurious diffraction orders, β_{tail}, due to all reconstructed spot tail,

and β_{NRN}, due to non-local region noise. These four crosstalk components can be calculated as

$$\beta_{SS} = \frac{I_1^{SS}}{I_0}, \tag{5-14}$$

$$\beta_{SDO} = \frac{I_1^{SDO}}{I_0}, \tag{5-15}$$

$$\beta_{tail} = \frac{I_1^{tail}}{I_0}, \tag{5-16}$$

and

$$\beta_{NRN} = \frac{I_1^{NRN}}{I_0} \tag{5-17}$$

As mentioned in Chapter 4, the reconstructed spot tail crosstalk depends on the relative reconstructed spot spacing and relative detector width (with respect to the reconstructed spot width). The non-local region noise should decrease as the detector size decreases since a detector will collect less uniformly distributed non-local region noise. Taken together, these two crosstalk components can be greatly reduced either by using a larger

compression ratio (hence, a larger sub-DOE area and system size), or by decreasing the detector size. On the other hand, the strength of the signal sidelobes and spurious diffraction orders generated by a sub-DOE are only dependent on the design of the sub-DOE. These two higher order terms are the undesired interconnections. As can be seen in Eq. (5-2), (5-3), and (5-4), these two higher order terms have the same reconstructed spot sizes and spacings as the desired interconnections. Therefore, the crosstalk due to signal sidelobes and spurious diffraction orders for a given designed sub-DOE array can not be reduced by increasing the compression ratio or decreasing the detector size. Since $\beta = \beta_{SS} + \beta_{SDO} + \beta_{tail} + \beta_{NRN}$, increasing the compression ratio or decreasing the detector size can at best reduce the overall crosstalk to $\beta_{SS} + \beta_{SDO}$. Thus, $\beta_{SS} + \beta_{SDO}$ is the absolute lower bound of the overall interconnection crosstalk in the system.

To evaluate these crosstalk terms, a weighted interconnection with 128×128 nodes in both the input and output planes has been simulated. Nine different sub-DOE's (having 16 phase levels, 8×8 phase elements per period, and an average diffraction efficiency of 79%) were designed using the Gerchberg-Saxton algorithm, each of which connects an input node to the nearest 5×5 neighbors in the output plane with randomly chosen connection weights between zero and one. Each of the required 16,384 (= 128^2) sub-DOE's in the system were randomly selected from this set of nine sub-DOE designs. Also, for the system scale sizes of interest, we assume that the paraxial reconstruction region, $L \times L$, of each sub-DOE is of size equal to two DOE reconstruction periods in each dimension (as shown in Figure 5-3). Calculating from Eq. (5-14) to (5-17) for the designed interconnection system, Figure 5-4 shows the resulting crosstalk at the center

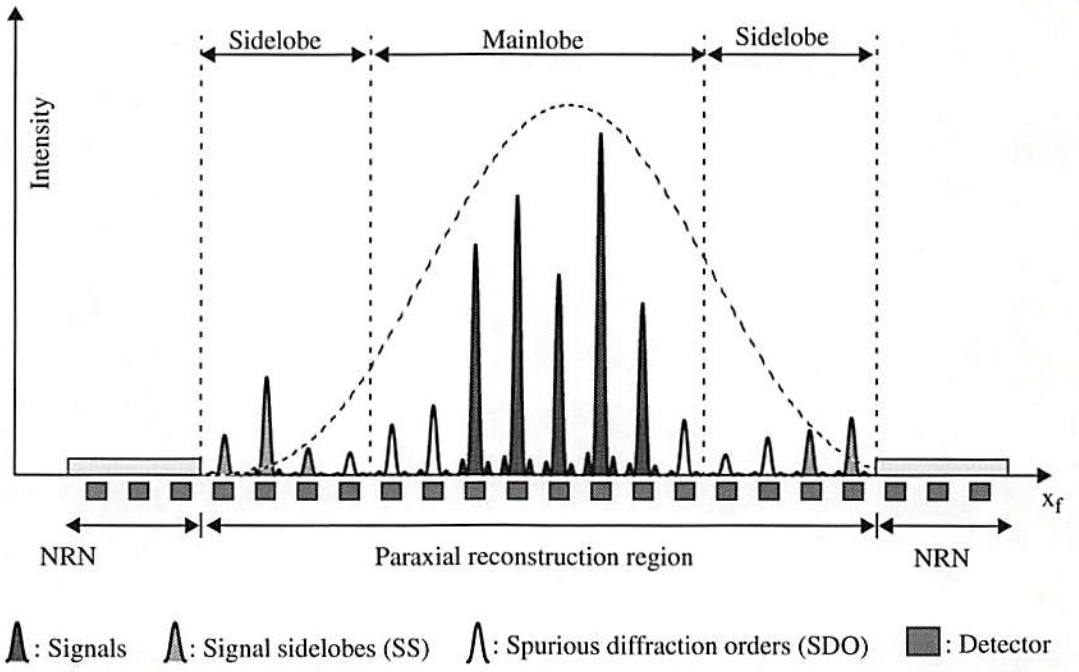

Figure 5-3: The reconstruction pattern form a DOE (with 8×8 phase elements in one period and assuming a paraxial reconstruction region of 2×2 sub-DOE period).

detector as a function of the C and K with the same constraint of $C \geq 2K + 1$ mentioned in Chapter 4. As predicted, the overall crosstalk is much higher than the crosstalk in a fully connected system for typical choices of parameters. In fact, β_{tail} of a limited-fanout system alone is comparable in magnitude (not shown in Figure 5-4) with the total crosstalk of a fully connected system because both of them are the crosstalk due to the reconstructed spot tails. The overall crosstalk still can be reduced by increasing the compression ratio (hence, a larger system size, since system lateral width W and propagation length Z are each proportional to C) or reducing the detector size (thereby reducing $\beta_{tail} + \beta_{NRN}$, as predicted above). For the interconnection system of this study, the crosstalk component due to signal sidelobes is about 0.11 and the crosstalk component

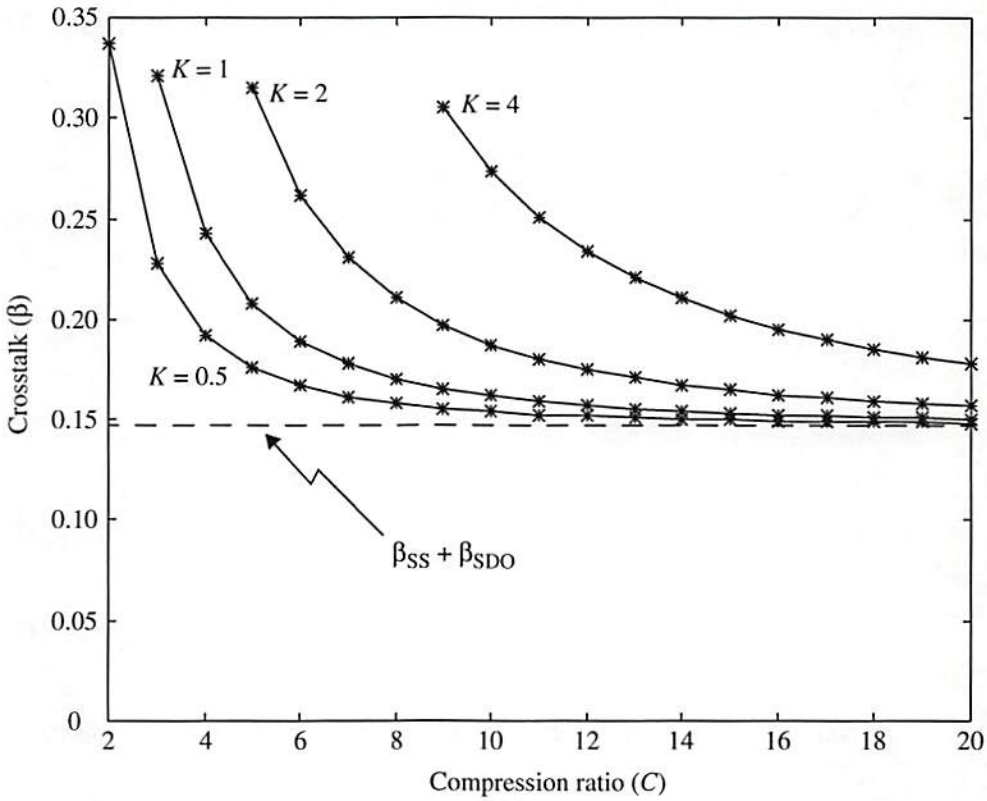

Figure 5-4: Crosstalk of the limited-fanout space-variant interconnection system as a function of C (compression ratio) for K (normalized detector width) = 0.5, 1, 2, 4. The sum of the major components, $\beta_{SS} + \beta_{SDO}$, is shown by the dash line.

due to spurious diffraction orders is about 0.04. Also shown in Figure 5-4 by the dashed line is the sum of these two crosstalk components, β_{SS} and β_{SDO}, which sets the absolute lower bound on interconnection crosstalk for the system. Given our definition of crosstalk, the resulting crosstalk level for the center detector is always greater than 0.15, which might be too high for some neural network applications. In the next section, we will develope a crosstalk reduction technique for the limited-fanout system.

5.3 Crosstalk Reduction for Limited-Fanout Space-Variant Systems by Modified DOE Designs

We now consider employing an alternative sub-DOE design to reduce the major crosstalk components β_{SS} and β_{SDO}. The high crosstalk problem in the system is due to the signal sidelobes and spurious diffraction orders from every sub-DOE reconstruction. These two components fall on, and not between, the detectors. Our approach to reduce the crosstalk is to rearrange the reconstruction in the sub-DOE design such that some of the signal sidelobes and spurious diffraction orders fall in off-detector locations in the output plane. We accomplish this by inserting $Y-1$ $(Y>1)$ spurious diffraction orders between every pair of signal orders in each dimension when designing the sub-DOE's. We assume Y is a prime integer in this thesis. The standard sub-DOE design in Section 5.2 corresponds to the case of $Y = 1$. In the next two sub-sections, we discuss the crosstalk reduction for spurious diffraction orders and signal sidelobes respectively.

5.3.1 Spurious Diffraction Order Crosstalk Reduction for the Limited-Fanout Space-Variant Interconnection System

For the crosstalk reduction parameter $Y>1$, the sub-DOE is designed using only every Y'th diffraction order in each dimension as a specified signal. The sub-DOEs are designed so that in the system these signals line up with the detectors. There are no detectors between these signal locations. The diffraction orders between each pair of adjacent signal orders are intended to be available as innocuous locations for noise energy, and therefore become additional spurious diffraction orders, which do not fall on detectors. If the sub-

DOE is designed appropriately, then most of the energy in spurious diffraction orders will fall on these off-detector locations.

Of course, outside of the signal diffraction order region, there will still be spurious diffraction orders that fall on detectors, as before. However, only 1 out of each set of Y^2 spurious diffraction orders will fall on the detectors. Here, we refer to the spurious diffraction orders that fall on any detector as the *detected spurious diffraction orders* and the spurious diffraction orders that will not fall on any detector as the *non-detected spurious diffraction orders*. In order to reduce the crosstalk substantially, the detected spurious diffraction orders need to be suppressed in the design process as much as possible. The DOE design algorithm has to be modified to incorporate this additional constraint. Figure 5-5 shows the modified Gerchberg-Saxton algorithm for suppressing the detected spurious diffraction orders when $Y > 1$. In this thesis, this modified algorithm is referred as the crosstalk-reduction Gerchberg-Saxton algorithm. A DOE illuminated by an input light beam will modulate only the phase profile of the input beam. After passing through the DOE, the light will diffract into many different diffraction orders. We refer to the plane immediately behind the DOE as the diffraction plane. The goal is to design the discrete phase distribution in one period of the DOE such that the intensity distributions of the diffracted orders will be a good approximation to the M^2 desired interconnection weights. Note that the design parameters are the number of phase elements in one period of the DOE in each dimension, J (= $BM\Delta$), and the number of discrete phase levels. As mentioned in Chapter 2, we choose J to be greater than M (hence, $B > 1$) to improve the performance of the DOE. The number of periods in the

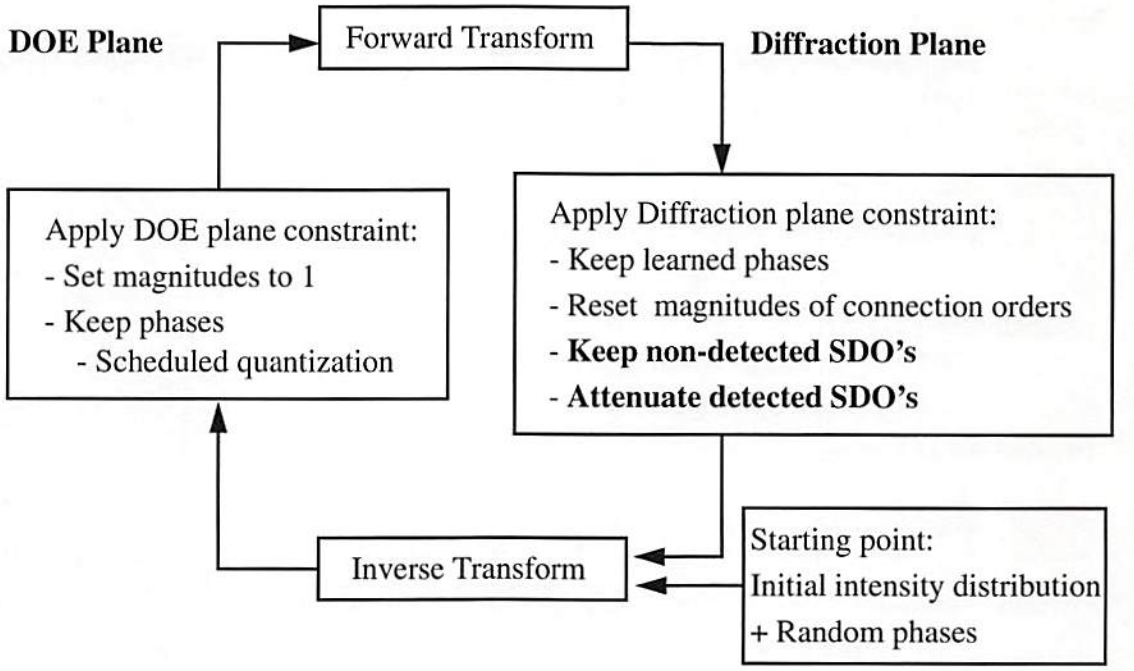

Figure 5-5: Schematic diagram of the crosstalk-reduction Gerchberg-Saxton algorithm.

DOE, size of the phase element, optical wavelength, and focal length of the lens will only affect the spot spacing and spot size after the lens focuses the diffracted light onto the reconstruction plane. Therefore, they are not the parameters in DOE design process. The DOE plane and the diffraction plane are related by Eq. (3-6). The complex amplitude of each diffraction order, $\underline{A}(r,s)$ (r and s are the indices of the diffraction orders with $r, s \in \{0, \pm 1, \pm 2, ...\}$), can be computed by

$$\underline{A}(r,s) = J^{-2}\,\mathrm{sinc}\left(\frac{r}{J}\right)\mathrm{sinc}\left(\frac{s}{J}\right)DFT\{\underline{h}(p,q)\}, \tag{5-18}$$

where $\underline{h}(p, q)$ (p and q are the indices of the phase elements with $p, q \in \{0, 1, ..., J-1\}$) is the complex amplitude transmittance distribution in one period of the DOE and DFT is the two-dimensional discrete Fourier transform. The intensity of each diffracted order is obtained by

$$I(r, s) = |\underline{A}(r, s)|^2. \tag{5-19}$$

As mentioned in Chapter 3, the number of independent diffraction orders is equal to the number of the phase elements in the DOE. Therefore, we only need to calculate the complex amplitude for $r, s \in \{\frac{-J}{2}, ..., \frac{J}{2}-1\}$ (assuming J is even). Similarly $\underline{h}(p, q)$ can be calculated by

$$\underline{h}(p, q) = DFT^{-1}\left\{\frac{J^2 \underline{A}(r, s)}{\operatorname{sinc}\left(\frac{r}{J}\right)\operatorname{sinc}\left(\frac{s}{J}\right)}\right\}, \tag{5-20}$$

Where DFT^{-1} is the inverse discrete Fourier transform. The design process begins with a set of random phase values, $\Phi(r, s)$, and the initial set of desired intensity distribution of diffraction orders, $I_0(r, s)$, defined as

$$I_0(r, s) = \begin{cases} W(u, v) & \text{for } (r, s) = (Yu, Yv) \\ 0 & \text{otherwise} \end{cases}, \tag{5-21}$$

where $W(u, v)$ represents the desired interconnection weights and (u, v) is the indices of the connection weight $(u, v \in \{\frac{-(M-1)}{2}, ..., \frac{(M-1)}{2}\}$, assuming M is odd). Note that the signals are now in the (Yu, Yv)'th diffraction orders of the DOE. The complex amplitude, $\underline{A}(r, s)$, of each diffraction order is synthesized as

$$\underline{A}(r, s) = \sqrt{I_0(r, s)}\, e^{j2\pi\Phi(r, s)}. \tag{5-22}$$

Then, the complex amplitude transmittance, $\underline{h}(p, q)$, of the DOE grating then is computed from Eq. (5-20) (labeled as inverse transform in Figure 5-5). In the DOE plane, the transmittance magnitude of each phase element is set to one (phase only) and the phase of each element is quantized to one of multiple levels by an appropriate quantization schedule. The purpose of the quantization schedule is to gradually quantize the phase of each element to one of the predefined phase levels for avoiding the stagnation problem [46]. Afterward, the complex amplitude, $\underline{A}(r, s)$, and intensity, $I(r, s)$, of each diffraction order are calculated from Eq. (5-18) and (5-19) (labeled as forward transform in Figure 5-5). In the diffraction plane, the learned phases of all diffraction orders from the iteration are saved for the next iteration. The intensities of the signals, $I(Yu, Yv)$, are reset to the corresponding connection weight, $W(u, v)$. The intensities of the non-detected spurious diffraction orders are saved since they are not on detectors. On the other hand, the intensities of the detected spurious diffraction orders are attenuated in order to reduce the spurious diffraction order crosstalk. This process then continues until it converges to an acceptable solution.

In this thesis, we want to compare the crosstalk levels for different values of Y with the same space-bandwidth product in each sub-DOE and the same system volume. For a given phase element size, the lateral width of each sub-DOE, S ($= CT = CBM\Delta$), and the lateral width of entire sub-DOE array, W ($= NS$), are therefore the same for different values of Y. In order to have the same volume for a given S and W, the propagation length, Z, has to stay constant. The propagation length (Z or f) due to the cascadability requirement over layers can be shown to be $\frac{S(BM\Delta)}{\lambda Y}$ (detailed in Section 5.4). Since Z is proportional to the ratio $\frac{B}{Y}$, the oversampling ratio (B) and the number of phase elements in one period of each sub-DOE in each dimension (J) must be increased in the same proportion as Y to hold constant the propagation length and system volume. Consequently, the corresponding compression ratios (C) needs to be decreased by a factor of Y to hold constant the space-bandwidth product of each sub-DOE. The crosstalk will be studied as a function of the compression ratio, C, and the normalized detector width, K (the ratio of the detector width, P, to the reconstructed spot width, $\frac{2\lambda f}{CT}$), for different Y values. The constraint on the values of C and K due to the allowable size of the detector can be derived as follows. Using the expressions for the propagation length (Z) and the lateral width of a sub-DOE (S), the reconstructed spot width, $\frac{2\lambda f}{CT}$, can be found as

$$S_w = \frac{2(BM\Delta)}{Y} = \frac{2T}{Y}, \tag{5-23}$$

where $T\,(=\,BM\Delta)$ is the period length of each sub-DOE in each dimension. The width of the detector, P, can then be expressed as $\frac{2K\,(BM\Delta)}{Y}$. Because the width of the detector (P) can not be greater than the width of the sub-DOE (S), C and K will be constrained by the relation $C \geq \frac{2K}{Y}$. Accordingly, a minimum C value that satisfies $C \geq \frac{2K}{Y} + 1$ was assumed.

Using the crosstalk-reduction Gerchberg-Saxton algorithm, the sub-DOEs were redesigned and additional sets of simulations (details described in Section 5.2) were performed for values of $Y = 2$. The number of the phase elements in one period of each sub-DOE is increased from 8×8 to 16×16 to hold constant the system volume for a given size of the sub-DOE. There is one spurious diffraction order in between every pair of signal orders in each dimension (Figure 5-3). Hence, only one out of every four

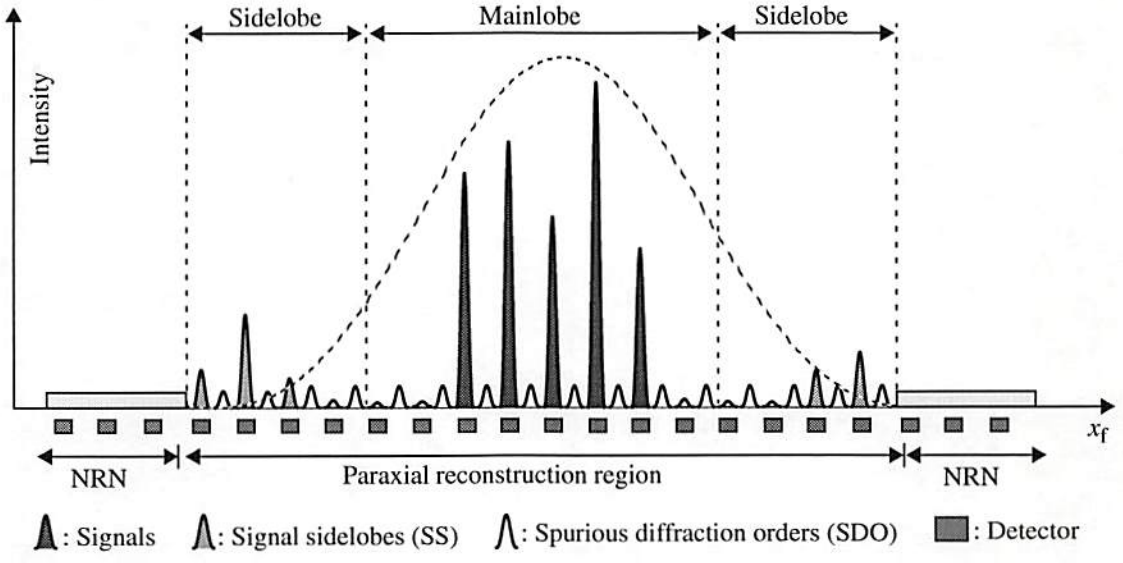

Figure 5-6: The reconstruction pattern form a DOE (with 16×16 phase elements in one period) for crosstalk reduction parameter (Y) = 2 case. Note that there is one spurious diffraction order between each pair of detectors but signal sidelobes still fall on detectors.

spurious diffraction orders will fall on the detectors. With the success of suppressing the detected spurious diffraction orders (at the expense of a stronger non-detected spurious diffraction orders), the spurious diffraction order crosstalk (β_{SDO}) should be reduced substantially. However, the signal sidelobe crosstalk (β_{SS}) will not be reduced since all of the signal sidelobes still fall on the detectors as shown in Figure 5-3. On the other hand, the tail crosstalk (β_{tail}) will increase slightly due to the tails of non-detected spurious diffraction orders between the detectors. Figure 5-7 shows the overall crosstalk as a function of the compression ratio, C, and normalized detector size, K. Similar to the crosstalk for the case of $Y = 1$, the crosstalk decreases as C increases or K decreases. However, the total crosstalk is lower than the crosstalk for the case of $Y = 1$ since one of the major crosstalk components has been reduced.

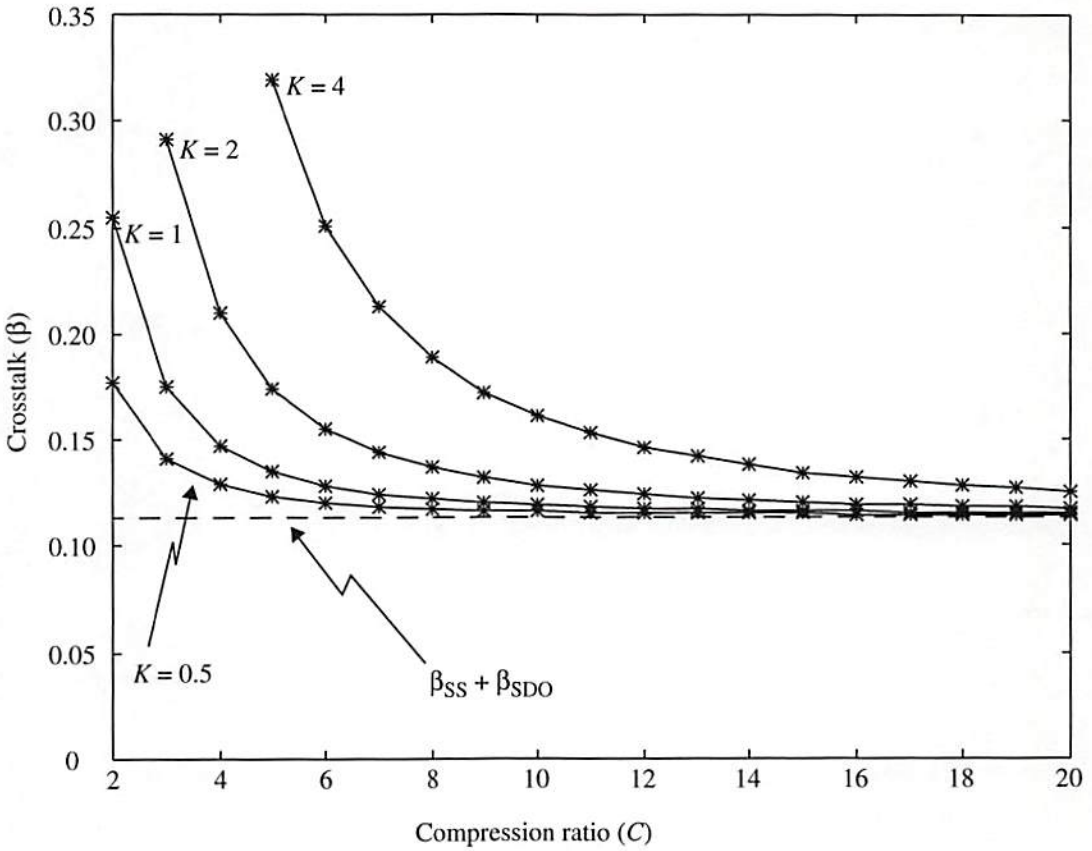

Figure 5-7: Crosstalk of the limited-fanout space-variant interconnection system for crosstalk reduction parameter (Y) = 2 as a function of C (compression ratio) and K (normalized detector width) = 0.5, 1, 2, 4. The sum of the major components, $\beta_{SS} + \beta_{SDO}$, is shown by the dash line.

5.3.2 Signal Sidelobe Crosstalk Reduction for the Limited-Fanout Space-Variant Interconnection System

Unlike the spurious diffraction orders, signal sidelobes can not be suppressed by DOE design process because they are sidelobes of the signals. This can be justified by finding the strength of the signal sidelobe relative to the connection order in the sub-DOE. As mentioned in Section 5.3.1, the connection orders are in the (Yu, Yv)'th diffraction orders of each sub-DOE for the case of $Y > 1$. For a given connection order (fixed values of Y, u, and v), the corresponding signal sidelobe in the (k_1, k_2)'th sidelobe is in the $(k_1 BM + Yu, k_2 BM + Yv)$'th diffraction orders of the sub-DOE. Using Eq. (3-16), the intensity ratio of the signal sidelobe order in the (k_1, k_2)'th sidelobe to the corresponding connection order can be found as

$$\frac{I(k_1 BM + Yu, k_2 BM + Yv)}{I(Yu, Yv)} = \left| \frac{\left(\frac{Yu}{BM}\right)}{k_1 + \left(\frac{Yu}{BM}\right)} \right|^2 \times \left| \frac{\left(\frac{Yv}{BM}\right)}{k_2 + \left(\frac{Yv}{BM}\right)} \right|^2 . \qquad (5\text{-}24)$$

Note that the relative strength of a given signal sidelobe diffraction order (u, v, k_1 and k_2 fixed) is dependent only on the ratio, $\frac{B}{Y}$, as can be seen from Eq. (5-24). The relative strength of the signal sidelobes can not be reduced by the design process and it can be reduced only by increasing the ratio, $\frac{B}{Y}$, of the sub-DOE. On the other hand, we want to fix the ratio, $\frac{B}{Y}$, to hold constant the propagation length and system volume for different Y

values (Section 5.3.1). Therefore, a different approach other than increasing the ratio, B/Y, for reducing signal sidelobe crosstalk is needed.

The goal is to move some or all of the signal sidelobes to off-detector locations. First, we exam that the spatial relation between signal sidelobes and the detector locations. It can be shown that the signal sidelobes of each sub-DOE will always fall on the detectors if the number of the phase elements in one period of each sub-DOE in each dimension, J ($= BM$), is divisible by the crosstalk reduction parameter Y. On the other hand, if J is not divisible by Y, then, within the paraxial reconstruction region of each sub-DOE, only one out of each set of Y^2 signal sidelobes will be detected and the remaining signal sidelobes become non-detected. However, since we have to increase J by a factor of Y to hold constant the propagation length and system volume, J will be always divisible by Y. Therefore, we will only increase J by a factor that is slightly different from Y to make it not divisible by Y. The resultant propagation length and system volume will only be approximately the same as those for $Y = 1$. We have analyzed this crosstalk reduction approach for the case of $Y = 3$. The sub-DOEs were redesigned and additional sets of simulations (details described in Section 5.2) were performed. For the reason stated above, we have used 22×22 (rather than 24×24) phase elements for one period of each sub-DOE. There are two spurious diffraction orders in between every pair of signal orders in each dimension. Hence, only one out of every nine spurious diffraction orders will fall on the detectors. We use the crosstalk-reduction algorithm to suppress the detected spurious diffraction orders in the sub-DOEs design process to further reduce the spurious diffraction order crosstalk. On the other hand, because we have chosen the paraxial

reconstruction region, $L \times L$, of each sub-DOE to be of size equal to two DOE reconstruction periods in each dimension for the system size of interest, there is no signal sidelobe falling on the detector for the case of $Y = 3$ (Figure 5-8). Therefore, the signal

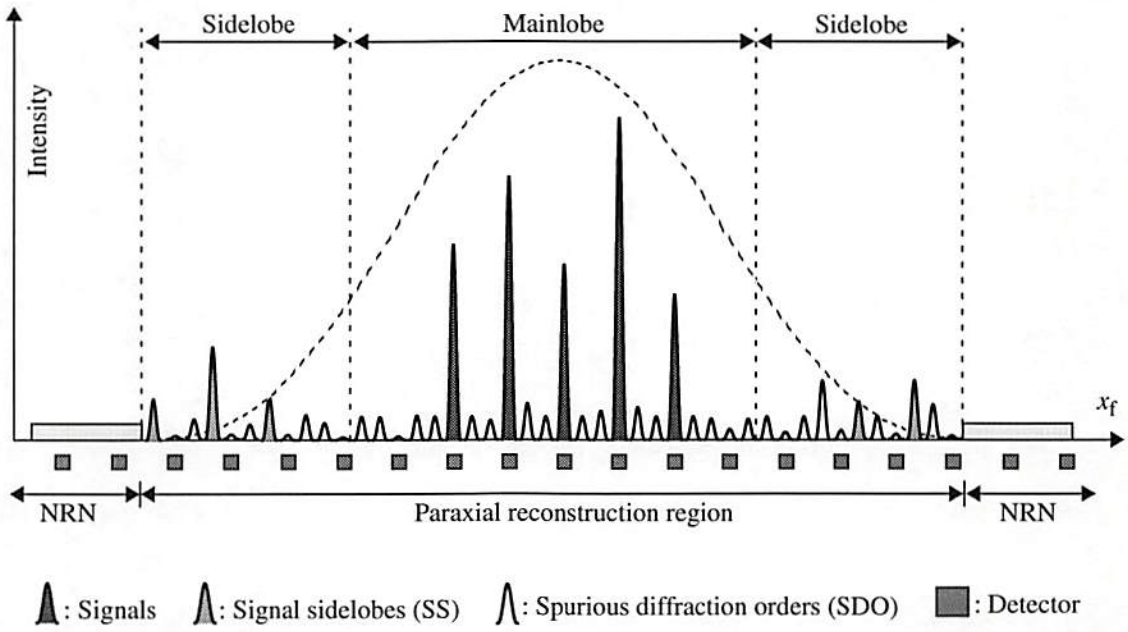

Figure 5-8: The reconstruction pattern form a DOE (with 22×22 phase elements in one period) for crosstalk reduction parameter (Y) = 3 case. Note that there are two spurious diffraction orders between each pair of detectors and signal sidelobes do not fall on detectors.

sidelobes within the paraxial reconstruction region of each sub-DOE become non-detected and the signal sidelobe crosstalk (β_{SS}) is entirely eliminated. In Figure 5-9 the simulated major crosstalk components, β_{SS} and β_{SDO}, and their sum are shown as a function of the crosstalk reduction parameter, Y. As can be seen in Figure 5-9, the sum of the major crosstalk components ($\beta_{SS} + \beta_{SDO}$) is reduced by a factor of 27 in going from $Y = 1$ to $Y = 3$. This significant reduction is due to the use of non-detected spurious diffraction

orders and non-detected signal sidelobes as well as the suppression of the detected spurious diffraction orders. Figure 5-10 shows the total crosstalk for the case of $Y = 3$ as

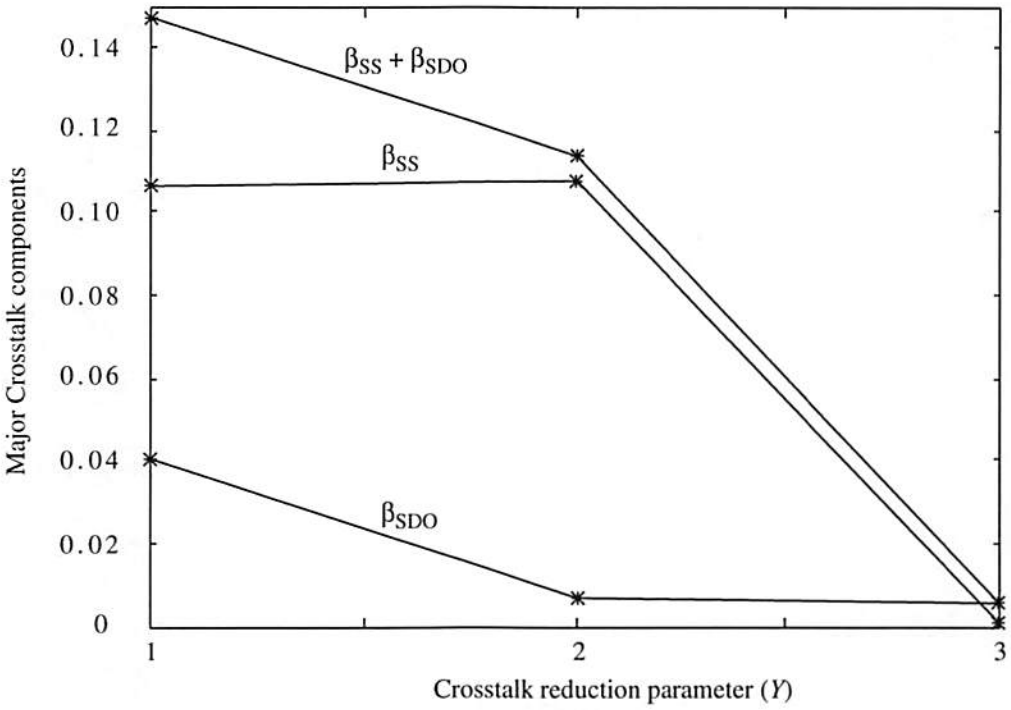

Figure 5-9: The major crosstalk components ($\beta_{SS} + \beta_{SDO}$, β_{SS}, and β_{SDO}) of the limited-fanout space-variant interconnection system as a function of crosstalk reduction parameter (Y).

a function of the compression ratio, C, and normalized detector size, K. The total crosstalk still decreases as C increases or K decreases. These results show that β_{SS} and β_{SDO} have been reduced to the point where they are no longer the dominant components of the total crosstalk. The other crosstalk components, β_{tail} and β_{NRN}, have increased somewhat (compared with the case of $Y = 1$ in Figure 5-4), because of a lower average sub-DOE diffraction efficiency (79% for $Y = 1$; 58% for $Y = 2$; 52% for $Y = 3$) and

the tails of non-detected spurious diffraction orders. Even so, the total crosstalk is significantly reduced for most parameter values of interest.

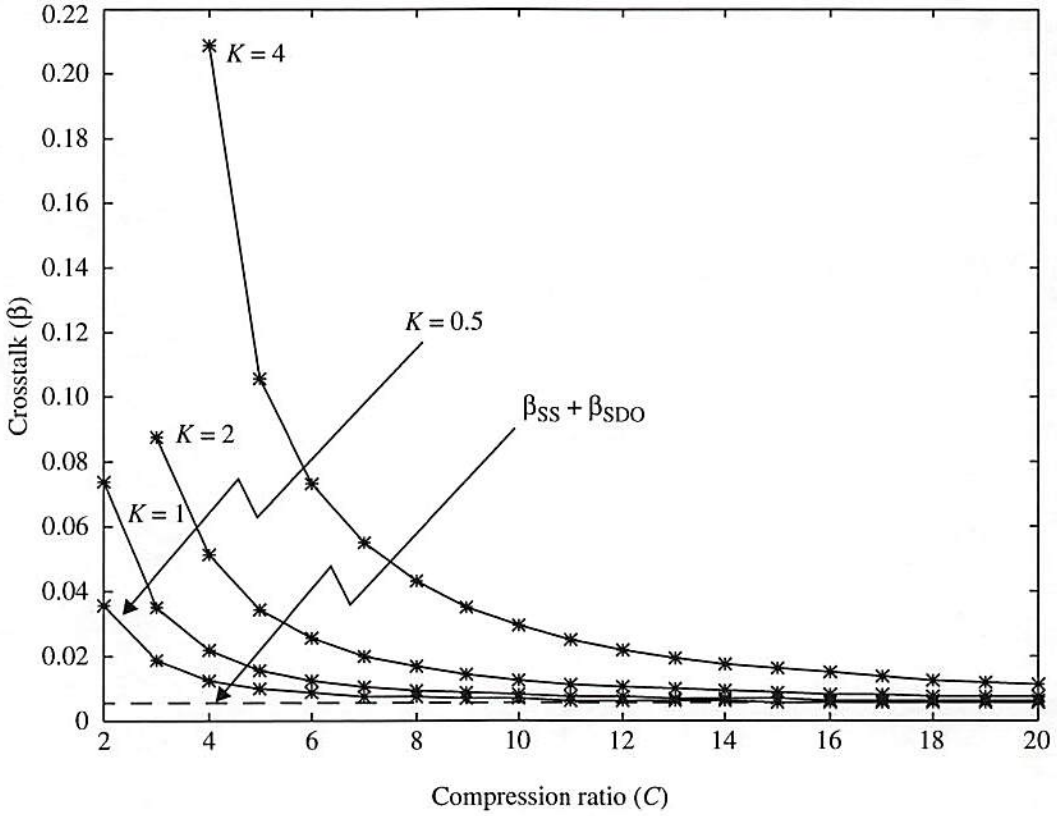

Figure 5-10: Crosstalk of the limited-fanout space-variant interconnection system for crosstalk reduction parameter (Y) = 3 (with 22×22 phase elements in one period) as a function of C (compression ratio) and K (normalized detector width) = 0.5, 1, 2, 4. The sum of the major components, $\beta_{SS} + \beta_{SDO}$, is shown by the dash line.

We have shown that, by using non-detected spurious diffraction orders and a crosstalk-reduction design algorithm, spurious diffraction order crosstalk can be significantly reduced. In addition, signal sidelobe crosstalk can be reduced by choosing a proper value of J that is not divisible by the crosstalk reduction parameter used. Although

the simulations are preformed for systems with a small paraxial reconstruction region for each sub-DOE, the crosstalk reduction technique should directly applicable to systems with a larger paraxial reconstruction region for each sub-DOE.

5.4 Advantages and Limitations of the Crosstalk Reduction Technique

The idea of inserting spurious diffraction orders in between signal diffraction orders can be usefully extended to any prime integer $Y > 3$ with a proper choice of J. If there are signal sidelobes within the paraxial reconstruction region of each sub-DOE, there is only one out of each set of Y^2 signal sidelobe orders falling on the detectors. Hence, a larger Y is preferred to reduce the signal sidelobe crosstalk. The reduction of spurious diffraction order crosstalk, β_{SDO}, will depend on how effectively the DOE design algorithms suppress the detected spurious diffraction orders. Our design programs tended to reduce β_{SDO} as Y increased, showing additional preference for larger Y. However, as mentioned, the compression ratio (C) becomes smaller as Y gets larger for a given sub-DOE size and system volume. Consequently, the spacing between adjacent reconstructed spots becomes smaller as Y increases. At some value of Y, the reconstructed spots will begin to overlap, and crosstalk performance will degrade. This will constrain the maximum allowable value of Y for a given set of physical dimensions. For neuron array devices that have smarter pixels in the output plane, the increase in device area for electronics corresponds to an increase in the detector-to-detector spacing (Figure 5-1), so that larger values of Y can be accommodated.

There is another advantage for using $Y > 1$, which is the potential to reduce the propagation distance (and hence the system volume). Because the connection orders are

now in the (Yu, Yv)'th diffraction orders of each sub-DOE, the diffraction angle of each connection order relative to the optical axis of the sub-DOE becomes

$$(\theta_{Yu}, \theta_{Yv}) = \left(\tan^{-1}\left(\frac{Yu\lambda}{T}\right), \tan^{-1}\left(\frac{Yv\lambda}{T}\right)\right). \tag{5-25}$$

Therefore, within the paraxial region, the diffraction angle of a given connection order is proportional to Y for a given T and λ, so $Y > 1$ implies a shorter propagation length between layers. The propagation length (Z or f) can be calculated by forcing the reconstructed spot spacing between the connection orders (which now becomes $\frac{Y\lambda f}{T}$) equal to the detector pitch, which is the same as the width of a node or sub-DOE (S), i.e.,

$$\frac{Y\lambda f}{T} = S. \tag{5-26}$$

Substituting $T = BM\Delta$, the propagation length can be expressed as

$$Z = \frac{S(BM\Delta)}{\lambda Y}. \tag{5-27}$$

Because the propagation length is proportional to $\frac{B}{Y}$, in all the simulations discussed in Section 5.3, the oversampling ratio (B) was increased approximately in the same proportion as Y to hold constant the propagation length and study the effect of Y on

crosstalk. If we now increase the oversampling ratio by a proportion less than Y (*i.e.*, we increase B and Y such that the ratio B/Y decreases), then the propagation length will be reduced. Therefore, we will be able to reduce the propagation length and interconnection crosstalk at the same time. However, the crosstalk reduction will not be as significant as the reduction found in Section 5.3, since one period of the sub-DOE now contains fewer phase elements and it becomes more difficult for the design algorithm to suppress the detected spurious diffraction orders. This behavior has been verified by an additional simulation for the case of $Y = 3$ and a sub-DOE period of 16×16 phase elements. As expected, the overall crosstalk (Figure 5-11) is higher than the crosstalk for the case of $Y = 3$ and a sub-DOE period of 22×22 phase elements (as in Section 5.3), but the propagation length will be reduced by a factor of $\frac{3}{2}$. Taking the crosstalk reduction parameter (Y) into consideration, Table 5-3 shows the characteristics of limited-fanout systems with a given set of independent parameters (N, λ, Δ, S, B, M, Y), to explicitly show the dependence on Y. It can be seen that a shorter propagation length, smaller volume, and higher connection density can be obtained if we increase the oversampling ratio by a proportion less than Y.

Table 5-3: Limited-fanout space-variant systems with crosstalk reduction parameter, Y.

SBWP of sub-DOE array	W	Propagation length (Z)	Aspect ratio (W/Z)	f-number of lens	NI	System volume (ZW2)	Connection density (NI/ZW2)
$\frac{N^2S^2}{\Delta^2}$	NS	$\frac{S(BM\Delta)}{\lambda Y}$	$\left(\frac{Y}{B}\right)\left(\frac{\lambda}{\Delta}\right)\left(\frac{N}{M}\right)$	$M\left(\frac{B}{Y}\right)\left(\frac{\Delta}{\lambda}\right)$	M^2N^2	$\frac{S^3(BM\Delta)N^2}{\lambda Y}$	$\frac{\lambda MY}{S^3(B\Delta)}$

As an example, we have calculated the characteristics of a limited-fanout system in the first three rows of Table 5-4 with independent parameters $N = 128$, $M = 5$,

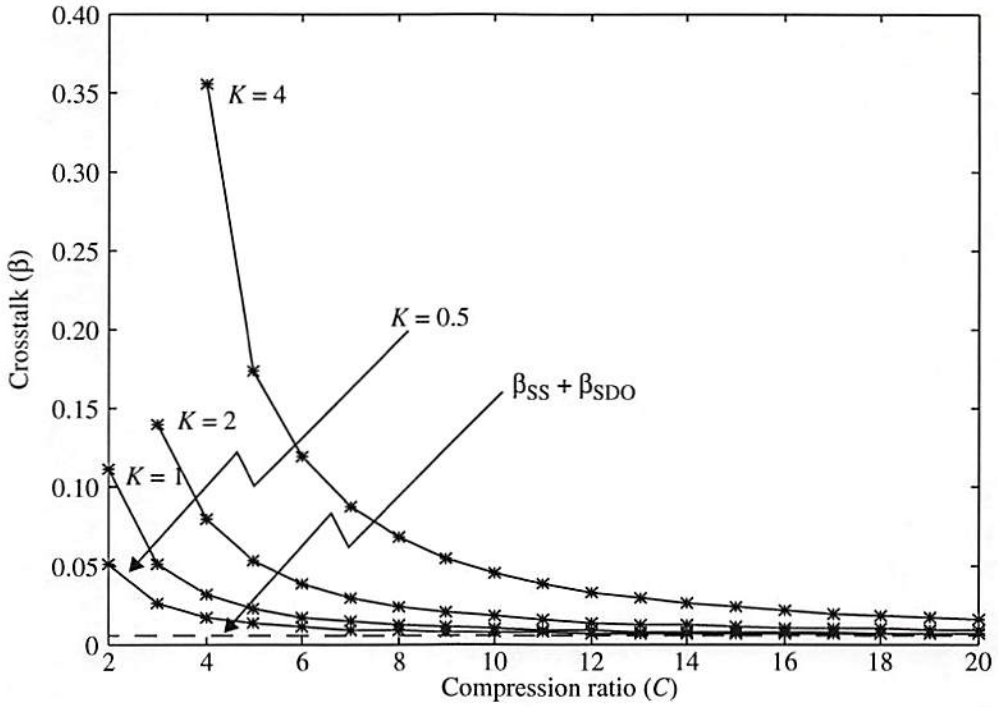

Figure 5-11: Crosstalk of the limited-fanout space-variant interconnection system (with 16×16 phase elements in one period) for crosstalk reduction parameter (Y) = 3 as a function of C (compression ratio) and K (normalized detector width) = 0.5, 1, 2, 4 (from bottom to top). The sum of the major components, $\beta_{SS} + \beta_{SDO}$, is shown by the dash line.

$\lambda = 850\text{nm}$, $\Delta = 2\mu\text{m}$, $S = 96\mu\text{m}$, and $P = 32\mu\text{m}$ for different values of Y. Note that, as mentioned in detail in Section 5.3.1 and 5.3.2, we want to compare the crosstalk levels for different values of Y while holding constant the space-bandwidth product in each sub-DOE, and holding approximately constant the propagation length (Z or f). Therefore, we have to increase the oversampling ratio (B) approximately by a factor of Y since the propagation length Z is proportional to the ratio $\frac{B}{Y}$. Consequently, the corresponding compression ratios (C) are decreased approximately by a factor of Y to hold constant the

space-bandwidth product of each sub-DOE. In Table 5-4, the compression ratio, C ($= \frac{S}{BM\Delta}$), and normalized detector width, K ($= \frac{P}{(2BM\Delta)/Y}$), are calculated from the values of the independent parameters for different Y. The interconnection crosstalk levels of the systems are then obtained from Figure 5-4, 5-7, 5-10 and 5-11 for the calculated values of C and K in each case. In the first three rows of Table 5-4, it can be seen that the overall crosstalk has been reduced by a factor of 2.6 in going from $Y = 1$ to $Y = 3$.

Table 5-4: An example for limited-fanout S-V systems, showing different values of the oversampling ratio (B) and the crosstalk reduction parameter (Y) for $N = 128$, $M = 5$, $\lambda = 850$ nm, $\Delta = 2$ μm, $S = 96$ μm, and $P = 32$ μm.

No. of phase elements in one period of sub-DOE	Detector width (P) (μm)	B	Y	Width of sub-DOE array (W) (mm)	Propagation length (Z) (mm)	C	Normalized detector width (K)	Crosstalk (β)
8 × 8	32	1.6	1	12.3	1.81	6.0	1.0	0.1889
16 × 16	32	3.2	2	12.3	1.81	3.0	1.0	0.1746
22 × 22	32	4.4	3	12.3	1.67	2.2	1.1	0.0736
22 × 22	16	4.4	3	12.3	1.67	2.2	0.55	0.0316

In going from $Y = 1$ to $Y = 3$ in Table 5-4, although the two major crosstalk components, β_{SS} and β_{SDO}, were greatly reduced (as shown in Figure 5-9), the total crosstalk β was reduced only moderately. This is because the other two crosstalk components (reconstructed tails and non-local region noise) can become dominant for sufficiently low values of compression ratio C (*cf.* Figure 5-4, 5-7, 5-10 and 5-11). As can be seen from these figures, the interconnection crosstalk due to reconstructed tails and non-local region noise can be reduced significantly by increasing the compression ratio or decreasing the normalized detector width K. Since the total crosstalk is the sum of all the

crosstalk components, a value of $Y > 1$ together with either a larger value of C or a smaller value of K can be used to greatly reduce the total crosstalk. For example, further crosstalk reduction can be achieved (compared with the third row of Table 5-4) by reducing detector width from 32μm to 16μm (thereby reducing the normalized detector width, K, by a factor of two) as shown in the fourth row of Table 5-4; the tradeoff is a reduction in the detected light energy. This reduces the total crosstalk by a factor of 6 when compared with Y=1 (first row of Table 5-4). If instead we decrease the crosstalk by increasing the compression ratio C, the space-bandwidth product (or size) of each sub-DOE (or node) must become larger for a fixed sub-DOE period length (fixed B, M, and Δ). Table 5-5 shows the resulting characteristics for systems with the same set of independent parameter values as in the first three rows of Table 5-4 except a larger sub-DOE width of 192μm. A crosstalk reduction factor of 7.3 has been achieved in going from $Y = 1$ to $Y = 3$ because the crosstalk components due to reconstructed tails and non-local region noise are greatly suppressed.

Table 5-5: Same example as in the first three rows of Table 5-4 for limited-fanout S-V systems except for a larger sub-DOE width (S = 192 μm). Different values of the oversampling ratio (B) and the crosstalk reduction parameter (Y) are shown for N = 128, M = 5, λ = 850 nm, Δ = 2 μm, and P = 32 μm.

No. of phase elements in one period of sub-DOE	Detector width (P) (μm)	B	Y	Width of sub-DOE array (W) (mm)	Propagation length (Z) (mm)	C	Normalized detector width (K)	Crosstalk (β)
8 × 8	32	1.6	1	24.6	3.6	12.0	1.0	0.1572
16 × 16	32	3.2	2	24.6	3.6	6.0	1.0	0.1287
22 × 22	32	4.4	3	24.6	3.3	4.4	1.1	0.0214

If minimization of the overall system size is important, the period length of the sub-DOE can be reduced while keeping the space-bandwidth product of the sub-DOE constant. This increases the diffraction angles, resulting in a decrease in propagation length. For example, Table 5-6 is based on the same values of independent parameters as the third row of Table 5-4, except the oversampling ratio B (and therefore the sub-DOE period length) is smaller in Table 5-6. As can be seen in this example, the tradeoff is a slight increase in total crosstalk. It is also possible to reduce the crosstalk and propagation length at the same time, and this can be seen by comparing an 8×8 sub-DOE period length using $Y = 1$ (first row of Table 5-4) with a 16×16 sub-DOE period length using $Y = 3$ (Table 5-6).

Table 5-6: An example for limited-fanout S-V systems that will reduce crosstalk and propagation length at the same time for $N = 128$, $M = 5$, $\lambda = 850$ nm, $\Delta = 2$ μm, $S = 96$ μm, and $Y = 3$.

No. of phase elements in one period of sub-DOE	Detector width (P) (μm)	B	Y	Width of sub-DOE array (W) (mm)	Propagation length (Z) (mm)	C	Normalized detector width (K)	Crosstalk (β)
16×16	32	3.2	3	12.3	1.20	3.0	1.5	0.0933

The maximum crosstalk that can be tolerated in an actual system will vary depending on the neural network model and on the application. For applications that have tight crosstalk requirements, a high crosstalk reduction parameter together with either a large sub-DOE size or a small detector width can be used.

By using the crosstalk reduction technique, it is possible to obtain a much lower crosstalk level for limited-fanout systems than is possible with the standard design ($Y = 1$). Because the crosstalk can be reduced to a fairly low level (e.g., $\beta = 0.0214$ in

the third row of Table 5-5), the system will have a much higher accuracy. The fully connected space-variant system can also take the advantage of a shorter propagation length by using $Y > 1$, although the crosstalk will increase slightly due to the non-detected spurious diffraction orders in between detectors.

This crosstalk reduction method should reduce crosstalk for physically small systems with the propagation length on the order of millimeter (results shown above) as well as for physically larger systems. An additional set of simulations have been performed for a physically larger system by assuming that the propagation length is long enough such that the paraxial reconstruction region will be larger than the entire optoelectronic chip area. Under this assumption, there are three crosstalk components, i. e., β_{SS}, β_{SDO}, and β_{tail}. Figure 5-12 shows the interconnection crosstalk as a function of C and K. Figure 5-13, and 5-13 show that the crosstalk has been significantly reduced by using the crosstalk reduction technique.

Although the designs in this chapter were based on the Gerchberg-Saxton algorithm. However, because the principle of the crosstalk reduction technique is based on rearrangement of noise orders and suppression of the remaining detected noise orders, this technique should also be applicable to other design algorithms.

In conclusion, the characteristics of the DOE-based fully connected and limited-fanout S-V interconnection architectures for fixed-connection multilayer feed-forward neural networks were analyzed and compared in terms of the propagation length, system volume, connection density, and interconnection crosstalk. A crosstalk reduction method with the potential to reduce system size for the limited-fanout S-V architecture was

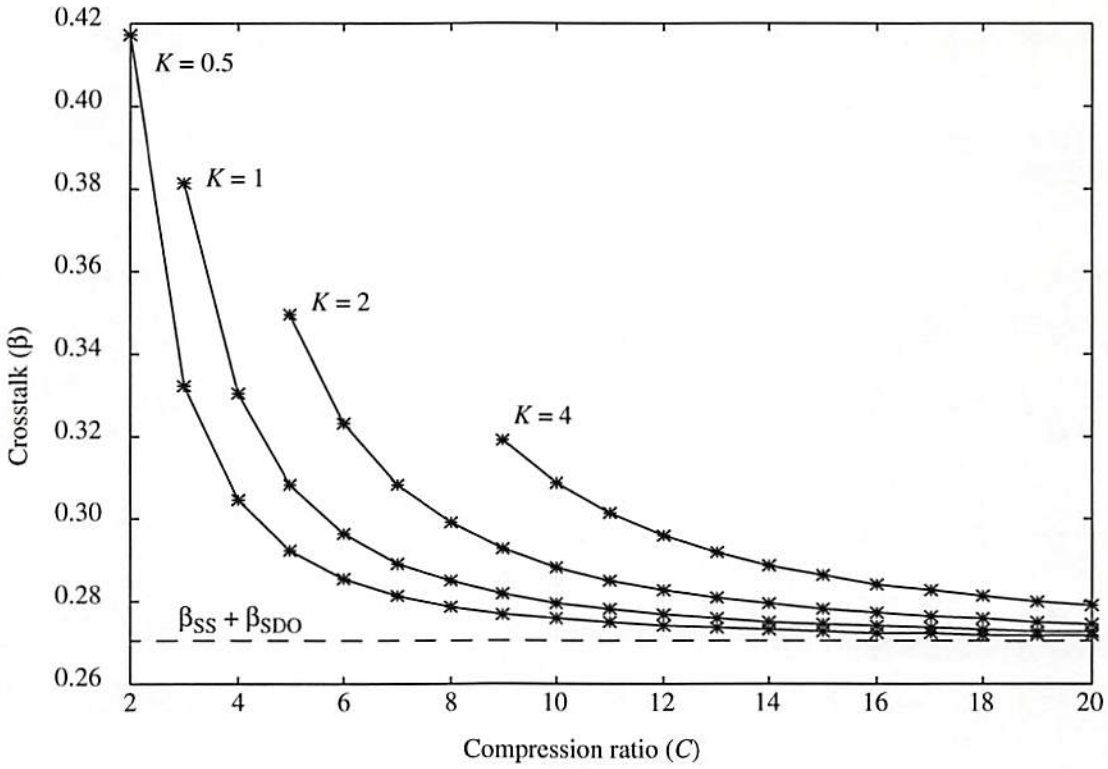

Figure 5-12: Crosstalk of the physically large limited-fanout space-variant interconnection system as a function of C (compression ratio) for K (normalized detector width) = 0.5, 1, 2, 4. The sum of the major components, $\beta_{SS} + \beta_{SDO}$, is shown by the dashed line.

described and a DOE design algorithm that incorporates this method was developed. Its validity was verified by simulating one layer of an interconnection system with 128×128 input nodes, 128×128 output nodes, and 5×5 nearest neighbor weighted interconnections from each input node to the output nodes. This method can reduce crosstalk for physically small systems with the propagation length on the order of millimeters (results shown) as well as for physically larger systems. Similar analyses should be applicable to other DOE-based analog (or digital) space-variant or space-invariant interconnection systems. Thus, by designing the sub-DOEs from the perspective

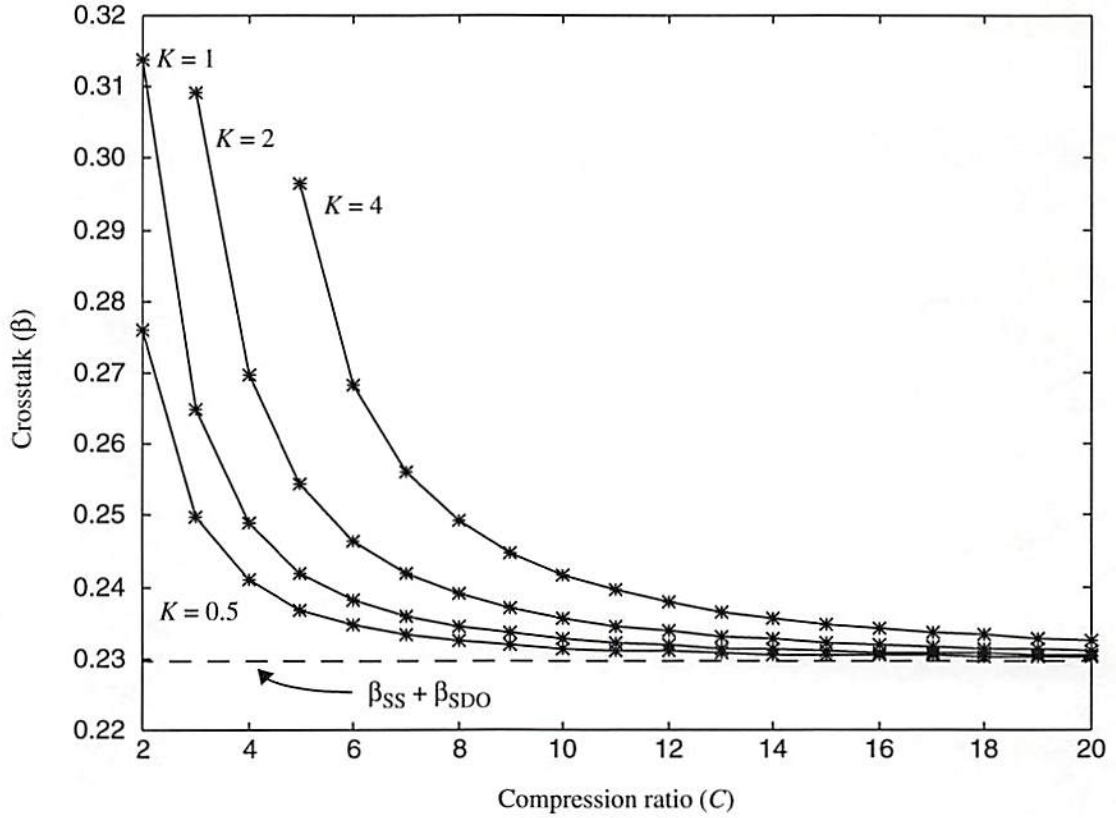

Figure 5-13: Crosstalk of the physically large limited-fanout space-variant interconnection system for crosstalk reduction parameter (Y) = 2 as a function of C (compression ratio) and K (normalized detector width) = 0.5, 1, 2, 4. The sum of the major components, β_{SS} + β_{SDO}, is shown by the dashed line.

of the interconnection system rather than from the perspective of each diffractive element, the interconnection system can achieve a higher accuracy and lower crosstalk level.

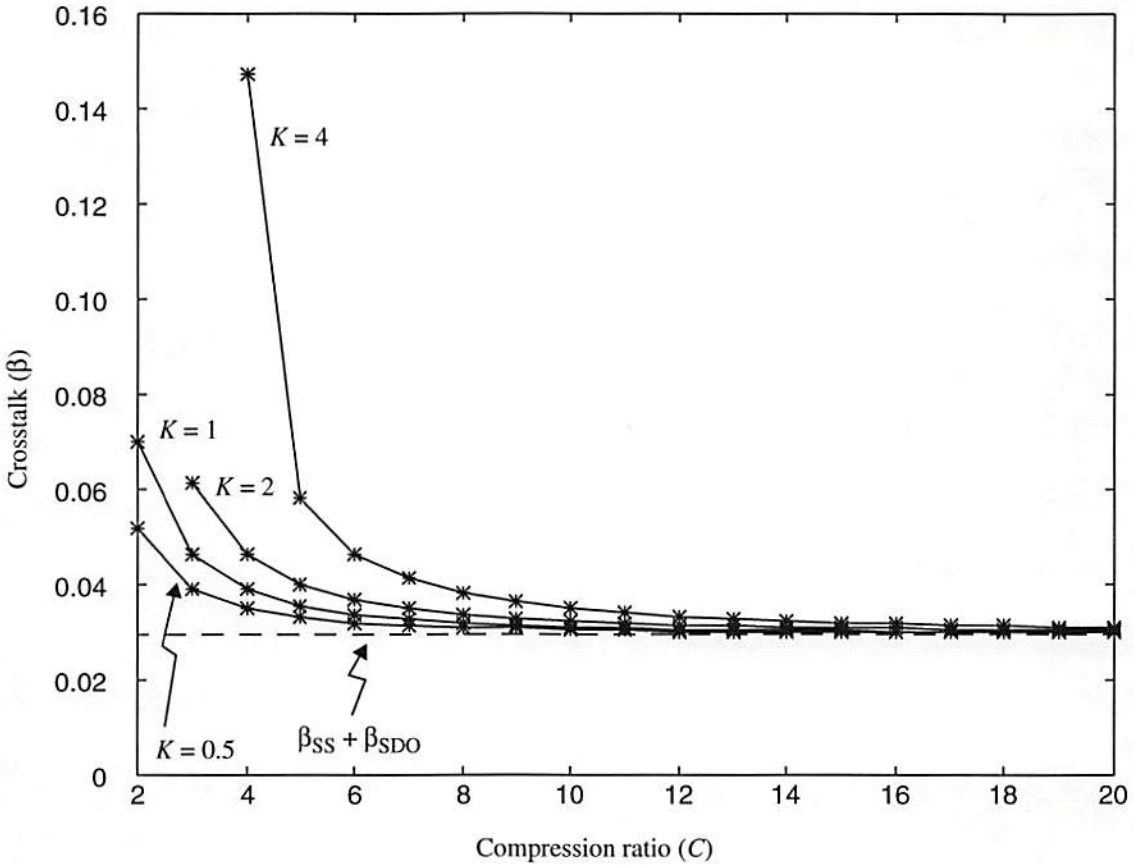

Figure 5-14: Crosstalk of the physically large limited-fanout space-variant interconnection system for crosstalk reduction parameter (Y) = 3 as a function of C (compression ratio) and K (normalized detector width) = 0.5, 1, 2, 4. The sum of the major components, $\beta_{SS} + \beta_{SDO}$, is shown by the dashed line.

Chapter 6

Reconstruction of Diffractive Optical Elements at Wavelengths Different from the Designed Wavelength

A DOE is usually designed and fabricated for one specific wavelength only. However, there are times that the illumination wavelength is not the same as the designed wavelength. For example, we might want to do initial testing on a DOE that is designed and fabricated for infrared wavelength using visible wavelengths for convenience. Also, when DOEs are in use in a system, the actual illumination wavelength may vary during operation of the system due to temperature or other effects. Therefore, it is important to understand how the reconstruction of a DOE will be affected when the illumination wavelength is different from the designed wavelength.

In this chapter, we will study the effect on the reconstruction of a DOE when the illumination wavelength is different from the designed optical wavelength. In Section 6.1, we will derive the relation between DOE reconstruction and illumination wavelength. We will see that, in general, the deviation on the reconstructed diffraction order is dependent on the phase distribution of a DOE. Also, this deviation varies from one diffraction order to another. Simulations on the effect of the wavelength change are performed for DOEs designed for a 3×3 spot array generator and a 3×3 triangular connection pattern.

From these simulations, we have found that binary-phase-level DOEs behave differently from multiple-phase-level DOEs when the illumination wavelength changes. The simulations show that binary-phase-level DOEs still generate symmetric reconstruction patterns even when the illumination wavelength changes. Furthermore, the percentage deviation of any of the off-axis reconstructed diffraction orders for a binary-phase-level DOE is the same for all the off-axis diffraction orders. In Section 6.2, we will develop a theoretical proof of this behavior of binary-phase-level DOEs. From the theoretical derivation, we have found that the percentage deviation of any of the off-axis reconstructed diffraction orders for a binary-phase-level DOE is only dependent on the ratio of the designed wavelength and the illumination wavelength. We have also considered the effect of wavelength change on DOE reconstruction for DOEs with phase delays other than 0 and π. The effect of wavelength change for those DOEs is somewhat different from the effect of wavelength change for the conventional DOEs.

6.1 The Relation between DOE Reconstruction and Illumination Wavelength

For a Z-phase-level DOE, the phase level index, $PI(p, q)$ (p and q are the indices of the phase elements with $p, q \in \{0, 1, ..., J-1\}$ and J is the number of the phase elements in one period of the DOE in each dimension), of a phase element can have any integer value from 0 to (Z-1). Each phase level index corresponds to one of the Z possible phase delays.

The phase delay of each phase element, $PD(p, q)$, is associated with the phase level index by

$$PD(p, q) = \frac{2\pi}{Z} \cdot PI(p, q). \tag{6-1}$$

Let's assume that the DOE is designed for a particular optical wavelength, λ_0. Then, in order to have the correct phase delays, each of the phase elements must be etched correctly. First, let's derive the correct etch depth, $d(p, q)$, for each of the phase elements. Shown in Figure 6-1 is a typical phase profile of a multiple-phase-level DOE. Let D represent the largest etch depth and n represent the index of refraction of the DOE material, then the phase delay of a phase element relative to the phase delay of the phase element with zero phase index can be calculated as

$$\begin{aligned} PD(p, q) &= \frac{2\pi}{\lambda_0} \cdot [D - d(p, q)] \cdot n + \frac{2\pi}{\lambda_0} \cdot d(p, q) - \frac{2\pi}{\lambda_0} \cdot D \\ &= \frac{2\pi}{\lambda_0} \cdot [D - d(p, q)] \cdot (n - 1) \end{aligned} . \tag{6-2}$$

Note that we have assumed that the index of refraction is 1 outside the DOE in Eq. (6-1). We can find the expression for the largest etch depth by using the fact that the phase index is (Z-1) and the phase delay is $\frac{2\pi}{Z} \cdot (Z - 1)$ when there is no etch needed ($d(p, q) = 0$). From Eq. (6-1) and Eq. (6-2), the largest etch depth can be shown as follows

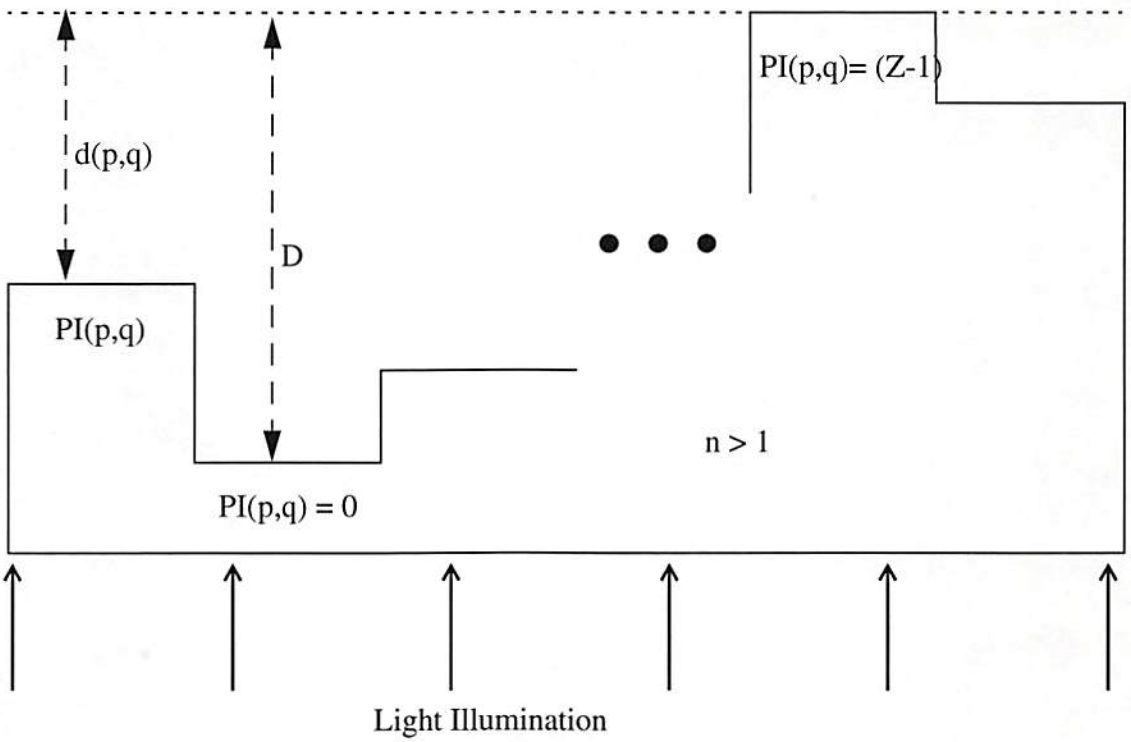

Figure 6-1: Discrete phase profile of a multiple-phase-level DOE. D is the largest etch depth and n is the index of refraction.

$$D = \frac{\lambda_0 (Z-1)}{Z(n-1)}. \tag{6-3}$$

Combining Eq. (6-2) and Eq. (6-3), we can immediately obtain the expression for the correct etch depth as

$$d(p,q) = \frac{\lambda_0}{Z(n-1)} \cdot [\,(Z-1) - PI(p,q)]\,. \tag{6-4}$$

Table 6-1 shows the relations of the phase indices, relative phase delays, and the correct etch depth for the designed wavelength, λ_0, for a multiple-phase-level DOE. Note that,

Table 6-1: The relation of the phase indices, relative phase delays, and the correct etch depth.

$PI(p, q)$	0	1	...	$(Z-2)$	$(Z-1)$
$PD(p, q)$	0	$\frac{2\pi}{Z}$	...	$\frac{2\pi}{Z}(Z-2)$	$\frac{2\pi}{Z}(Z-1)$
$d(p, q)$	$\frac{\lambda_0(Z-1)}{(n-1)Z}$	$\frac{\lambda_0(Z-2)}{(n-1)Z}$	...	$\frac{\lambda_0}{(n-1)Z}$	0

when the DOE is binary-phase-level, there is only one etch depth and it is equal to the designed wavelength, λ_0, if $n = 1.5$.

From the expression for the phase delay in Eq. (6-1), the complex transmission function of each phase element, $\underline{h}(p, q)$, can be expressed as

$$\underline{h}(p, q) = \exp\left[j\frac{2\pi}{Z} \cdot PI(p, q)\right]. \tag{6-5}$$

and the reconstructed diffraction intensity pattern, $I(r, s)$, can be calculated by

$$\begin{aligned} I(r, s, \lambda_0) &= \left| J^{-2}\,\mathrm{sinc}\left(\frac{r}{J}\right)\mathrm{sinc}\left(\frac{s}{J}\right) DFT\{\underline{h}(p, q)\}\right|^2 \\ &= \left| J^{-2}\,\mathrm{sinc}\left(\frac{r}{J}\right)\mathrm{sinc}\left(\frac{s}{J}\right) DFT\left\{\exp\left[j\frac{2\pi}{Z} \cdot PI(p, q)\right]\right\}\right|^2, \end{aligned} \tag{6-6}$$

and the intensity of the zeroth diffraction order can be calculated as

$$I(0, 0, \lambda_0) = \left| J^{-2} \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp\left[j \frac{2\pi}{Z} \cdot PI(p, q) \right] \right|^2 . \tag{6-7}$$

Now, if the DOE is illuminated by light with wavelength λ_1, then from Eq. (6-2), the phase delay $PD(p, q)$ with respect to $PI(p, q) = 0$ can be calculated as

$$\begin{aligned} PD(p, q) &= \frac{2\pi}{\lambda_1} \cdot [D - d(p, q)] \cdot n + \frac{2\pi}{\lambda_1} \cdot d(p, q) - \frac{2\pi}{\lambda_1} \cdot D \\ &= \frac{2\pi}{\lambda_1} \cdot [D - d(p, q)] \cdot (n - 1) \\ &= \frac{2\pi}{\lambda_1} \cdot \left[\frac{\lambda_0 (Z - 1)}{Z (n - 1)} - \frac{\lambda_0}{Z (n - 1)} \cdot [(Z - 1) - PI(p, q)] \right] \cdot (n - 1) \\ &= \left(\frac{2\pi}{Z} \right) \left(\frac{\lambda_0}{\lambda_1} \right) PI(p, q) \end{aligned} , \tag{6-8}$$

Basically, all the phase delays are now scaled by a constant factor, (λ_0 / λ_1), due to the wavelength change, and they are now equally spaced between 0 and $\left(\frac{2\pi}{Z} \right) \cdot \left(\frac{\lambda_0}{\lambda_1} \right) \cdot (Z - 1)$. The complex transmission function of each phase element, $\underline{h}(p, q)$, now becomes

$$\underline{h}(p, q) = \exp\left[j \left(\frac{2\pi}{Z} \right) \cdot \left(\frac{\lambda_0}{\lambda_1} \right) \cdot PI(p, q) \right] . \tag{6-9}$$

The reconstructed diffraction intensity pattern, $I(r, s, \lambda_1)$, now becomes

$$
\begin{aligned}
I(r, s, \lambda_1) &= \left| J^{-2} \operatorname{sinc}\left(\frac{r}{J}\right) \operatorname{sinc}\left(\frac{s}{J}\right) DFT\{\underline{h}(p, q)\} \right|^2 \\
&= \left| J^{-2} \operatorname{sinc}\left(\frac{r}{J}\right) \operatorname{sinc}\left(\frac{s}{J}\right) DFT\left\{ \exp\left[j\left(\frac{2\pi}{Z}\right) \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot PI(p, q) \right] \right\} \right|^2,
\end{aligned}
\tag{6-10}
$$

and the intensity of the zeroth diffraction order can be calculated as

$$
I(0, 0, \lambda_1) = \left| J^{-2} \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp\left[j\left(\frac{2\pi}{Z}\right) \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot PI(p, q) \right] \right|^2 . \tag{6-11}
$$

From above, it is clear that the relation between $I(r, s, \lambda_1)$ and $I(r, s, \lambda_0)$ are, in general, dependent on both the phase distribution of the DOE and the diffraction order index (r, s).

Let's examine a few general properties of DOE reconstruction with incorrect illumination wavelength. First, when the wavelength change is such that the phase delay of each phase element becomes the original phase delay (when illuminating with the designed wavelength, λ_0) plus an integer multiples of 2π, then the diffraction efficiency of each diffraction order will become the same as the original diffraction efficiencies. From Eq. (6-1) and Eq. (6-8), we can see that if

$$
2\pi\left[\frac{PI(p, q)}{Z}\right]\left(\frac{\lambda_0}{\lambda_1}\right) = 2\pi\left[\frac{PI(p, q)}{Z}\right] + 2k\pi \tag{6-12}
$$

(k is some positive integer), then $I(r, s, \lambda_1) = I(r, s, \lambda_0)$. Equation (6-12) can be simplified to

$$\left(\frac{\lambda_0}{\lambda_1}\right) = kZ + 1 . \tag{6-13}$$

Note that this phenomenon only occurs for $(\lambda_0/\lambda_1) \geq 1$ and does not occur for $(\lambda_0/\lambda_1) < 1$, since in that case, the phase delay of each phase element with the incorrect illumination wavelength will be always smaller than the original phase delay and Eq. (6-12) and Eq. (6-13) will not be satisfied.

From properties of Discrete Fourier Transform (*DFT*), if $G(r, s) = DFT\{\underline{h}(p, q)\}$, then $G^*(-r, -s) = DFT\{\underline{h}^*(p, q)\}$ (* denotes complex conjugate). When the wavelength change is such that the phase delay of each phase element becomes the negative of the original phase delay (when illuminating with λ_0) plus an integer multiple of 2π, the reconstructed diffraction pattern will be spatially reversed from the original diffraction pattern. From Eq. (6-1) and Eq. (6-8), we can see that if

$$2\pi\left[\frac{PI(p, q)}{Z}\right]\left(\frac{\lambda_0}{\lambda_1}\right) = -2\pi\left[\frac{PI(p, q)}{Z}\right] + 2k\pi , \tag{6-14}$$

then $I(r, s, \lambda_1) = I(-r, -s, \lambda_0)$. Eq. (6-14) can be simplified to

$$\left(\frac{\lambda_0}{\lambda_1}\right) = kZ - 1 . \tag{6-15}$$

Note that this phenomenon only occurs for $(\lambda_0/\lambda_1) \geq 1$ and does not occur for $(\lambda_0/\lambda_1) < 1$, since in that case, the phase delay of each phase element with the incorrect illumination wavelength will be always smaller than the original phase delay and Eq. (6-14) and (6-15) will not be satisfied.

Lastly, when the wavelength change is such that the phase delay of each phase element becomes an integer multiple of 2π, all the light will go to the zeroth diffraction order. In this case, the phase delay of every phase element is essentially the same and there will be no diffraction pattern. From Eq. (6-8), we can see that this condition happens when

$$2\pi\left[\frac{PI(p,q)}{Z}\right]\left(\frac{\lambda_0}{\lambda_1}\right) = 2k\pi. \tag{6-16}$$

Eq. (6-16) can be satisfied if

$$\left(\frac{\lambda_0}{\lambda_1}\right) = kZ. \tag{6-17}$$

Note that this phenomenon only occurs for $\left(\frac{\lambda_0}{\lambda_1}\right) \geq 1$ and does not occur for $\left(\frac{\lambda_0}{\lambda_1}\right) < 1$.

The effect of wavelength change on the DOE reconstruction has been simulated for two groups of special purpose DOEs. The first group of DOEs are designed for a 3×3 spot array generator. The spot array generator has a target fan-out pattern of

$$\text{Sopt Array} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{6-18}$$

The DOEs have 8×8 phase elements in one period and two, four, and eight phase levels. The designed wavelength for all the DOEs is $850nm$. We have simulated the effect of wavelength changes on the DOE reconstruction from $800nm$ to $900nm$. Table 6-2 lists the non-uniformity (*NU*), minimum signal-to-noise ratio (*MinSNR*), and diffraction efficiency (η) of the original designed DOEs with two, four, and eight phase levels. Those performance indicators of DOEs are defined as

$$NU = \frac{\max(\text{signal intensity}) - \min(\text{signal intensity})}{\max(\text{signal intensity}) + \min(\text{signal intensity})} \times 100\%, \tag{6-19}$$

$$MinSNR = \frac{\min(\text{signal intensity})}{\max(\text{noise intensity})}, \tag{6-20}$$

and

$$\eta = \frac{\text{Light Energy in Signal Orders}}{\text{Total Light Energy}} \times 100\%, \tag{6-21}$$

In Eq. (6-20), the maximum of noise intensity is the largest intensity among spurious diffraction orders. Note that for non-uniformity, lower numbers are better, whereas for minimum signal-to-noise ratio and diffraction efficiency, higher numbers are better. In

each simulation, the diffraction efficiency, non-uniformity, minimum signal-to-noise ratio, diffraction efficiency of each desired diffraction order, percentage error of each desired diffraction order, and ratio of the diffraction efficiencies between each desired diffraction order and the zeroth diffraction order are plotted as a function of the illumination wavelength. Figure 6-2, 6-3, and 6-4 show the simulation results for these DOEs.

Table 6-2: DOE performance for the spot array generator. An 850nm wavelength was assumed for design, etch, and reconstruction.

J	Z	Non-uniformity (%)	MinSNR	Diffraction Efficiency (%)
8	2	4.3	6.4	58.5
8	4	3.6	7.8	70.9
8	8	2.7	21.1	81.3

The second group of DOEs are designed for a 3×3 triangular connection pattern. The triangular connection pattern has a target fan-out pattern of

$$Triangular = \begin{bmatrix} 0.25 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{bmatrix}. \tag{6-22}$$

The DOE has 16×16 phase elements in one period. Table 6-3 lists the non-uniformity, minimum signal-to-noise ratio, and diffraction efficiency of the original designed DOEs with two, four, and eight phase levels. In each simulation, the diffraction efficiency, non-uniformity, minimum signal-to-noise ratio, diffraction efficiency of each desired diffraction order, percentage error of each desired diffraction order, and ratio of the diffraction

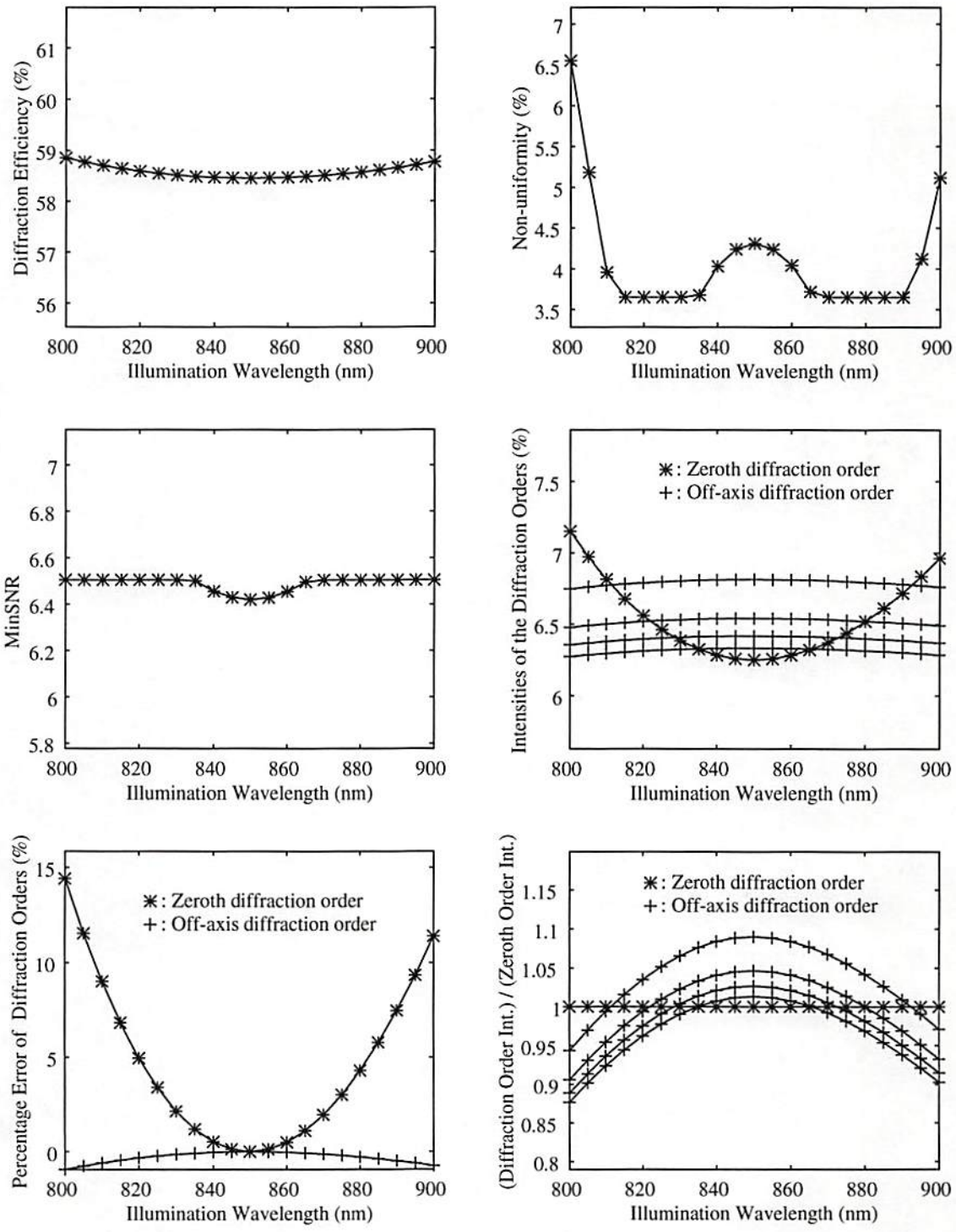

Figure 6-2: The effect of wavelength changes on DOE reconstruction for the two-phase-level DOE designed and etched at 850 nm for the spot array target pattern.

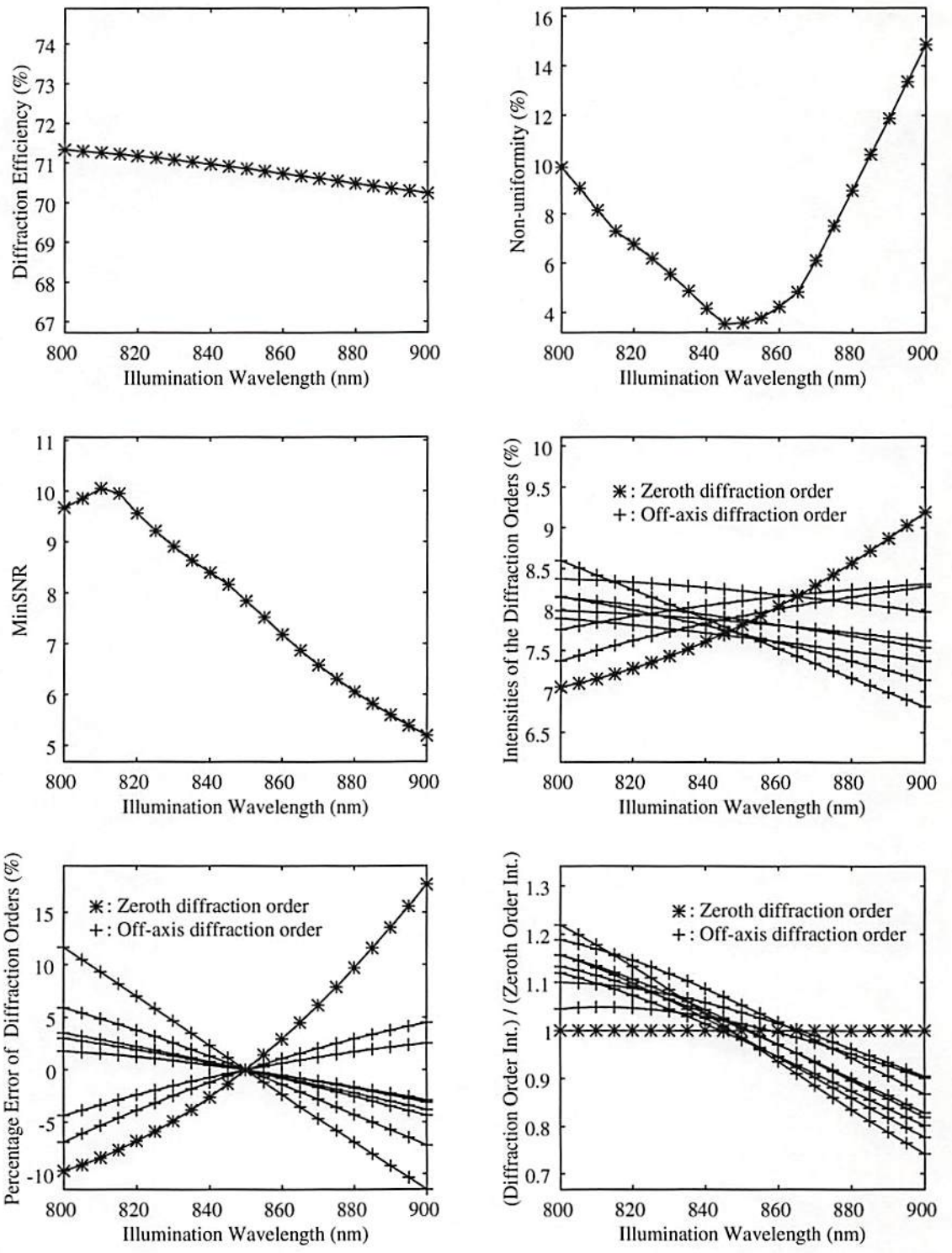

Figure 6-3: The effect of wavelength changes on DOE reconstruction for the four-phase-level DOE designed and etched at 850 nm for the spot array target pattern.

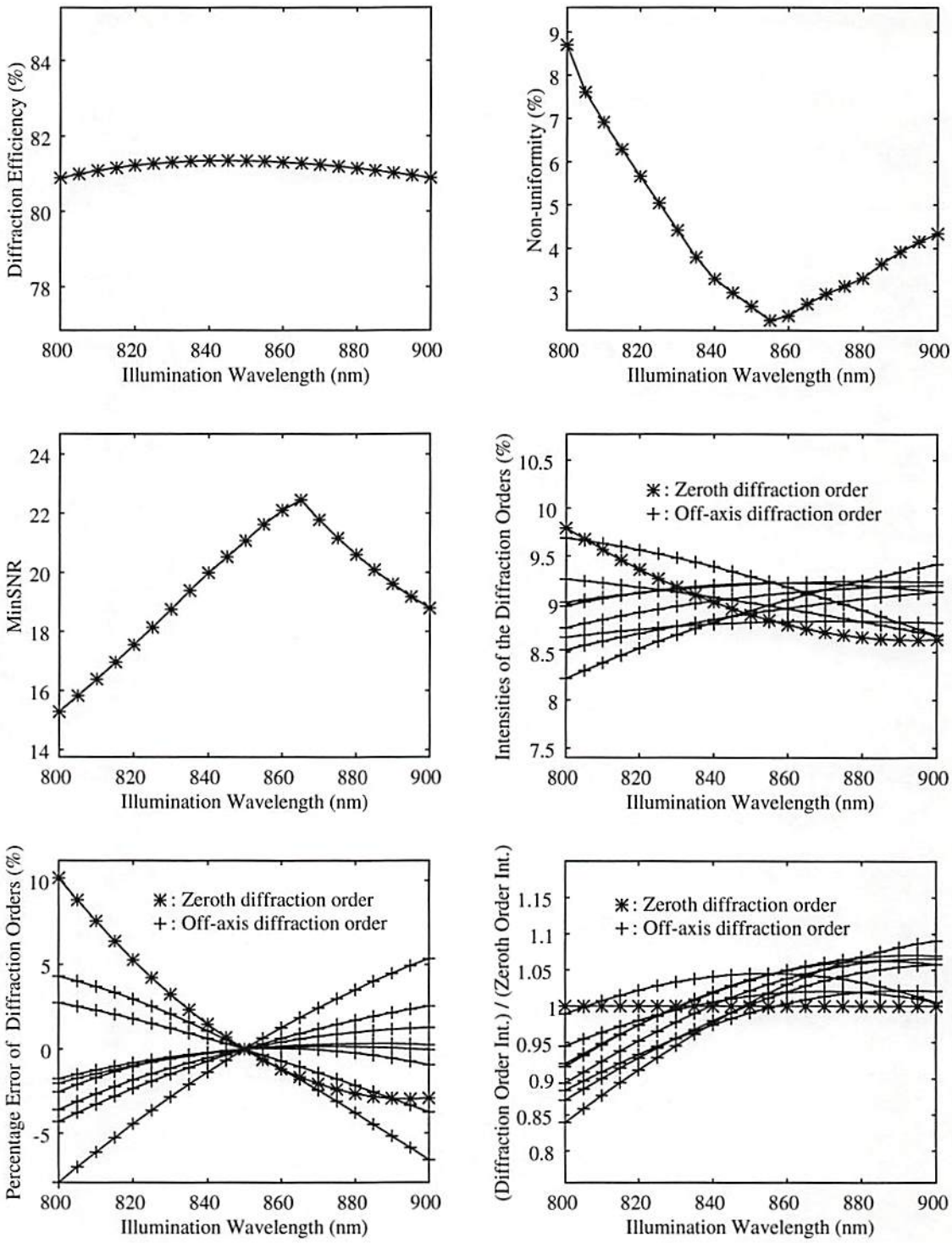

Figure 6-4: The effect of wavelength changes on DOE reconstruction for the eight-phase-level DOE designed and etched at 850 nm for the spot array target pattern.

Table 6-3: Performance of the DOEs designed for the triangular pattern. An 850nm wavelength was assumed for design, etch, and reconstruction.

J	Z	Non-uniformity (%)	MinSNR	Diffraction Efficiency (%)
16	2	1.7	9.1	62.6
16	4	2.9	17.4	69.4
16	8	1.7	27.4	76.3

efficiencies between each desired diffraction order and the zeroth diffraction order are plotted as a function of the illumination wavelength. Figure 6-5, 6-6, and 6-7 show the results of the simulation for the triangular pattern.

From Figure 6-2 to Figure 6-7, we can see that, for the DOEs considered here, the DOE performance (the non-uniformity, minimum signal-to-noise ratio, and diffraction efficiency) usually degrades if the illumination wavelength is different from the designed wavelength. From the percentage error plots in the figures for four or eight-phase-level DOEs, we can see that the percentage deviation from the original diffraction intensity for each diffraction order is dependent on the DOE phase distribution, and this percentage deviation also varies from one diffraction order to another. This phenomenon is consistent with the results we obtained from Eq. (6-10) and (6-11) for multiple-phase-level DOEs. On the other hand, from percentage error plots in the figures for two-phase-level DOEs, we can see that the percentage deviation from the original diffraction intensity for each off-axis diffraction order seems the same for all the off-axis diffraction orders. Furthermore, the minimum signal-to-noise ratio plot in Figure 6-5 stays unchanged when the wavelength changes. The minimum signal-to-noise ratio plot in Figure 6-2 also stays unchanged when the zeroth diffraction order is the not smallest signal order. Even further, the two-phase-level DOEs still generate symmetric diffraction patterns when the

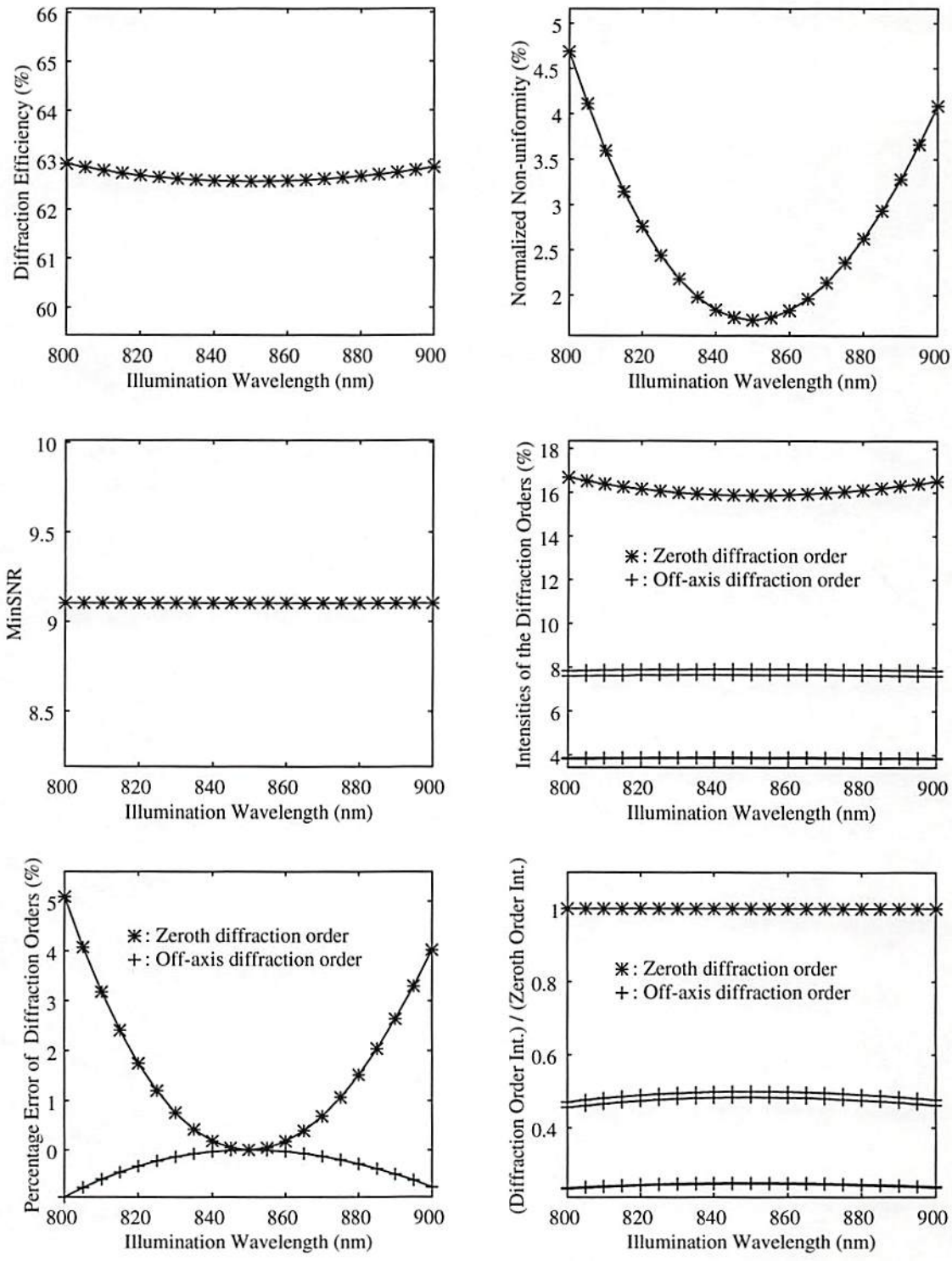

Figure 6-5: The effect of wavelength changes on DOE reconstruction for the two-phase-level DOE designed and etched at 850 nm for the triangular target pattern.

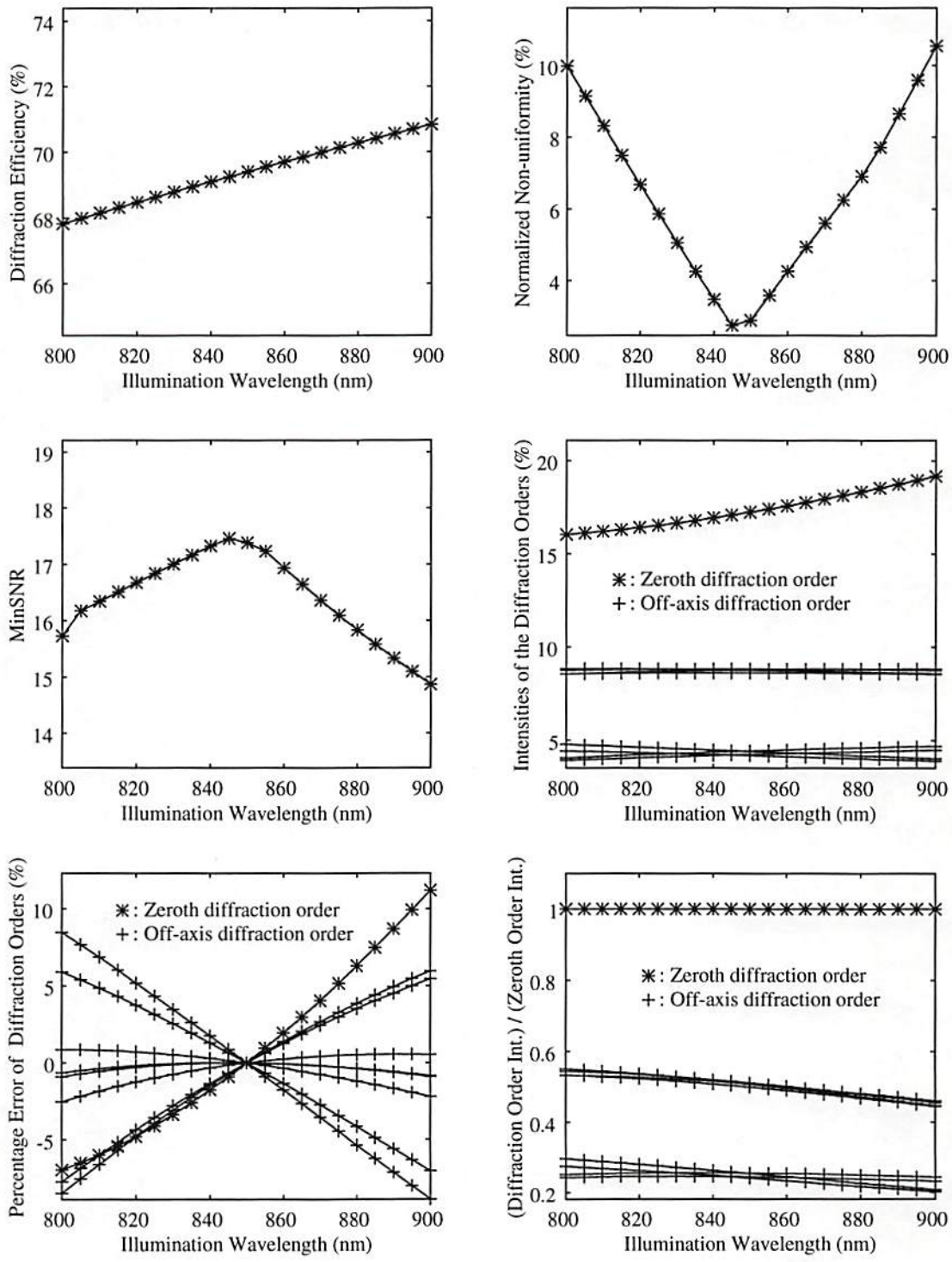

Figure 6-6: The effect of wavelength changes on DOE reconstruction for the four-phase-level DOE designed and etched at 850 nm for the triangular target pattern.

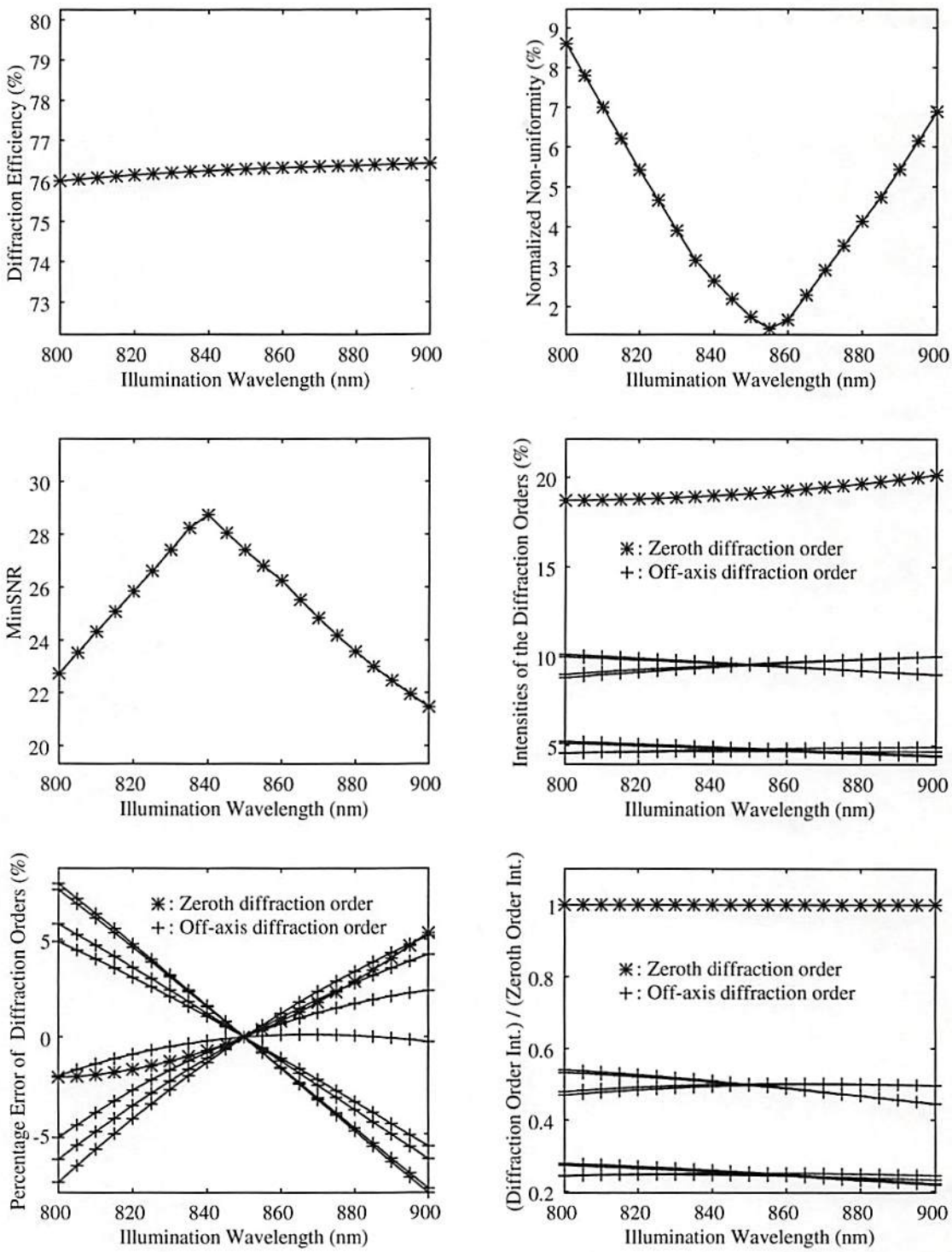

Figure 6-7: The effect of wavelength changes on DOE reconstruction for the eight-phase-level DOE designed and etched at 850 nm for the triangular target pattern.

illumination wavelength changes as can be seen from Figure 6-2 and Figure 6-5. Therefore, it seems that two-phase-level DOEs have special properties different from four and eight-phase-level DOEs when illumination wavelength changes. These phenomena are not obvious from the results we obtained in Eq. (6-10) and (6-11). In the next section, we will further investigate these special properties for two-phase-level DOEs.

6.2 Special Properties for Binary DOEs When Wavelength Changes

Let's first investigate the phenomenon that two-phase-level DOEs always generate symmetric intensity patterns regardless of the illuminating wavelength. Assuming that we illuminate a two-phase-level DOE with the designed wavelength λ_0, the intensities of the (r, s) 'th and $(-r, -s)$ 'th diffraction orders can be calculated from Eq. (6-6) as

$$I(r, s, \lambda_0) = \left| S \bullet DFT\{ \exp[j\pi \cdot PI(p, q)] \} \right|^2, \tag{6-23}$$

and

$$\begin{aligned}
& I(-r, -s, \lambda_0) \\
&= \left| S \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp[j\pi \cdot PI(p, q)] \bullet e^{-j2\pi \frac{-rp - sq}{J}} \right|^2 \\
&= \left| S \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp[j\pi \cdot PI(p, q)] \bullet e^{j2\pi \frac{rp + sq}{J}} \right|^2 \\
&= \left| \left[S \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp[-j\pi \cdot PI(p, q)] \bullet e^{-j2\pi \frac{rp + sq}{J}} \right]^* \right|^2 \\
&= \left| S \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp[-j\pi \cdot PI(p, q)] \bullet e^{-j2\pi \frac{rp + sq}{J}} \right|^2 ,
\end{aligned} \tag{6-24}$$

where

$$S \equiv \frac{\sin c\left(\frac{r}{J}\right) \bullet \sin c\left(\frac{s}{J}\right)}{J^2} \tag{6-25}$$

Since the DOE is binary phase level, the phase level index can only be 0 or 1. Therefore $\exp[-j\pi \cdot PI(p, q)]$ and $\exp[j\pi \cdot PI(p, q)]$ are essentially the same and the reconstructed intensity will be symmetric upon spatial inversion about the origin ($I(-r, -s, \lambda_0) = I(r, s, \lambda_0)$). Basically, this result comes directly from the property that the Fourier transform of a real sequence is always hermitian symmetric since the transmittance of a binary-phase-level DOE is real.

When the illumination wavelength changes to λ_1, the intensities of (r, s) 'th and $(-r, -s)$ 'th diffraction orders can be calculated from Eq. (6-10) as

$$I(r, s, \lambda_1) = \left| S \bullet DFT\left\{ \exp\left[j\pi \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot PI(p, q) \right] \right\} \right|^2, \tag{6-26}$$

and

$$\begin{aligned}
& I(-r, -s, \lambda_1) \\
&= \left| S \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp\left[j\pi \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot PI(p, q) \right] \bullet e^{-j2\pi \frac{-rp - sq}{J}} \right|^2 \\
&= \left| S \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp\left[j\pi \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot PI(p, q) \right] \bullet e^{j2\pi \frac{rp + sq}{J}} \right|^2 \\
&= \left| \left[S \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp\left[-j\pi \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot PI(p, q) \right] \bullet e^{-j2\pi \frac{rp + sq}{J}} \right]^* \right|^2 \\
&= \left| S \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp\left[-j\pi \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot PI(p, q) \right] \bullet e^{-j2\pi \frac{rp + sq}{J}} \right|^2 .
\end{aligned} \tag{6-27}$$

Although the DOE under consideration is binary, $\exp[j\pi \cdot (\lambda_0/\lambda_1) \cdot PI(p,q)]$ and $\exp[-j\pi \cdot (\lambda_0/\lambda_1) \cdot PI(p,q)]$ are not identical if $(\lambda_0/\lambda_1) \neq$ integer. Therefore, we can not use the property of Fourier transform as above to prove symmetry of the reconstructed intensity pattern. We can rewrite Eq. (6-26) to be

$$
\begin{aligned}
& I(r,s,\lambda_1) \\
&= \left| S \bullet \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} \exp\left[j\pi \cdot \left(\frac{\lambda_0}{\lambda_1} \right) \cdot PI(p,q) \right] \bullet e^{-j2\pi \frac{rp+sq}{J}} \right|^2 \\
&= \left| S \bullet \left[\sum_{\substack{p=0 \\ PI(p,q)=0}}^{(J-1)} \sum_{q=0}^{(J-1)} e^{-j2\pi \frac{rp+sq}{J}} + \sum_{\substack{p=0 \\ PI(p,q)=1}}^{(J-1)} \sum_{q=0}^{(J-1)} e^{j\pi \left(\frac{\lambda_0}{\lambda_1} \right)} \bullet e^{-j2\pi \frac{rp+sq}{J}} \right] \right|^2 .
\end{aligned}
\tag{6-28}
$$

Define

$$
Q \equiv \sum_{\substack{p=0 \\ PI(p,q)=0}}^{(J-1)} \sum_{q=0}^{(J-1)} e^{-j2\pi \frac{rp+sq}{J}} , \tag{6-29}
$$

and we can further simplify Eq. (6-28) to

$$
\begin{aligned}
& I(r,s,\lambda_1) \\
&= \left| S \bullet \left[Q + \sum_{\substack{p=0 \\ PI(p,q)=1}}^{(J-1)} \sum_{q=0}^{(J-1)} e^{j\pi \left(\frac{\lambda_0}{\lambda_1} \right)} \bullet e^{-j2\pi \frac{rp+sq}{J}} \right] \right|^2 \\
&= \left| S \bullet \left[Q + e^{j\pi \left(\frac{\lambda_0}{\lambda_1} \right)} \bullet \left\{ \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} e^{-j2\pi \frac{rp+sq}{J}} - Q \right\} \right] \right|^2 .
\end{aligned}
\tag{6-30}
$$

Using the identity

$$\sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} e^{-j2\pi \frac{rp+sq}{J}} = 0 \qquad (r,s) \neq (0,0)\,, \tag{6-31}$$

Eq. (6-30) can be expressed as

$$\begin{aligned} & I(r,s,\lambda_1) \\ &= \left| S \bullet \left[Q + e^{j\pi \left(\frac{\lambda_0}{\lambda_1}\right)} \bullet \{0-Q\} \right] \right|^2 \qquad (r,s) \neq (0,0) \\ &= S^2 \bullet |Q|^2 \bullet \left| 1 - \exp\left(j\pi \left(\frac{\lambda_0}{\lambda_1} \right) \right) \right|^2 \end{aligned} \,. \tag{6-32}$$

By the same procedure, Eq. (6-26) can be expressed as

$$I(-r,-s,\lambda_1) = S^2 \bullet |Q|^2 \bullet \left| 1 - \exp\left(-j\pi \left(\frac{\lambda_0}{\lambda_1} \right) \right) \right|^2 \qquad (r,s) \neq (0,0)\,. \tag{6-33}$$

Because $1 - \exp[-j\pi(\lambda_0/\lambda_1)]$ and $1 - \exp[j\pi \cdot (\lambda_0/\lambda_1)]$ are vectors with different directions but the same magnitudes (as shown in Figure 6-8), the reconstructed intensity pattern will be symmetric regardless the wavelength of illumination. The simulations in Figure 6-2 and Figure 6-5 are consistent with these theoretical results.

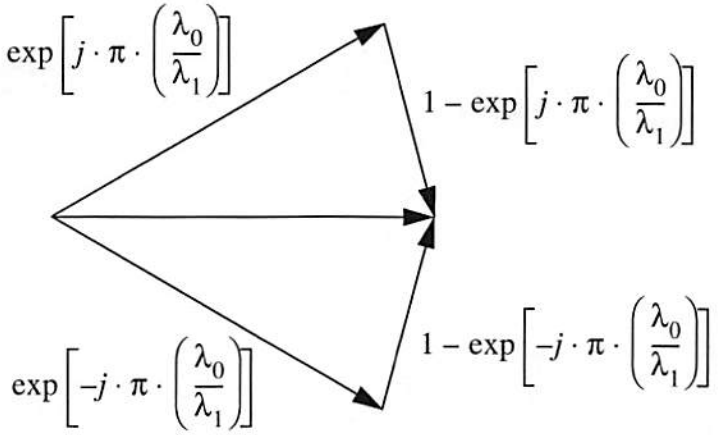

Figure 6-8: The diagram for proving that a two-phase-level DOE always generates a symmetric pattern even when the illumination wavelength changes.

Let's now find the relation between the intensity of each diffraction order with the correct wavelength and the same diffraction order with incorrect wavelength. By using the following identity

$$\sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} e^{-j2\pi\frac{rp+sq}{J}} = \begin{cases} 0 & (r,s) \neq (0,0) \\ J^2 & (r,s) = (0,0) \end{cases}, \tag{6-34}$$

We can find the diffraction intensity of any diffraction order from Eq. (6-30) and Eq. (6-32) as

$$
\begin{aligned}
& I(r, s, \lambda_1) \\
&= \left| S \bullet \left[Q + \sum_{\substack{p=0 \\ PI(p,q)=1}}^{(J-1)} \sum_{q=0}^{(J-1)} e^{j\pi\left(\frac{\lambda_0}{\lambda_1}\right)} \bullet e^{-j2\pi\frac{rp+sq}{J}} \right] \right|^2 \\
&= \left| S \bullet \left[Q + e^{j\pi\left(\frac{\lambda_0}{\lambda_1}\right)} \bullet \left\{ \sum_{p=0}^{(J-1)} \sum_{q=0}^{(J-1)} e^{-j2\pi\frac{rp+sq}{J}} - Q \right\} \right] \right|^2 \\
&= \begin{cases} S^2 \bullet |Q|^2 \bullet \left| 1 - \exp\left(j\pi\frac{\lambda_0}{\lambda_1} \right) \right|^2 & (r, s) \neq (0, 0) \\ S^2 \bullet \left| Q \bullet \left(1 - \exp\left(j\pi\frac{\lambda_0}{\lambda_1} \right) \right) + J^2 \bullet \exp\left(j\pi\frac{\lambda_0}{\lambda_1} \right) \right|^2 & (r, s) = (0, 0) \end{cases} \quad . \quad (6\text{-}35)
\end{aligned}
$$

Let M_0 represent the number of zero-phase-level-index phase elements in one period of the DOE and let M_1 represent the number of one-phase-level-index phase elements in one period of the DOE. The expression of Q for the zeroth diffraction order ($(r, s) = (0, 0)$) can be simplified as

$$
Q \equiv \sum_{\substack{p=0 \\ PI(p,q)=0}}^{(J-1)} \sum_{q=0}^{(J-1)} e^{-j2\pi\frac{rp+sq}{J}} = \sum_{\substack{p=0 \\ PI(p,q)=0}}^{(J-1)} \sum_{q=0}^{(J-1)} 1 = M_0 \, . \quad (6\text{-}36)
$$

Using the relation

$$
J^2 = M_0 + M_1 , \quad (6\text{-}37)
$$

Eq. (6-35) can be further simplified to

$$I(r,s,\lambda_1)=\begin{cases} S^2 \bullet |Q|^2 \bullet \left|1-\exp\left(j\pi\dfrac{\lambda_0}{\lambda_1}\right)\right|^2 & (r,s)\neq(0,0) \\ S^2 \bullet \left|M_0+M_1 \bullet \exp\left(j\pi\dfrac{\lambda_0}{\lambda_1}\right)\right|^2 & (r,s)=(0,0) \end{cases} . \qquad (6\text{-}38)$$

We can find $I(r,s,\lambda_0)$ by substituting $\lambda_1 = \lambda_0$ in Eq. (6-38) and simplifying it to

$$I(r,s,\lambda_0)=\begin{cases} S^2 \bullet |Q|^2 \bullet 2^2 & (r,s)\neq(0,0) \\ S^2 \bullet |M_0-M_1|^2 & (r,s)=(0,0) \end{cases} . \qquad (6\text{-}39)$$

Now, the ratio between $I(r,s,\lambda_1)$ and $I(r,s,\lambda_0)$ can be found from Eq. (6-38) and Eq. (6-39) as

$$\begin{aligned} &\frac{I(r,s,\lambda_1)}{I(r,s,\lambda_0)} \\ &= \begin{cases} \dfrac{\left|1-\exp\left(j\pi\dfrac{\lambda_0}{\lambda_1}\right)\right|^2}{4} & (r,s)\neq(0,0) \\ \dfrac{\left|M_0+M_1 \bullet \exp\left(j\pi\dfrac{\lambda_0}{\lambda_1}\right)\right|^2}{|M_0-M_1|^2} & (r,s)=(0,0) \end{cases} . \qquad (6\text{-}40) \\ &= \begin{cases} \sin^2\left[\left(\dfrac{\pi}{2}\right)\left(\dfrac{\lambda_0}{\lambda_1}\right)\right] & (r,s)\neq(0,0) \\ \dfrac{\left|M_0+M_1 \bullet \exp\left(j\pi\dfrac{\lambda_0}{\lambda_1}\right)\right|^2}{|M_0-M_1|^2} & (r,s)=(0,0) \end{cases} \end{aligned}$$

This result indicates that when the illuminating wavelength is different from the designed wavelength, the diffraction efficiency of any off-axis order can be calculated by

$$I(r, s, \lambda_1) = I(r, s, \lambda_0) \bullet \sin^2\left[\left(\frac{\pi}{2}\right)\left(\frac{\lambda_0}{\lambda_1}\right)\right] \qquad (r, s) \neq (0, 0) . \qquad (6\text{-}41)$$

This relation between the diffraction intensity at wavelength λ_1 and the originally designed diffraction intensity for a given off-axis order is independent of the phase distribution pattern of the DOE and is dependent only on the ratio of the designed wavelength and the actual illumination wavelength. Furthermore, the ratio between $I(r, s, \lambda_1)$ and $I(r, s, \lambda_0)$ is the same for all the off-axis diffraction orders and this ratio is periodic over the ratio of the wavelengths. It is also interesting to note from Eq. (6-41) that $I(r, s, \lambda_1)$ is always less than or equal to $I(r, s, \lambda_0)$. This means that if the illuminating wavelength is deviating from the designed wavelength, then the diffraction efficiencies of all off-axis diffraction orders will decrease by the same percentage. Therefore, the minimum signal-to-noise ratio (the weakest signal intensity to the strongest noise intensity) will remain the same for any illuminating wavelength if the zeroth diffraction order is not the weakest signal diffraction order. This phenomenon also implies that we can design a DOE such that the zeroth diffraction order is not part of the desired reconstruction pattern. Then, even if the illumination wavelength changes due to laser conditions or other reasons, the proportionality among the desired diffraction orders will remain unchanged. The simulations in Figure 6-2 and Figure 6-5 are in good agreement with our theoretical results. The intensity ratio for any off-axis diffraction order as a function of the wavelength ratio is plotted in Figure 6-9.

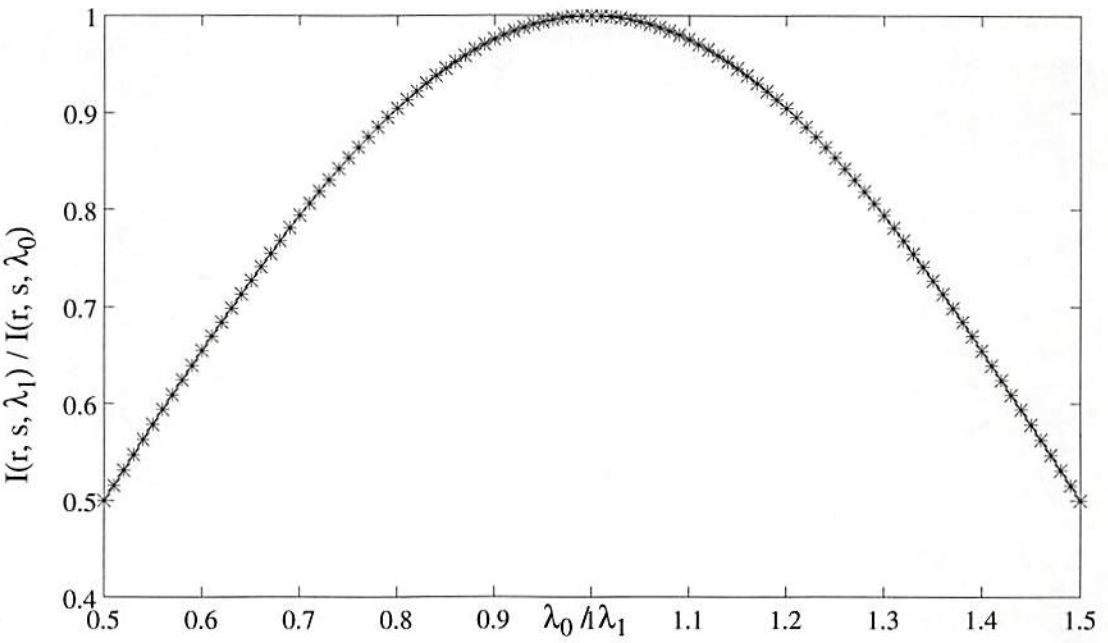

Figure 6-9: The intensity ratio of any off-axis diffraction order for a binary-phase-level DOE designed for wavelength λ_0 and reconstructed at wavelength λ_1 as a function of the wavelength ratio.

Since the diffraction efficiencies of all the off-axis diffraction orders decrease when illumination wavelength changes, the additional light energy must go to the zeroth order. Therefore, the diffraction efficiency of the zeroth order always increases when there is a wavelength change. This is also consistent with the result obtained in Figure 6-2 and Figure 6-5. From Eq. (6-38) and (6-40), the new diffraction efficiency of the zeroth order can be calculated by

$$I(0,0,\lambda_1)$$

$$= \begin{cases} I(0,0,\lambda_0) \bullet \dfrac{\left| M_0 + M_1 \bullet \exp\left(j\pi \dfrac{\lambda_0}{\lambda_1} \right) \right|^2}{\left| M_0 - M_1 \right|^2} & I(0,0,\lambda_0) \neq 0 \\ \dfrac{J^{-2}}{(M_0 - M_1)^2} \bullet \left| M_0 + M_1 \bullet \exp\left(j\pi \dfrac{\lambda_0}{\lambda_1} \right) \right|^2 & I(0,0,\lambda_0) = 0 \end{cases} . \qquad (6\text{-}42)$$

Note that, unlike for the off-axis diffraction orders, this relation is dependent on the phase distribution pattern of the DOE. Figure 6-10 show the intensity profiles of the two-phase-level DOE (designed for the triangular pattern and 850 *nm* wavelength) illuminated by the correct (850 *nm*) and the incorrect (633 *nm*) wavelengths. It can be clearly seen that, with the incorrect illumination wavelength, the intensity of the zeroth order is increased substantially while the intensities of all the off-axis diffraction orders are decreased by the same percentage.

The above analysis for binary-phase-level DOEs assumes that the designed relative phase delay is either 0 or π. However, it is possible to have a relative phase delay other than π (detailed in Chapter 8). In that case, the expressions we have derived above need to be modified. Assuming that the relative phase delay for the DOE is either 0 or φ, then by replacing π in Eq. (6-35) with φ, Eq. (6-38) becomes

$$I(r,s,\lambda_1) = \begin{cases} S^2 \bullet |Q|^2 \bullet \left| 1 - \exp\left[j\varphi \dfrac{\lambda_0}{\lambda_1} \right] \right|^2 & (r,s) \neq (0,0) \\ S^2 \bullet \left| M_0 + M_1 \bullet \exp\left[j\varphi \dfrac{\lambda_0}{\lambda_1} \right] \right|^2 & (r,s) = (0,0) \end{cases} , \qquad (6\text{-}43)$$

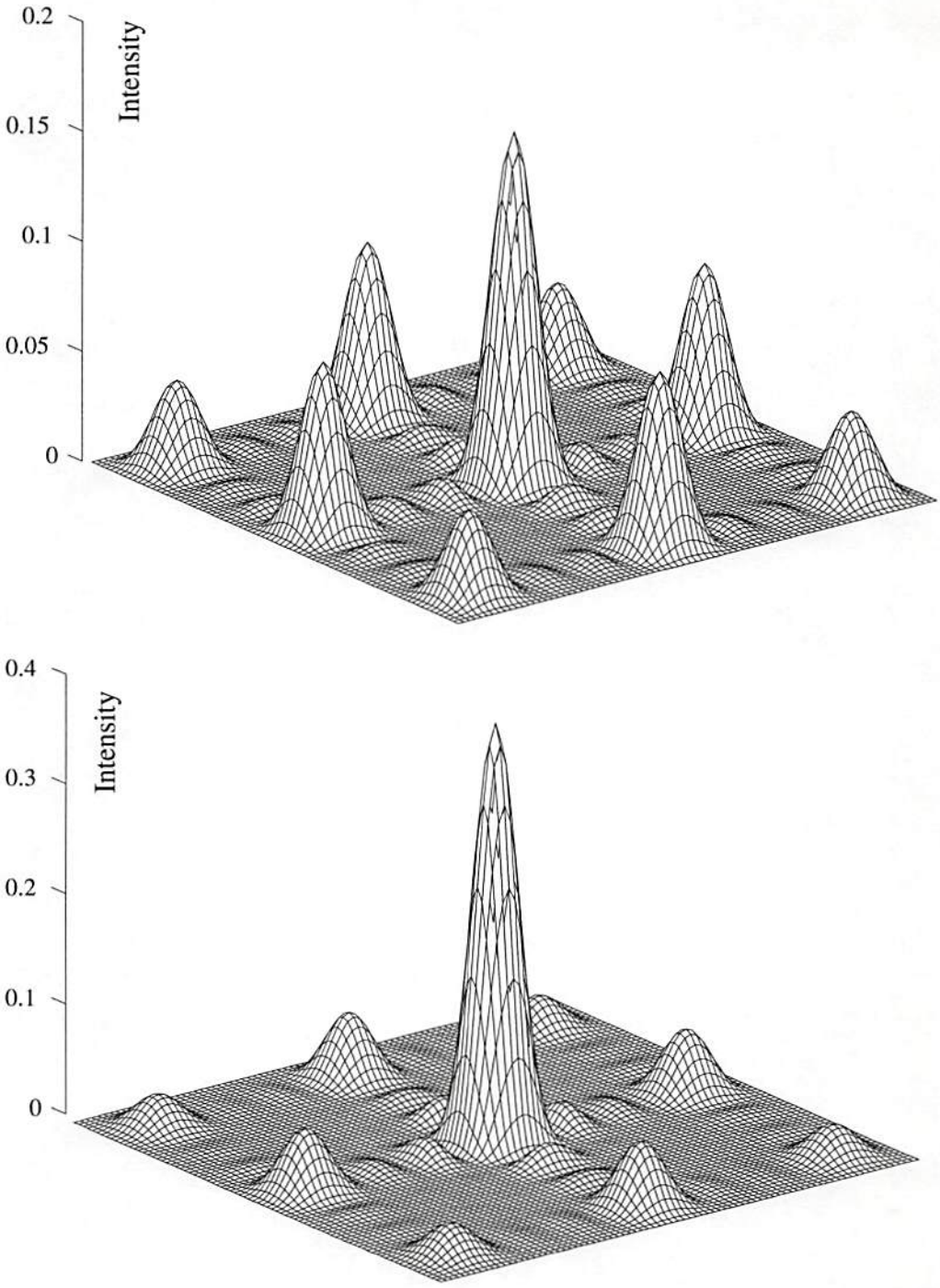

Figure 6-10: The intensity profiles of the two-phase-level DOE reconstruction (designed for the triangular pattern and 850 *nm* wavelength) illuminated with 850 *nm* (top) and 633 *nm* (bottom) wavelengths.

and Eq. (6-39) becomes

$$I(r,s,\lambda_0)=\begin{cases} S^2 \bullet |Q|^2 \bullet |1-\exp[j\varphi]|^2 & (r,s)\neq(0,0) \\ S^2 \bullet \left|M_0+M_1\cdot\exp[j\varphi]\right|^2 & (r,s)=(0,0) \end{cases} . \qquad (6\text{-}44)$$

Now, the ratio between $I(r,s,\lambda_1)$ and $I(r,s,\lambda_0)$ can be found from Eq. (6-43) and Eq. (6-44) as

$$\frac{I(r,s,\lambda_1)}{I(r,s,\lambda_0)}=\begin{cases} \dfrac{\left|1-\exp\left[j\varphi\dfrac{\lambda_0}{\lambda_1}\right]\right|^2}{|1-\exp[j\varphi]|^2} & (r,s)\neq(0,0) \\ \dfrac{\left|M_0+M_1\bullet\exp\left[j\varphi\dfrac{\lambda_0}{\lambda_1}\right]\right|^2}{\left|M_0+M_1\cdot\exp[j\varphi]\right|^2} & (r,s)=(0,0) \end{cases} . \qquad (6\text{-}45)$$

This result indicates that when the illuminating wavelength is different from the designed wavelength, the new diffraction efficiency of any off-axis order can be calculated by

$$I(r,s,\lambda_1) = I(r,s,\lambda_0)\bullet\frac{\left|1-\exp\left[j\varphi\dfrac{\lambda_0}{\lambda_1}\right]\right|^2}{|1-\exp[j\varphi]|^2} \qquad (r,s)\neq(0,0), \qquad (6\text{-}46)$$

and the new diffraction efficiency of the zeroth order can be calculated by

$$
\begin{aligned}
& I(0,0,\lambda_1) \\
&= \begin{cases} I(0,0,\lambda_0) \bullet \dfrac{\left| M_0 + M_1 \bullet \exp\left[j\varphi \dfrac{\lambda_0}{\lambda_1} \right] \right|^2}{\left| M_0 + M_1 \cdot \exp\left[j\varphi \right] \right|^2} & I(0,0,\lambda_0) \neq 0 \\ J^{-2} \bullet \left| M_0 + M_1 \bullet \exp\left[j\varphi \dfrac{\lambda_0}{\lambda_1} \right] \right|^2 & I(0,0,\lambda_0) = 0 . \end{cases}
\end{aligned}
\tag{6-47}
$$

The relation between the new diffraction intensity and the originally designed diffraction intensity for a given off-axis order is still independent of the phase distribution pattern of the DOE and is dependent only on the ratio of the designed wavelength to the actual illumination wavelength, and the phase delay φ. The ratio between $I(r, s, \lambda_1)$ and $I(r, s, \lambda_0)$ is still the same for all the off-axis diffraction orders. This means that if the illuminating wavelength is deviating from the designed wavelength, then the diffraction efficiencies of all off-axis diffraction orders will be changed by the same percentage. It is also interesting to note that, from Eq. (6-46) that $I(r, s, \lambda_1)$ for this non-conventional binary-phase-level DOE is not always less than or equal to $I(r, s, \lambda_0)$ as is the case for the conventional binary-phase-level DOE. From Eq. (6-46), for the case of $\varphi < \pi$, we can also find that the intensity of every off-axis diffraction order will increase if $\lambda_1 < \lambda_0$ and will decrease if $\lambda_1 > \lambda_0$. On the other hand, for the case of $\varphi > \pi$, the intensity of every off-axis diffraction order will increase if $\lambda_1 > \lambda_0$ and will decrease if $\lambda_1 < \lambda_0$. Therefore, it is possible that the intensity of the zeroth diffraction order will be decreased when there is a wavelength shift. This is quite different from the case for the conventional binary-phase-level DOE. The intensity ratio for any off-axis diffraction order as a function of the

wavelength ratio for an unconventional binary-phase-level DOE with phase delays of 0 and 0.82π is plotted in Figure 6-11.

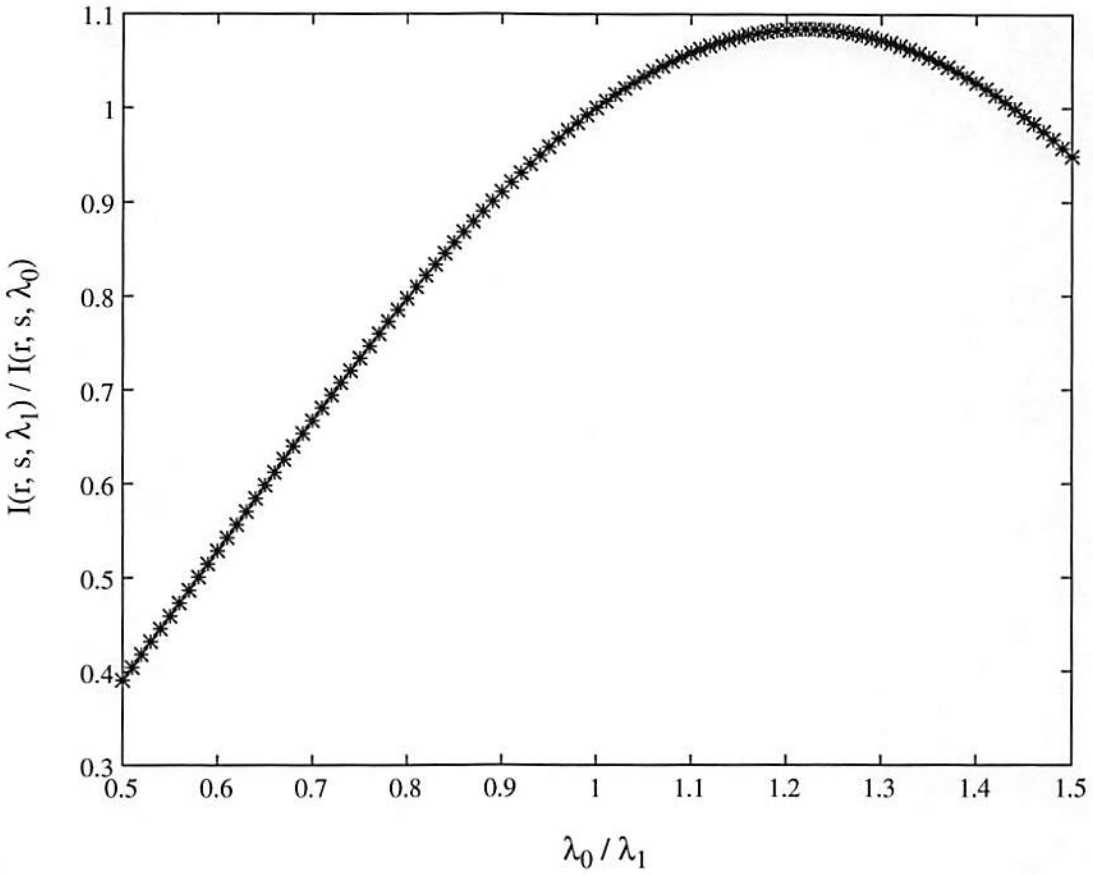

Figure 6-11: The intensity ratio of any off-axis diffraction order as a function of the wavelength ratio, for an unconventional binary-phase-level DOE with phase delays of 0 and 0.82π, designed for wavelength λ_0 and reconstructed at wavelength λ_1.

Chapter 7

The Effect of Etch Depth Error on Reconstruction of Diffractive Optical Elements

To fabricate a multiple-phase-level diffractive optical element, a few etching (or deposition) procedures that repeatedly etch into (or deposit onto) the DOE substrate are needed. In this chapter, we will investigate the effect of etch depth error during fabrication on the reconstruction of a diffractive optical element. We assumed the DOE is made by etch for the analysis in this chapter. A similar analysis can be done for DOEs fabricated by deposition. In Section 7.1, we will derive a general expression for DOE reconstruction when etch depth error occurs for multiple-phase-level DOEs. We will apply the results from Section 7.1 to the case of binary-phase-level DOEs in Section 7.2. We will verify that a binary-phase-level DOE with etch depth error appears equivalent to a binary-phase-level DOE with incorrect illumination wavelength and, therefore, has identical properties. The effect on the reconstruction of a binary-phase-level DOE when there are wavelength changes and etch depth error at the same time is also considered in Section 7.2. We have also considered the effect of wavelength change and etch depth error on the reconstruction for binary-phase-level DOEs with phase delays other than 0 and π. The effect of wavelength change and etch depth error for the unconventional binary-phase-level DOEs

is somewhat different from the effect of wavelength change and etch depth error for the conventional binary-phase-level DOEs. Section 7.3 and 7.4 consider multiple-phase-level DOEs (with Z possible phase levels) fabricated using binary-optics process ($\log_2 Z$ etch steps). In Section 7.3, we will study the case in which there is a constant percentage error in each etch step, and in Section 7.4 we will discuss the case that the etch depth errors in the etch steps are independent of each other.

7.1 Reconstruction of Diffractive Optical Elements Under Etch Depth Error

To generate the correct phase delay in each phase element, we have to etch into the physical area of that phase element on the DOE substrate with a correct etch depth. As derived in Chapter 6, for a Z-phase-level DOE, the correct depth, $d(p, q)$, of phase element (p, q), can be expressed as

$$d(p, q) = \frac{\lambda_0}{Z(n-1)} \cdot [(Z-1) - PI(p, q)], \tag{7-1}$$

where $PI(p, q)$ is the phase level index (p and q are the indices of the phase elements with $p, q \in \{0, 1, ..., J-1\}$), n is the index of refraction of the DOE substrate (we have assumed the index of refraction is one outside the DOE), and λ_0 is the designed wavelength. From Eq. (7-1), we can see that the correct etch depth for any phase element is just some integer multiple of $\frac{\lambda_0}{Z(n-1)}$. This integer is determined by

$[\,(Z-1)-PI(p,q)]$ and is in the range of $\{0, 1, 2, \ldots, (Z-1)\}$. Let's express $[\,(Z-1)-PI(p,q)]$ by its binary representation as

$$[\,(Z-1)-PI(p,q)] = \sum_{i=1}^{\log_2 Z} e_i(p,q)\cdot 2^{i-1}. \tag{7-2}$$

Using Eq. (7-2), the correct etch depth of phase element (p, q) in Eq. (7-1) can then be expressed as

$$\begin{aligned} d(p,q) &= \frac{\lambda_0}{Z(n-1)}\cdot\sum_{i=1}^{\log_2 Z} e_i(p,q)\cdot 2^{i-1} \\ &= \sum_{i=1}^{\log_2 Z} e_i(p,q)\cdot\left[2^{i-1}\cdot\frac{\lambda_0}{Z(n-1)}\right] \\ &= \sum_{i=1}^{\log_2 Z} e_i(p,q)\cdot d_i \end{aligned}, \tag{7-3}$$

in which

$$d_i \equiv 2^{i-1}\cdot\frac{\lambda_0}{Z(n-1)}. \tag{7-4}$$

From Eq. (7-3), we can see that, to fabricate a Z-phase-level DOE, we only need to perform the etch at most $\log_2 Z$ times rather than Z times and the correct depth for i'th etch step, d_i, is defined in Eq. (7-4). In this binary-optics fabrication process [10], at the i'th etch step, if the i'th bit of the binary representation of $[\,(Z-1)-PI(p,q)]$ for any phase element is one, then an etch of a depth of d_i is performed otherwise no etch is needed. In this sense,

the maximum possible etch depth, D, is just the sum of all the etch depths from every etch step and it can be easily shown to be

$$D = \sum_{i=1}^{\log_2 Z} d_i = \frac{\lambda_0 (Z-1)}{Z(n-1)}. \tag{7-5}$$

Let's now calculate the relative phase delay, $PD(p, q)$, in terms of the etch depth. The phase delay (relative to the phase element with a zero phase level index) introduced by each phase element with phase level index $PI(p, q)$ then can be calculated by summing the phase delays due to both non-etched and etched steps then subtracting the phase delay for the phase element with a zero phase level index as follows:

$$\begin{aligned}
&PD(p, q) \\
&= \frac{2\pi n}{\lambda_0}\left[D - \sum_{i=1}^{\log_2 Z} e_i(p, q)\, d_i\right] + \frac{2\pi}{\lambda_0}\sum_{i=1}^{\log_2 Z} e_i(p, q)\, d_i - \frac{2\pi}{\lambda_0} D \\
&= \frac{2\pi}{\lambda_0}\left[nD - n\sum_{i=1}^{\log_2 Z} e_i(p, q)\, d_i + \sum_{i=1}^{\log_2 Z} e_i(p, q)\, d_i - D\right] \\
&= \frac{2\pi}{\lambda_0}\left[(n-1)D - (n-1)\sum_{i=1}^{\log_2 Z} e_i(p, q)\, d_i\right] \\
&= \frac{2\pi}{\lambda_0}(n-1)\left[D - \sum_{i=1}^{\log_2 Z} e_i(p, q)\, d_i\right] \\
&= \frac{2\pi}{\lambda_0}(n-1)\left[\sum_{i=1}^{\log_2 Z} d_i - \sum_{i=1}^{\log_2 Z} e_i(p, q)\, d_i\right] \\
&= \frac{2\pi}{\lambda_0}(n-1)\sum_{i=1}^{\log_2 Z} [1 - e_i(p, q)]\, d_i
\end{aligned} \tag{7-6}$$

Due to fabrication error, the actual etch depth in i'th etch step will be different from d_i. Assuming the actual etch depth in i'th etch step can be expressed as $[1 + r_i]\, d_i$, where r_i is the percentage fabrication error, then the actual phase delay in Eq. (7-6) can be expressed as

$$PD(p, q) = \frac{2\pi}{\lambda_0}(n-1) \sum_{i=1}^{\log_2 Z} [1 - e_i(p, q)]\, [1 + r_i]\, d_i . \tag{7-7}$$

Substituting the expression for d_i given in Eq. (7-4), Eq. (7-7) becomes

$$\begin{aligned} & PD(p, q) \\ &= \frac{2\pi}{\lambda_0}(n-1) \sum_{i=1}^{\log_2 Z} [1 - e_i(p, q)] \cdot [1 + r_i] \cdot 2^{i-1} \cdot \frac{\lambda_0}{Z(n-1)} \\ &= \frac{2\pi}{Z} \sum_{i=1}^{\log_2 Z} [1 - e_i(p, q)] \cdot [1 + r_i] \cdot 2^{i-1} \end{aligned} \quad . \tag{7-8}$$

Note that in this chapter, we use the zero phase level as a reference for zero phase delay and zero phase delay error. Therefore, the phase delay of a given phase element is associated with the etch depth error of the etch steps for which the given phase element is not etched. This is because the phase delay is calculated with respect to the zeroth phase level ($PI(p, q) = 0$) and the phase elements with a zero phase level will be etched in all etch steps. Therefore, the etch depth error of any given phase element will also appear in the etch depth error of the zeroth phase level elements. The complex transmission function of

each phase element, $\underline{h}(p, q)$, which is the complex exponential function of the phase delay, $\exp[j \cdot PD(p, q)]$, now becomes

$$\underline{h}(p, q) = \exp\left[j\left(\frac{2\pi}{Z}\right) \cdot \sum_{i=1}^{\log_2 Z} [1 - e_i(p, q)] \cdot [1 + r_i] \cdot 2^{i-1} \right]. \tag{7-9}$$

The reconstructed diffraction intensity pattern under etch depth error, $II(r, s)$, can then be calculated by

$$\begin{aligned} & II(r, s) \\ &= \left| J^{-2} \operatorname{sinc}\left(\frac{r}{J}\right) \operatorname{sinc}\left(\frac{s}{J}\right) DFT\{\underline{h}(p, q)\} \right|^2. \end{aligned} \tag{7-10}$$

7.2 The Effect of Etch Depth Error on Reconstruction for Binary-Phase-Level Diffractive Optical Elements

For binary phase level DOEs, there is only one etch step. Because the binary representation for $[(Z-1) - PI(p, q)]$ is just $e_1(p, q)$ in this case, the phase delay in Eq. (7-8) can be directly simplified to

$$\begin{aligned} & PD(p, q) \\ &= \frac{2\pi}{Z} \sum_{i=1}^{\log_2 Z} [1 - e_i(p, q)] \cdot [1 + r_i] \cdot 2^{i-1} \\ &= \frac{2\pi}{2} \cdot [1 - e_1(p, q)] \cdot [1 + r_1] \\ &= \pi \cdot PI(p, q) \cdot [1 + r_1] \end{aligned} \quad . \qquad (7\text{-}11)$$

Comparing Eq. (7-11) with Eq. (6-8), it is clear that the effect of etch depth error for binary phase level DOEs is equivalent to the effect due to wavelength change. This is because both cases are effectively changing the relative phase delay in the DOEs. The $\left(\frac{\lambda_0}{\lambda_1}\right)$ term in Eq. (6-8) is simply replaced by $[1 + r_1]$ in Eq. (7-11). Therefore, the effect of etch depth error on DOE reconstructions should be similar to the effect of wavelength change. With similar derivation steps with the derivation steps in Section 6.2, the intensity of the reconstructed diffraction order with etch depth error for binary phase level DOEs can be shown to be

$$II(r,s) = \begin{cases} S^2 \bullet |Q|^2 \bullet \left|1 - \exp\left(j\pi\left[1 + r_1\right]\right)\right|^2 & (r,s) \neq (0,0) \\ S^2 \bullet \left|M_0 + M_1 \bullet \exp\left(j\pi\left[1 + r_1\right]\right)\right|^2 & (r,s) = (0,0) \end{cases}, \qquad (7\text{-}12)$$

with

$$S \equiv \frac{\operatorname{sinc}\left(\frac{r}{J}\right) \bullet \operatorname{sinc}\left(\frac{s}{J}\right)}{J^2}, \qquad (7\text{-}13)$$

and

$$Q \equiv \sum_{\substack{p=0 \\ PI(p,q)=0}}^{(J-1)} \sum_{q=0}^{(J-1)} e^{-j2\pi\frac{rp+sq}{J}}. \qquad (7\text{-}14)$$

In Eq. (7-12), M_0 represents the number of zero-phase-level-index phase elements in one period of the DOE and M_1 represents the number of one-phase-level-index phase elements in one period of the DOE. The ratio of the reconstructed diffraction intensity for a given diffraction order between the DOE with etch depth error, $II(r,s)$, and the DOE with correct etch depth, $I(r,s)$, can be similarly found as

$$\frac{II(r,s)}{I(r,s)} = \begin{cases} Sin^2\left[\left(\frac{\pi}{2}\right)(1+r_1)\right] & (r,s) \neq (0,0) \\ \dfrac{\left|M_0 + M_1 \bullet \exp\left[j\pi(1+r_1)\right]\right|^2}{\left|M_0 - M_1\right|^2} & (r,s) = (0,0) \end{cases} . \qquad (7\text{-}15)$$

Therefore, for binary phase level DOEs, the percentage error of diffraction intensity for any off-axis diffraction order is dependent only on the etch depth error and not dependent on the phase distribution of the DOE. For a given etch depth error, this percentage error is the same for all the off-axis diffraction orders. On the other hand, unlike for the off-axis diffraction orders, the percentage error of diffraction intensity for the zeroth diffraction order is dependent on the phase distribution pattern of the DOE. When there is any etch depth error, the intensities of all off-axis reconstructed diffraction orders always decrease and the intensity of zeroth diffraction order always increases. The percentage error for all orders increases as the etch depth error increases.

The effect of etch depth error on reconstruction for binary phase level DOEs has been simulated for the DOEs designed for the 3×3 spot array generator and the 3×3 triangular connection pattern as mentioned in detail in Chapter 6. In each simulation, the diffraction efficiency, non-uniformity, minimum signal-to-noise ratio, diffraction efficiency of each desired diffraction order, percentage error of each desired diffraction order, and ratio of the diffraction efficiencies between each desired diffraction order and the zeroth diffraction order are plotted as a function of the percentage etch depth error. In these simulations for binary phase level DOEs, a maximum etch depth error of 10% is assumed. As expected, the diffraction efficiencies of all the orders (except for the zeroth

order) decrease by the same proportion when the etch depth error occurs and the additional light energy from all the non-zeroth orders goes to the zeroth diffraction order. Figure 7-1 and Figure 7-2 show the simulation results for the binary phase level DOEs. The results are the same as those of Figure 6-2 and 6-5, except the range and sense of the abscissa is different. The percentage error for the reconstructed diffraction intensity plots from those figures show consistent agreement with the theoretical results in Eq. (7-12) and Eq. (7-15).

When both the wavelength and the etch depth are incorrect, the reconstruction of the DOE will be affected by both factors. The phase delay can be shown to be

$$PD(p,q) = \pi \cdot PI(p,q) \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot [1 + r_1], \tag{7-16}$$

and the intensity of the reconstructed diffraction order can be shown to be

$$II(r,s,\lambda_1) = \begin{cases} S^2 \bullet |Q|^2 \bullet \left|1 - \exp\left\{j\pi\left(\frac{\lambda_0}{\lambda_1}\right)[1 + r_1]\right\}\right|^2 & (r,s) \neq (0,0) \\ S^2 \bullet \left|M_0 + M_1 \bullet \exp\left\{j\pi\left(\frac{\lambda_0}{\lambda_1}\right)[1 + r_1]\right\}\right|^2 & (r,s) = (0,0) \end{cases}. \tag{7-17}$$

The ratio of the reconstructed diffraction intensity for a given diffraction order between the DOE with etch depth error and incorrect wavelength, $II(r,s,\lambda_1)$, and the DOE with correct etch depth and wavelength, $I(r,s,\lambda_0)$, can be similarly found as

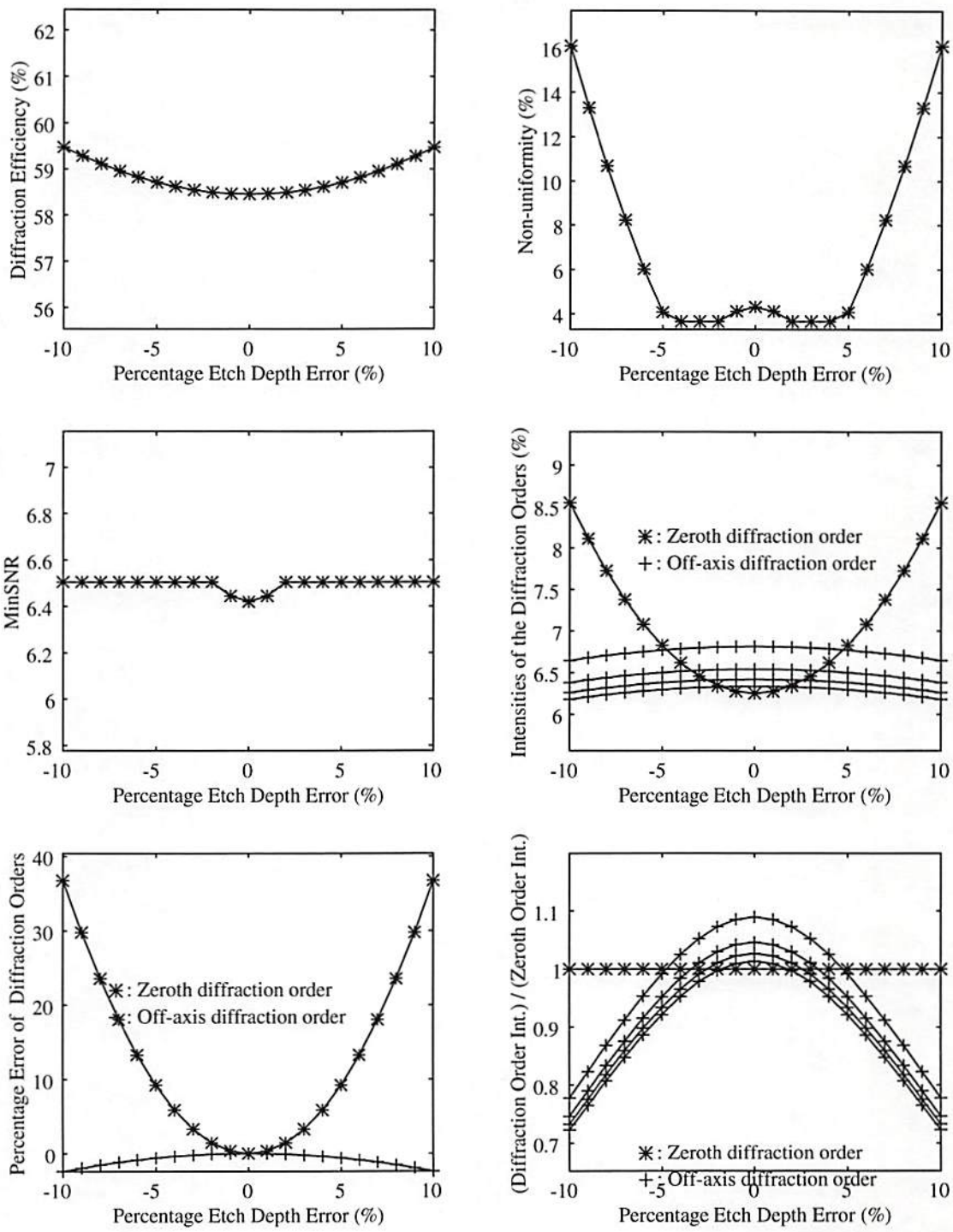

Figure 7-1: The effect of etch depth error on DOE reconstruction for the two-phase-level DOE designed for the spot array target pattern.

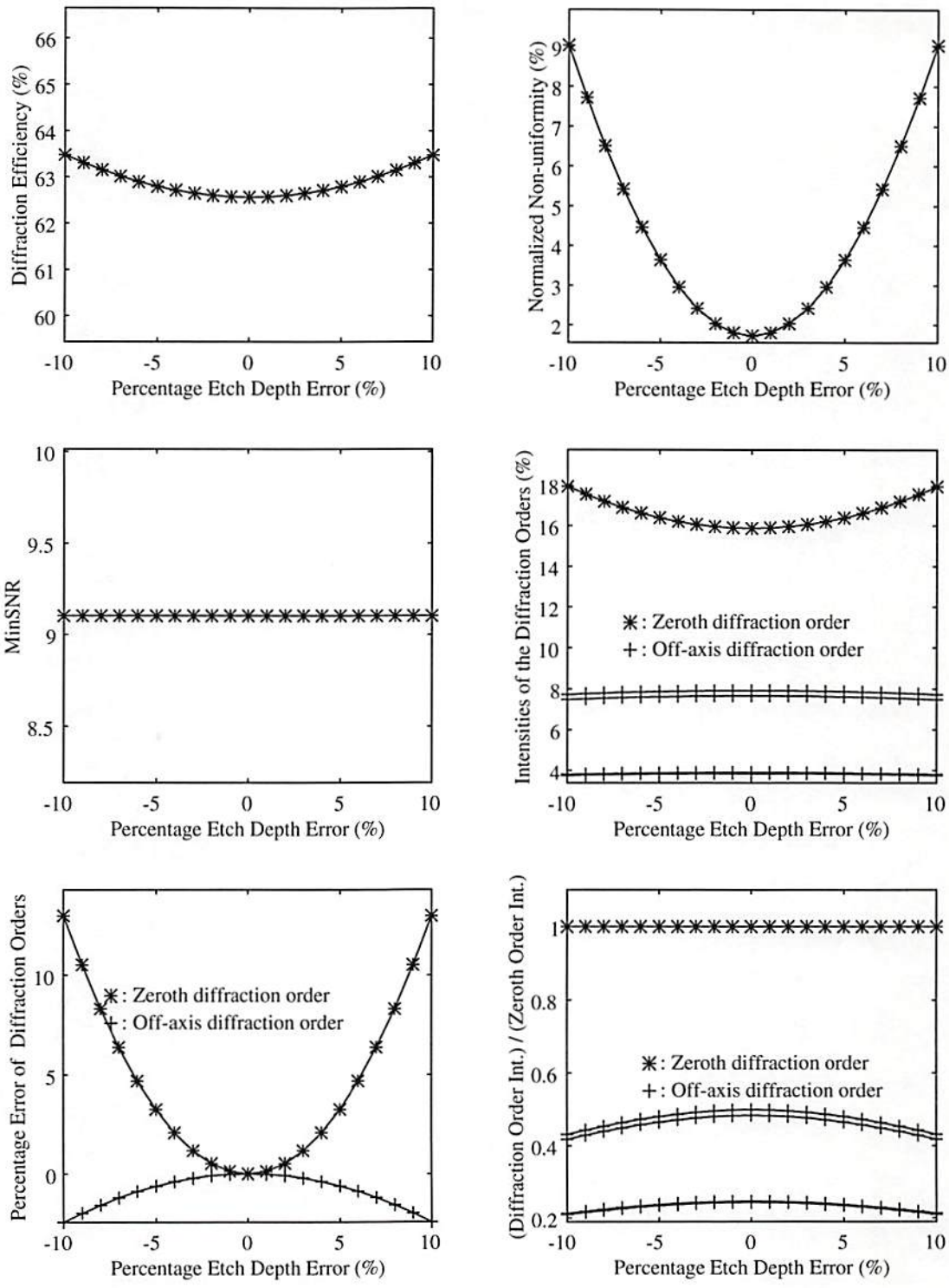

Figure 7-2: The effect of etch depth error on DOE reconstruction for the two-phase-level DOE designed for the triangular target pattern.

$$\frac{II(r,s,\lambda_1)}{I(r,s,\lambda_0)} = \begin{cases} \sin^2\left[\left(\frac{\pi}{2}\right)\left(\frac{\lambda_0}{\lambda_1}\right)(1+r_1)\right] & (r,s) \neq (0,0) \\ \dfrac{\left|M_0 + M_1 \bullet \exp\left\{j\pi\left(\frac{\lambda_0}{\lambda_1}\right)[1+r_1]\right\}\right|^2}{|M_0 - M_1|^2} & (r,s) = (0,0) \end{cases} . \quad (7\text{-}18)$$

From Eq. (7-18), we can also see that the intensity ratio between any two off-axis diffraction orders is a constant regardless the correctness of wavelength and etch depth. Let (r, s) and (r', s') represent two different off-axis diffraction orders, then we will have the following relation

$$\frac{II(r,s,\lambda_1)}{II(r',s',\lambda_1)} = \frac{I(r,s,\lambda_0)}{I(r',s',\lambda_0)} . \quad (7\text{-}19)$$

Hence, the intensity ratio of any two off-axis diffraction orders is independent of the illuminating wavelength and etch depth error for binary DOEs. This result for binary-phase-level DOEs can be used to help characterize the DOE fabrication processes.

The above results can also be used to measure the etch depth error in fabrication process. All we have to do is just use any convenient illumination wavelength to obtain a reconstruction pattern. Then measure the intensity of any off-axis diffraction order and calculate the etch depth error from Eq. (7-18). Furthermore, we can derive a way of cancelling the effect of etch depth error on the reconstruction of DOEs by using an appropriate illumination wavelength. Once the etch depth error has been measured, we then can just intentionally use an appropriate incorrect illumination wavelength to cancel

the effect of the etch depth error and retrieve the designed diffraction pattern. The illumination wavelength, λ_1, that can cancel the effect of the etch depth error must satisfy

$$\left(\frac{\lambda_0}{\lambda_1}\right) \cdot (1 + r_1) = 1 . \tag{7-20}$$

The above analysis for binary-phase-level DOEs assumes that the designed relative phase delay is either 0 or π. As mentioned in Chapter 6, it is possible to have a relative phase delay other than π (as detailed in Chapter 8). In that case, the expressions we have derived above need to be modified. Assuming that the relative phase delay for the DOE is either 0 or φ, then the phase delay, $PD(p, q)$, for the case of etch depth error and incorrect wavelength will be either 0 or $\left(\frac{\lambda_0}{\lambda_1}\right) \cdot (1 + r_1) \cdot \varphi$. Eq. (7-17) becomes

$$II(r, s, \lambda_1) = \begin{cases} S^2 \bullet |Q|^2 \bullet \left| 1 - \exp\left\{ j\varphi \left(\frac{\lambda_0}{\lambda_1}\right)[1 + r_1] \right\} \right|^2 & (r, s) \neq (0, 0) \\ S^2 \bullet \left| M_0 + M_1 \bullet \exp\left\{ j\varphi \left(\frac{\lambda_0}{\lambda_1}\right)[1 + r_1] \right\} \right|^2 & (r, s) = (0, 0) \end{cases}, \tag{7-21}$$

and Eq. (7-18) becomes

$$\frac{II(r, s, \lambda_1)}{I(r, s, \lambda_0)} = \begin{cases} \dfrac{\left| 1 - \exp\left[j\varphi \left(\frac{\lambda_0}{\lambda_1}\right)[1 + r_1] \right] \right|^2}{|1 - \exp[j\varphi]|^2} & (r, s) \neq (0, 0) \\ \dfrac{\left| M_0 + M_1 \bullet \exp\left[j\varphi \left(\frac{\lambda_0}{\lambda_1}\right)[1 + r_1] \right] \right|^2}{|M_0 + M_1 \cdot \exp[j\varphi]|^2} & (r, s) = (0, 0) \end{cases} . \tag{7-22}$$

The relation between the new diffraction intensity and the originally designed diffraction intensity for a given off-axis order is still independent of the phase distribution pattern of the DOE and is independent of the phase distribution of the DOE and dependent only on the wavelength change and etch depth error. The ratio between $II(r, s, \lambda_1)$ and $I(r, s, \lambda_0)$ is still the same for all the off-axis diffraction orders. This means that if the illuminating wavelength is deviating from the designed wavelength and/or there is an etch depth error, then the diffraction efficiencies of all off-axis diffraction orders will be changed by the same percentage. It is also interesting to note that, from Eq. (7-22) that $II(r, s, \lambda_1)$ for this non-conventional binary-phase-level DOE is not always less than or equal to $I(r, s, \lambda_0)$ as is the case for the conventional binary-phase-level DOE. Therefore, it is possible that the intensity of the zeroth diffraction order will be decreased when there is a wavelength shift and/or etch depth error. This is quite different from the case for the conventional binary-phase-level DOE.

The above results can also be used to measure the etch depth error due to the fabrication process as well as to cancel the effect of etch depth error on the reconstruction for unconventional binary-phase-level DOEs by using an appropriate illumination wavelength. The illumination wavelength, λ_1, that can cancel the effect of the etch depth error still needs satisfy Eq. (7-20).

7.3 Constant Percentage Etch Depth Error in Etch Steps for Multiple-Phase-Level Diffractive Optical Elements

For multiple phase level DOEs, more than one etch steps is needed. If the percentage etch depth error is the same in every etch step, the phase delay in Eq. (7-8) can be simplified to

$$\begin{aligned} & PD(p,q) \\ &= \frac{2\pi}{Z} \sum_{i=1}^{\log_2 Z} [1 - e_i(p,q)] \cdot [1 + r_i] \cdot 2^{i-1} \\ &= \frac{2\pi}{Z} \cdot (1 + r_{cp}) \cdot \sum_{i=1}^{\log_2 Z} \left[2^{i-1} - 2^{i-1} \cdot e_i(p,q) \right] \quad , \\ &= \frac{2\pi}{Z} \cdot (1 + r_{cp}) \cdot \{ [Z-1] - [(Z-1) - p(p,q)] \} \\ &= \frac{2\pi}{Z} \cdot (1 + r_{cp}) \cdot PI(p,q) \end{aligned} \tag{7-23}$$

where r_{cp} represents the constant percentage etch depth error. The reconstructed diffraction intensity pattern, $II(r,s)$, in Eq. (7-10) can then be simplified to

$$\begin{aligned} & II(r,s) \\ &= \left| J^{-2} \operatorname{sinc}\left(\frac{r}{J}\right) \operatorname{sinc}\left(\frac{s}{J}\right) DFT\{ \underline{h}(p,q) \} \right|^2 \\ &= \left| J^{-2} \operatorname{sinc}\left(\frac{r}{J}\right) \operatorname{sinc}\left(\frac{s}{J}\right) DFT\left\{ \exp\left[j\left(\frac{2\pi}{Z}\right) \cdot (1 + r_{cp}) \cdot PI(p,q) \right] \right\} \right|^2 \end{aligned} \tag{7-24}$$

Comparing the results obtained from Section 6.1 with Eq. (7-23) and Eq. (7-24), it can be seen that the case of a constant percentage etch depth error in every etch step is essentially equivalent to a wavelength change.

The effect of constant percentage etch depth error in every etch step for four and eight-phase-level DOEs has been simulated for the DOEs designed for the 3×3 spot array generator and the 3×3 triangular connection pattern as mentioned in detail in Chapter 6. In each simulation, the diffraction efficiency, Non-uniformity, minimum signal-to-noise ratio, diffraction efficiency of each desired diffraction order, percentage error of each desired diffraction order, and ratio of the diffraction efficiencies between each desired diffraction order and the zeroth diffraction order are plotted as a function of percentage etch depth error. In these simulations, a maximum etch depth error of 10% is assumed. Figure 7-3, 7-4, 7-5, and 7-6 show the simulation results. The results are the same as those of Figure 6-3, 6-4, 6-6, and 6-7, shown as a function of illumination wavelength, except here the range and sense of the abscissa is different.

From the simulations, we can see that the reconstructed intensity of each diffraction order deviates from its designed value when there is etch depth error. This deviation increases as the etch depth error increases. For a given diffraction order, the effects of the etch depth error on the reconstruction for multiple-phase level DOEs are dependent on both the phase distribution pattern of the DOE and the amount of the etch depth error. For a given etch depth error, this percentage reconstruction error also varies for different diffraction orders. The reconstructed intensity of the zeroth diffraction order could be lower than the designed value. This is different from the case for binary phase level

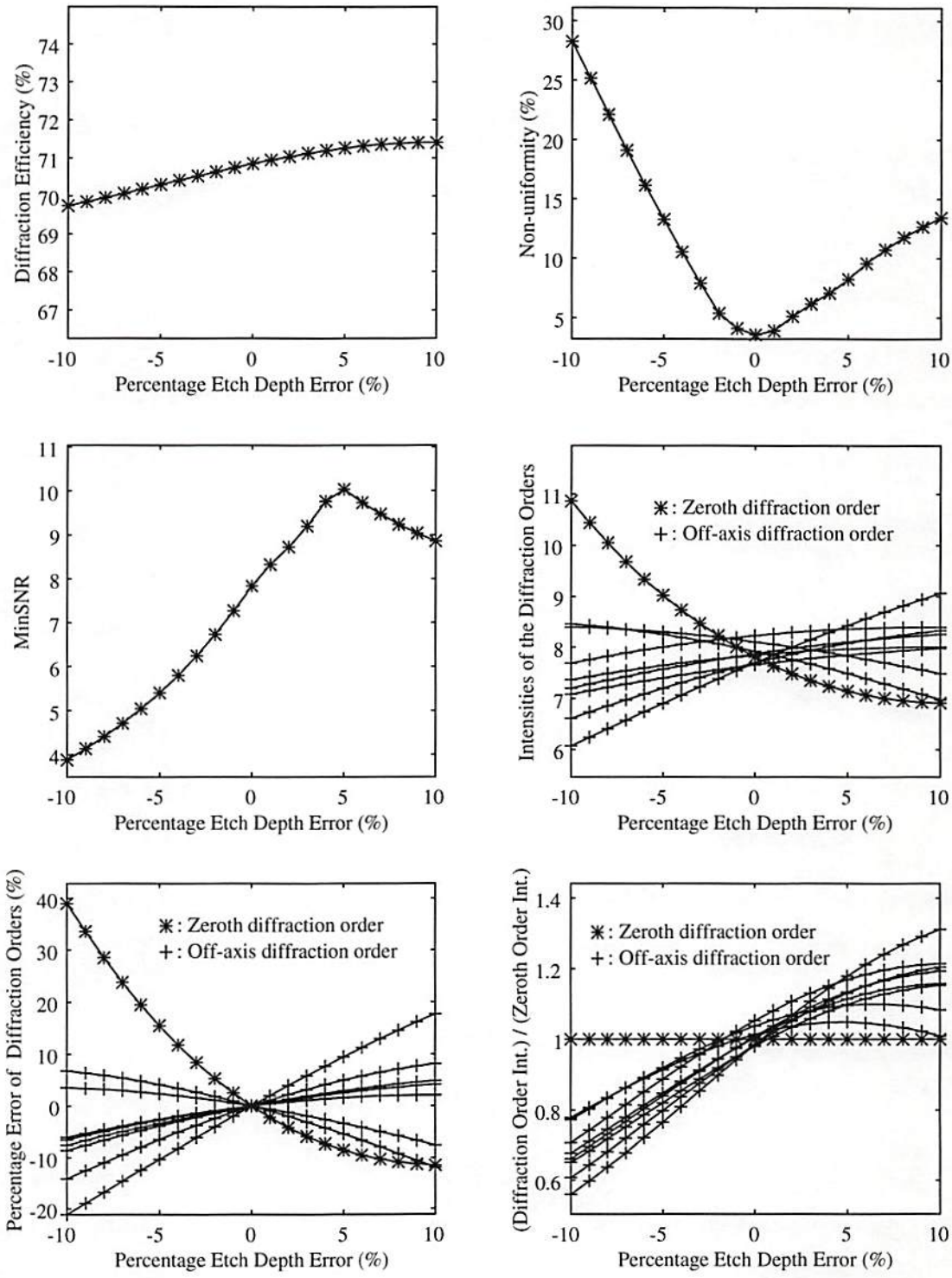

Figure 7-3: The effect of constant percentage etch depth error on DOE reconstruction for the four-phase-level DOE designed for the spot array target pattern.

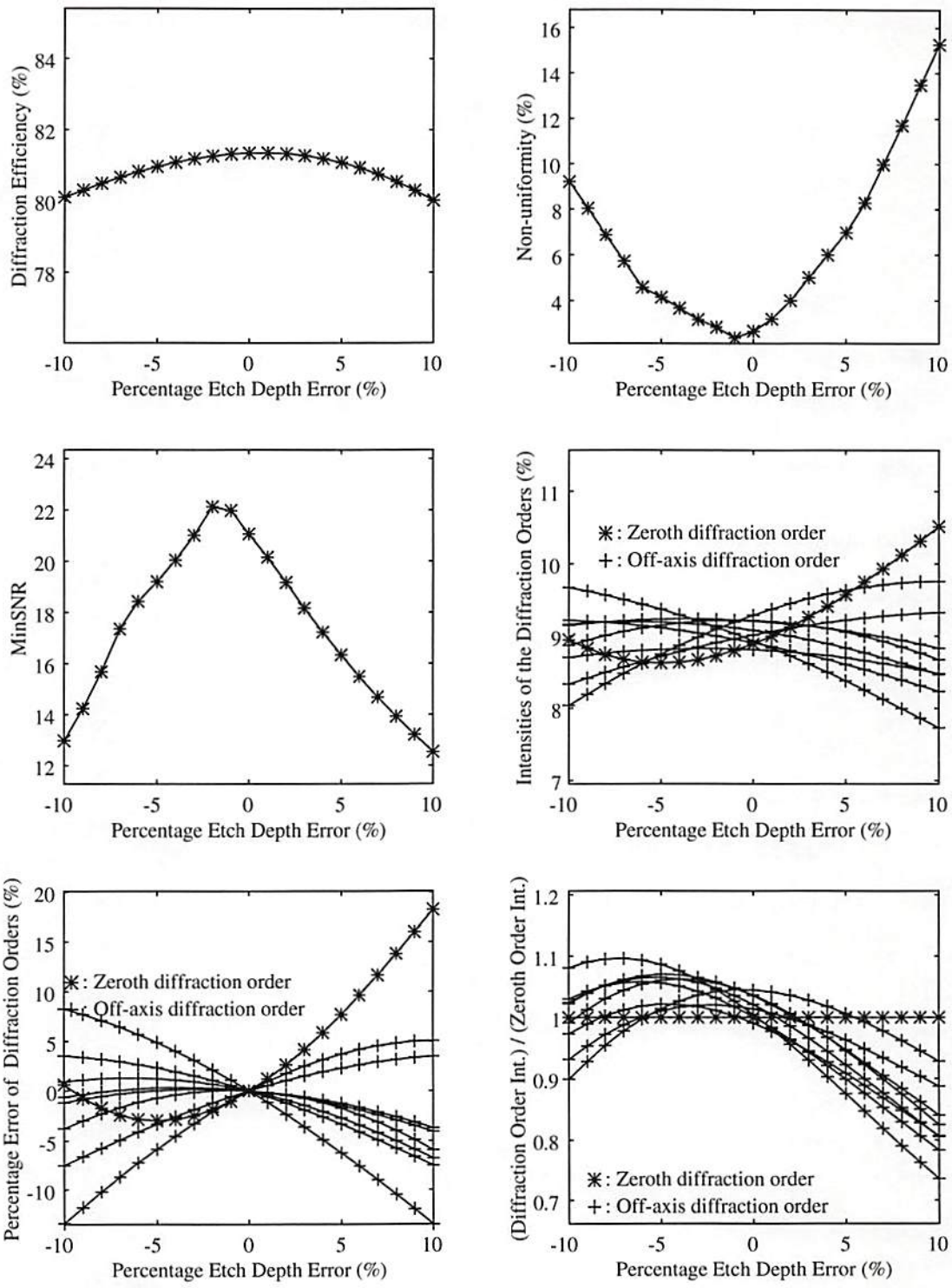

Figure 7-4: The effect of constant percentage etch depth error on DOE reconstruction for the eight-phase-level DOE designed for the spot array target pattern.

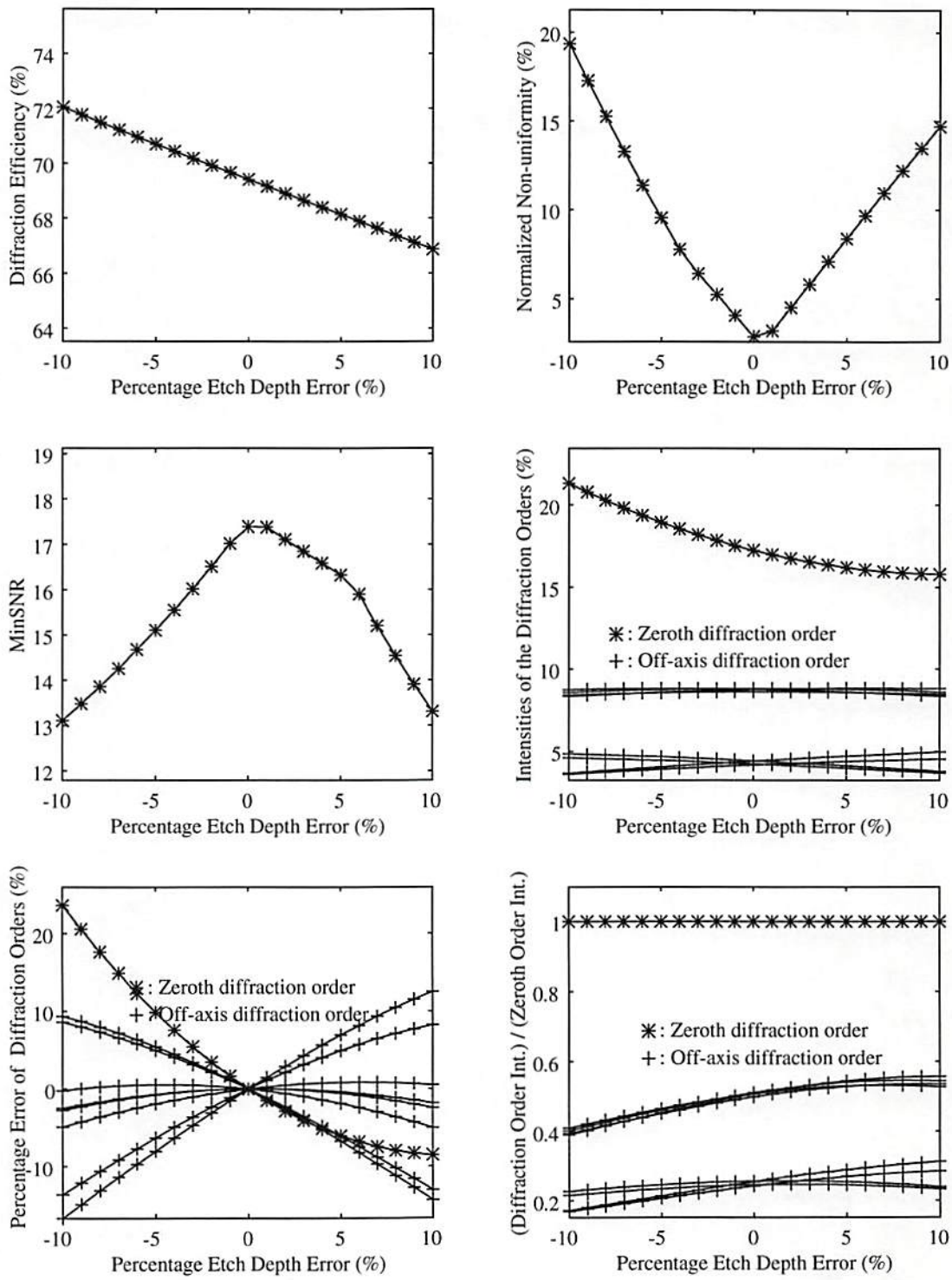

Figure 7-5: The effect of constant percentage etch depth error on DOE reconstruction for the four-phase-level DOE designed for the triangular connection pattern.

DOEs, in which the reconstructed intensity of the zeroth diffraction order is always higher than the designed value.

When both the wavelength and the etch depth are incorrect, the reconstruction of the DOE will be affected by both factors. The phase delay can be shown to be

$$PD(p,q) = \frac{2\pi}{Z} \cdot PI(p,q) \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot [1 + r_{cp}], \tag{7-25}$$

and the intensity of the reconstructed diffraction order can be shown to be

$$\begin{aligned} & II(r,s,\lambda_1) \\ &= \left| J^{-2} \operatorname{sinc}\left(\frac{r}{J}\right) \operatorname{sinc}\left(\frac{s}{J}\right) DFT\{\underline{h}(p,q)\} \right|^2 \\ &= \left| J^{-2} \operatorname{sinc}\left(\frac{r}{J}\right) \operatorname{sinc}\left(\frac{s}{J}\right) DFT\left\{ \exp\left[j\left(\frac{2\pi}{Z}\right) \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot (1 + r_{cp}) \cdot PI(p,q) \right] \right\} \right|^2 \end{aligned} \tag{7-26}$$

By using Eq. (7-26), we can measure the constant percentage etch depth error of the fabrication process and cancel the effect of the etch depth error on the reconstruction of the DOE by using an appropriate illumination wavelength as mentioned in detail in the last section. The illumination wavelength, λ_1, that can cancel the effect of the etch depth error must satisfy Eq. (7-20).

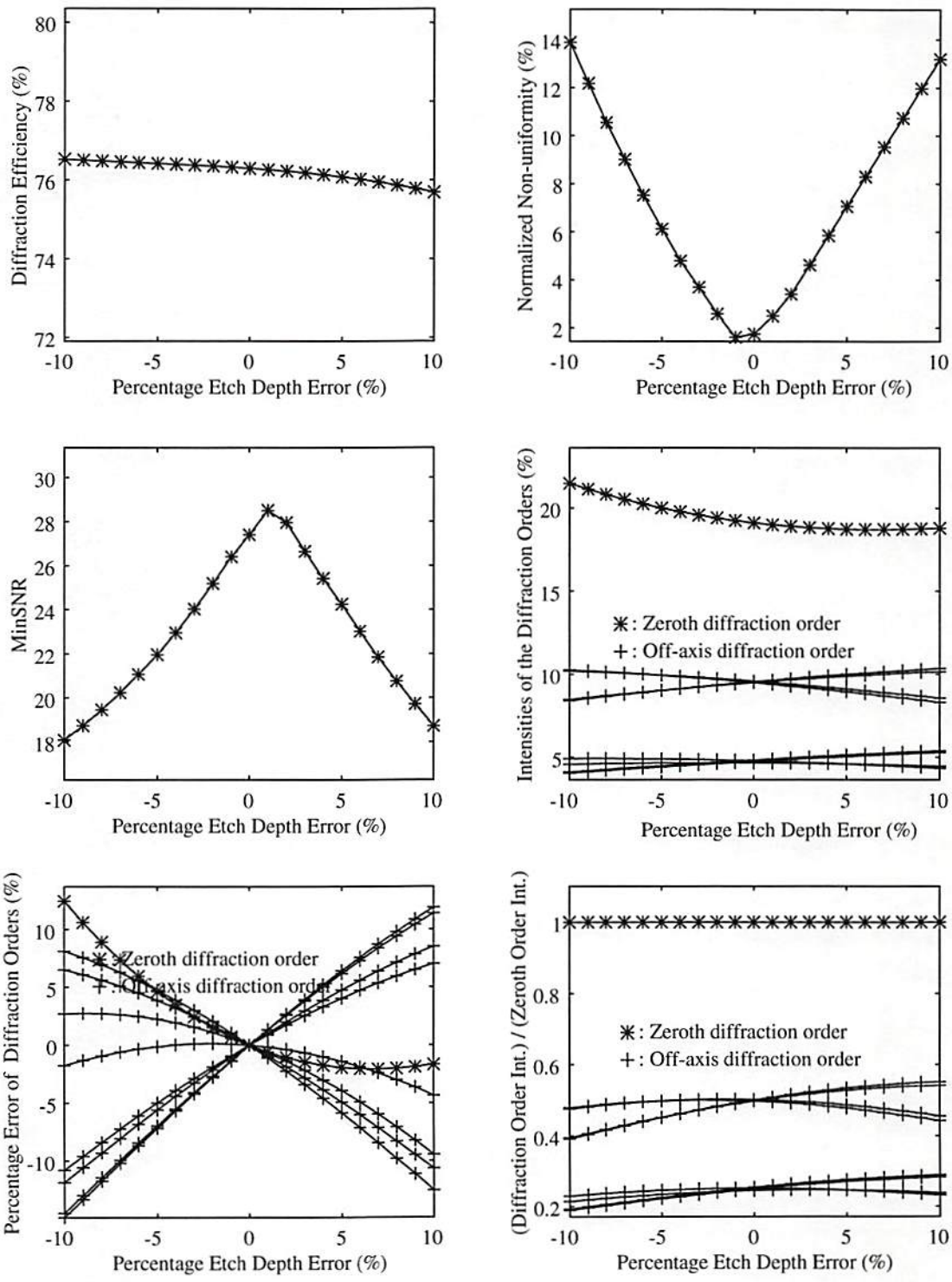

Figure 7-6: The effect of constant percentage etch depth error on DOE reconstruction for the eight-phase-level DOE designed for the triangular connection pattern.

7.4 Reconstruction of Four-Phase-Level Diffractive Optical Elements for Independent Etch Depth Error in Etch Steps

In the most general case, the etch depth error in each etch step is independent of the error in each of the other etch steps. Using Eq. (7-9), each possible phase level for $PI(p, q)$, the values and binary representation for $[(Z-1) - PI(p, q)]$, and the corresponding phase delay for four-phase level DOEs are listed in Table 7-1.

Table 7-1: Phase level index and phase delay calculation for four-phase level DOEs for independent etch depth errors.

$PI(p, q)$	$[(Z-1) - PI(p, q)]$	Binary representation	Phase delay
0	3	11	0
1	2	10	$\frac{\pi}{2} \cdot (1 + r_1)$
2	1	01	$\frac{\pi}{2} \cdot 2 \cdot (1 + r_2)$
3	0	00	$\frac{\pi}{2} \cdot [(1 + r_1) + 2 \cdot (1 + r_2)]$

Note that when $r_1 = -2r_2$, the etch depth errors for phase elements with phase level index of 3 will cancel and the corresponding phase delay will be correct although the phase delays for elements with phase level index of 1 or 2 are still incorrect.

The effect of independent etch depth error in the etch steps for four-phase-level DOEs has been simulated for the DOEs designed for the 3×3 spot array generator and the 3×3 triangular connection pattern as mentioned in detail in Chapter 6. In each

simulation, the diffraction efficiency, Non-uniformity, minimum signal-to-noise ratio, diffraction efficiency of each desired diffraction order, percentage error of each desired diffraction order, and ratio of the diffraction efficiencies between each desired diffraction order and the zeroth diffraction order are plotted as a function of percentage etch depth error. In these simulations, we assumed that the etch depth error in each etch step is independent of each other, the etch depth error of the most significant bit of the etch masks (r_2) is -10%, -5%, 5%, 10%, and $-|r_2| \le r_1 \le |r_2|$. Figure 7-7 to Figure 7-14 show the simulation results. In those figures, r_1 refers to the LSB (least significant bit) and r_2 refers to the MSB (most significant bit). Note that in Figure 7-7 to Figure 7-14, the points on the plots that correspond to a zero LSB etch depth error are not etch-error-free cases. To find the DOE reconstruction for the etch-depth-free cases, the reader should refer to Figure 6-3 and 6-6.

From the simulations above, we can see that, if the etch depth error in etch steps are independent, then the effect of etch depth error is not equivalent to the effect due to illumination wavelength changes. The percentage error of a given diffraction order is dependent on the etch depth error in all steps and the phase distribution of the DOE. For a given set of etch depth errors in the DOE, this percentage reconstruction error also varies for different diffraction orders. The reconstructed intensity of the zeroth diffraction order could be lower than the designed value. This is different from the case for binary phase level DOEs, in which case the reconstructed intensity of the zeroth diffraction order is always higher than the designed value.

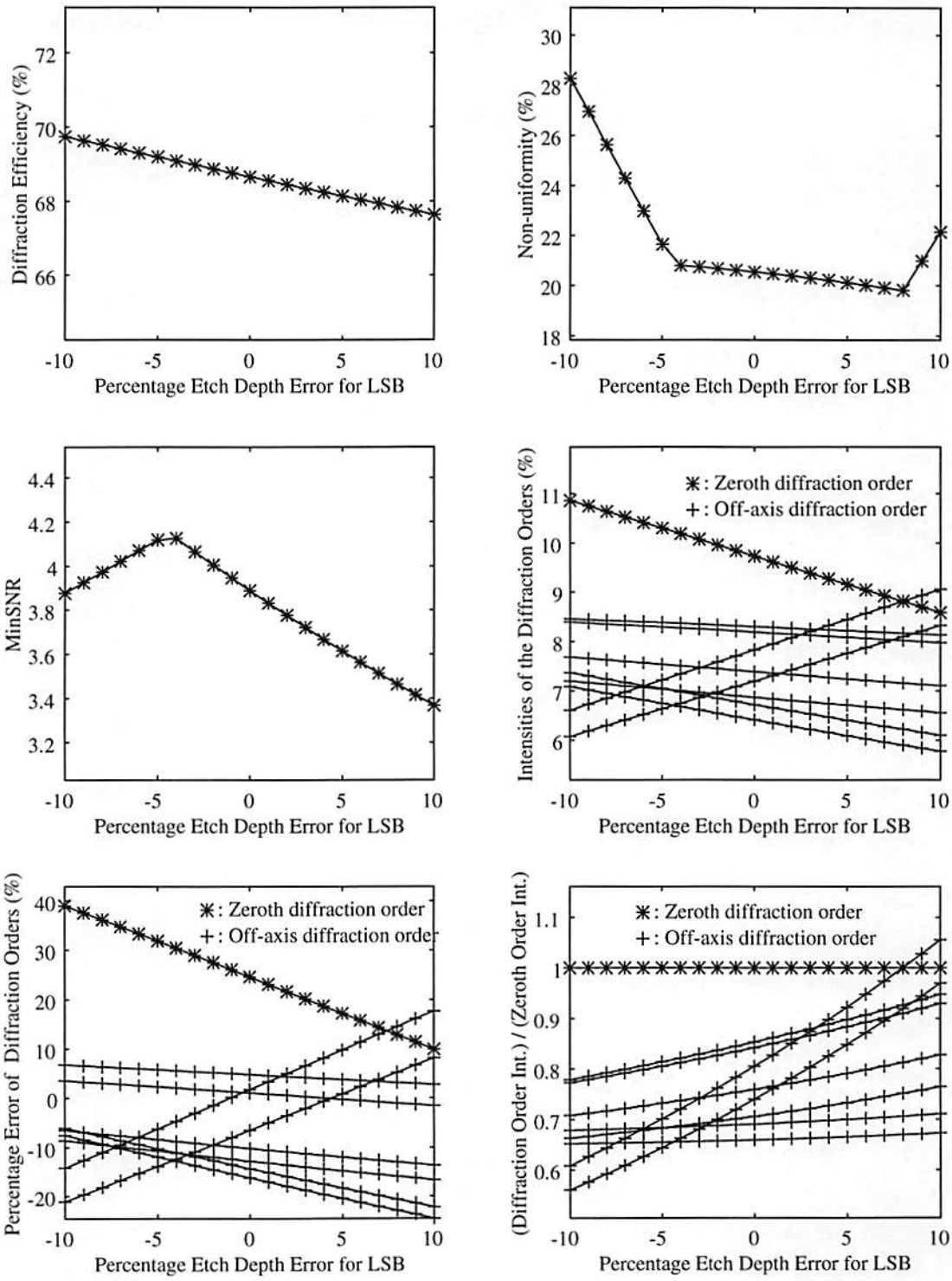

Figure 7-7: The effect of independent etch depth error in etch steps (-10% MSB) on DOE reconstruction for the four-phase-level DOE designed for the spot array pattern.

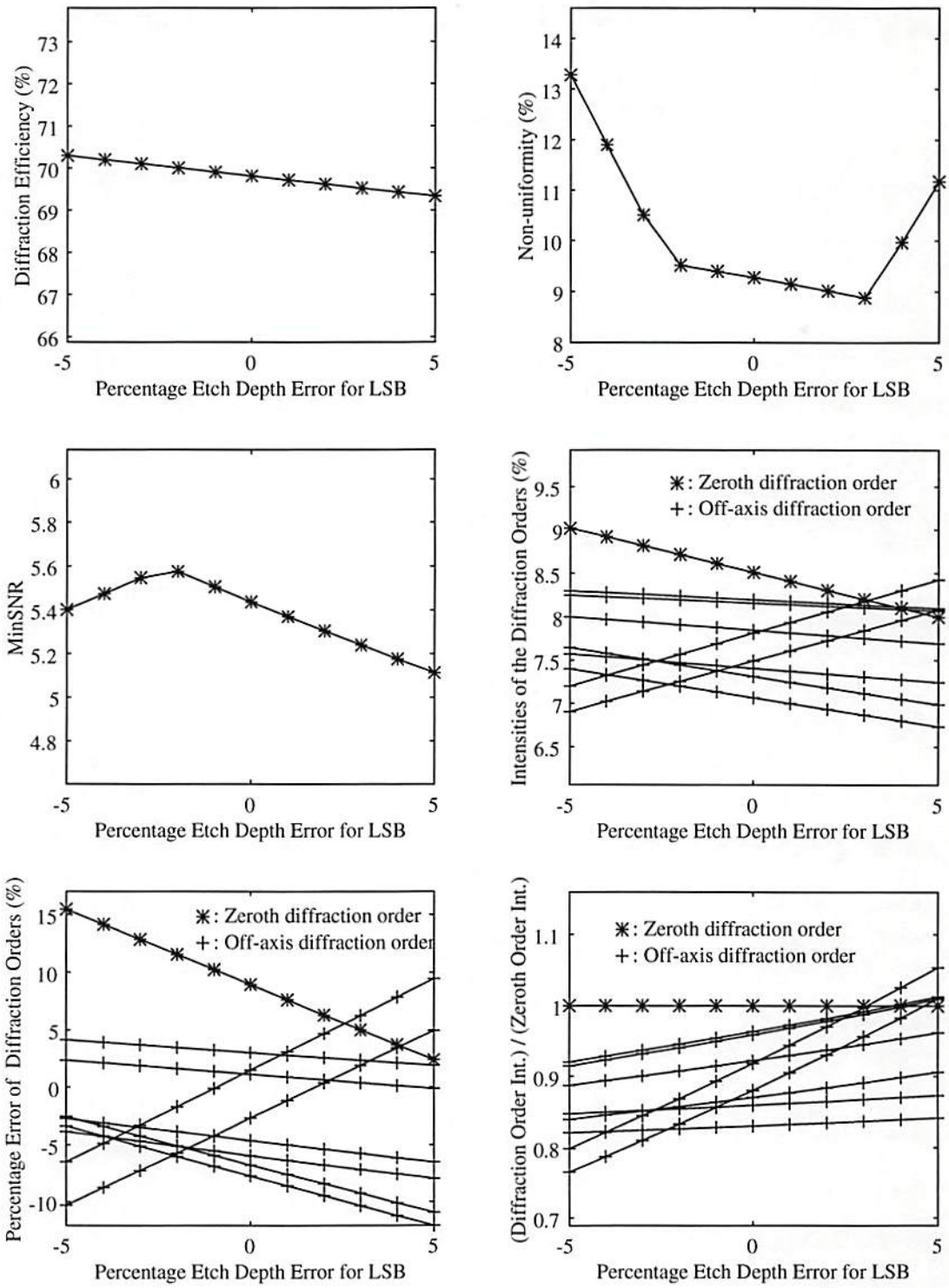

Figure 7-8: The effect of independent etch depth error in etch steps (-5% MSB) on DOE reconstruction for the four-phase-level DOE designed for the spot array pattern.

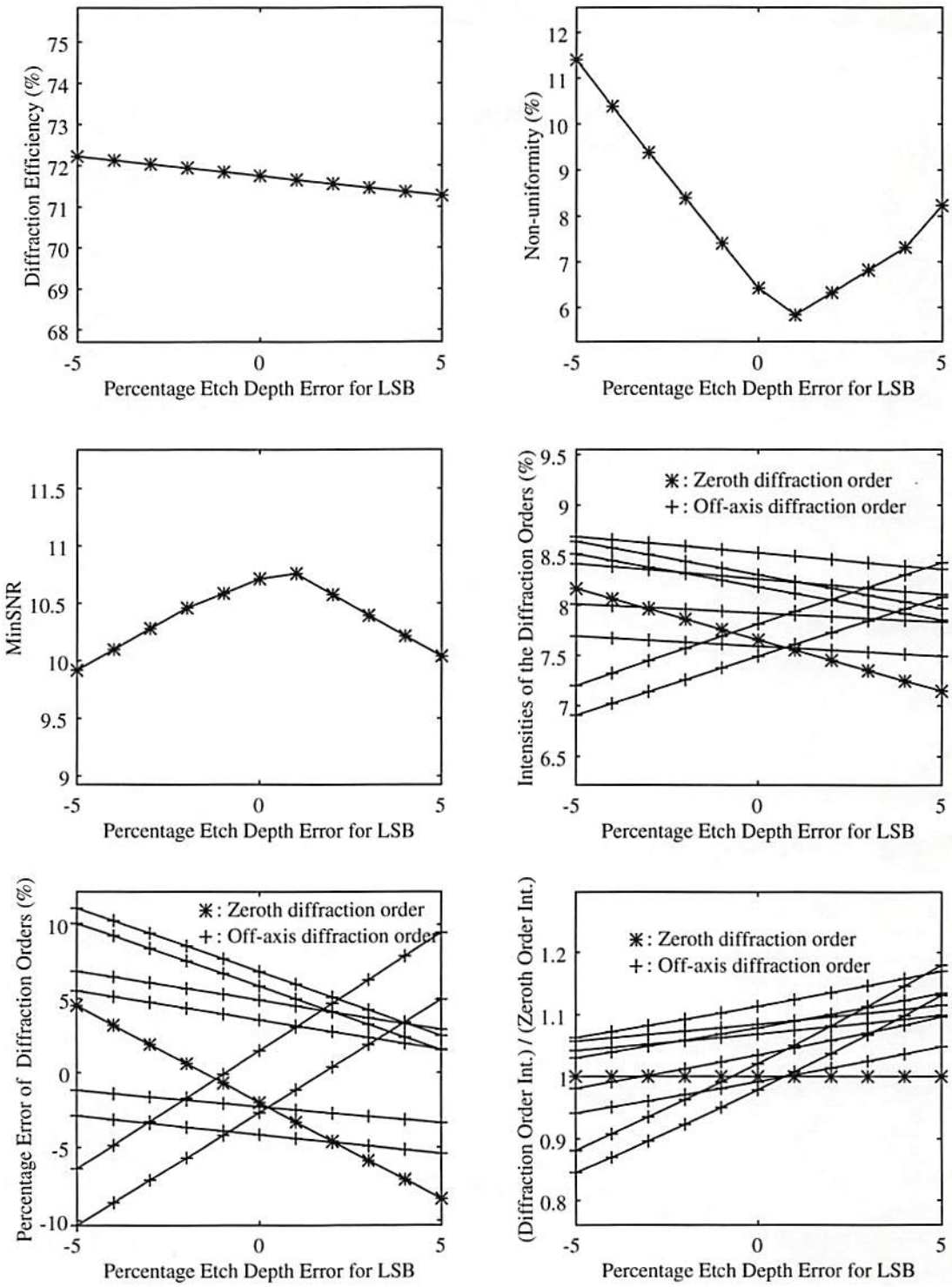

Figure 7-9: The effect of independent etch depth error in etch steps (5% MSB) on DOE reconstruction for the four-phase-level DOE designed for the spot array pattern.

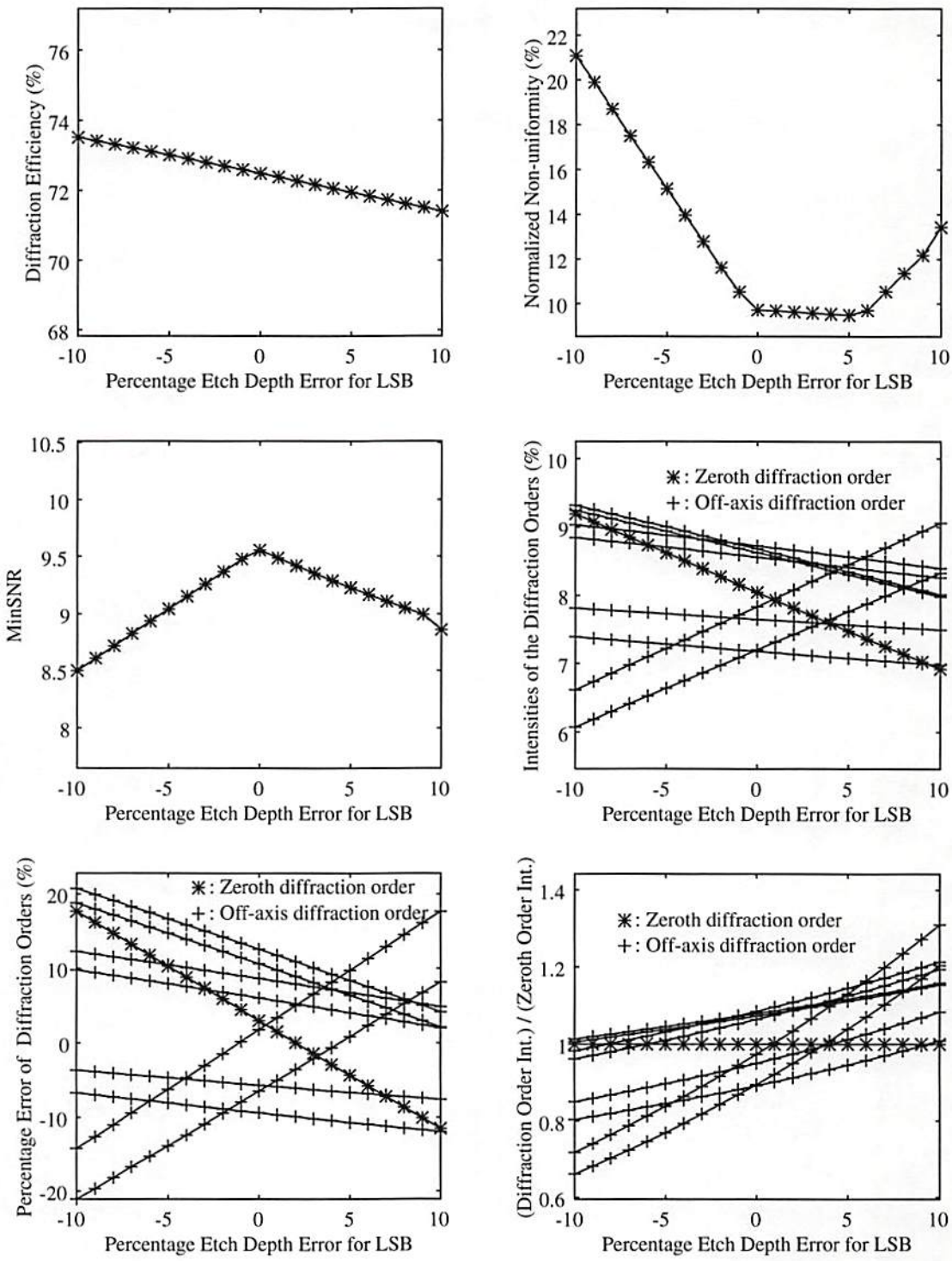

Figure 7-10: The effect of independent etch depth error in etch steps (10% MSB) on DOE reconstruction for the four-phase-level DOE designed for the spot array pattern.

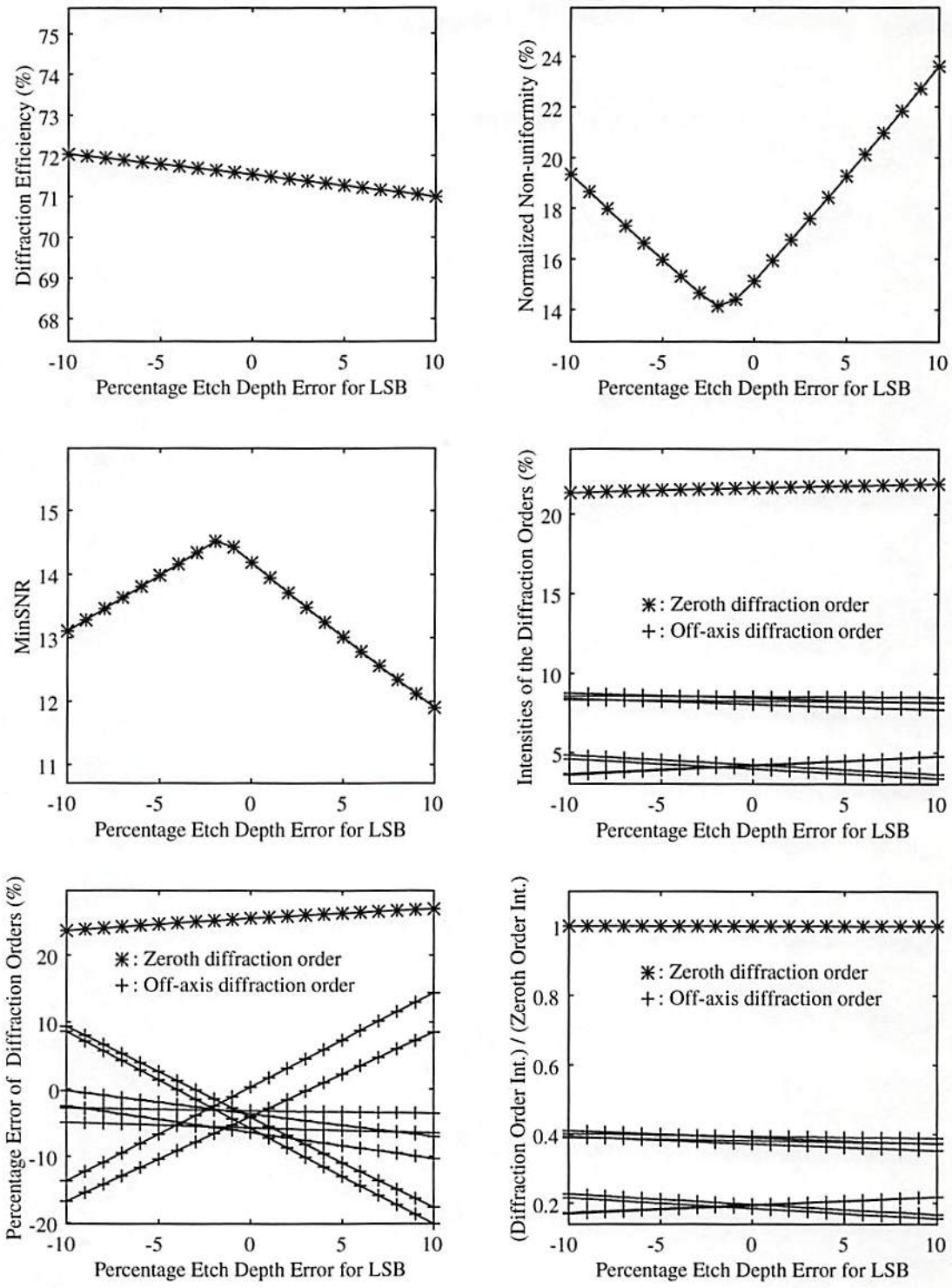

Figure 7-11: The effect of independent etch depth error in etch steps (-10% MSB) on DOE reconstruction for the four-phase-level DOE designed for the triangular pattern.

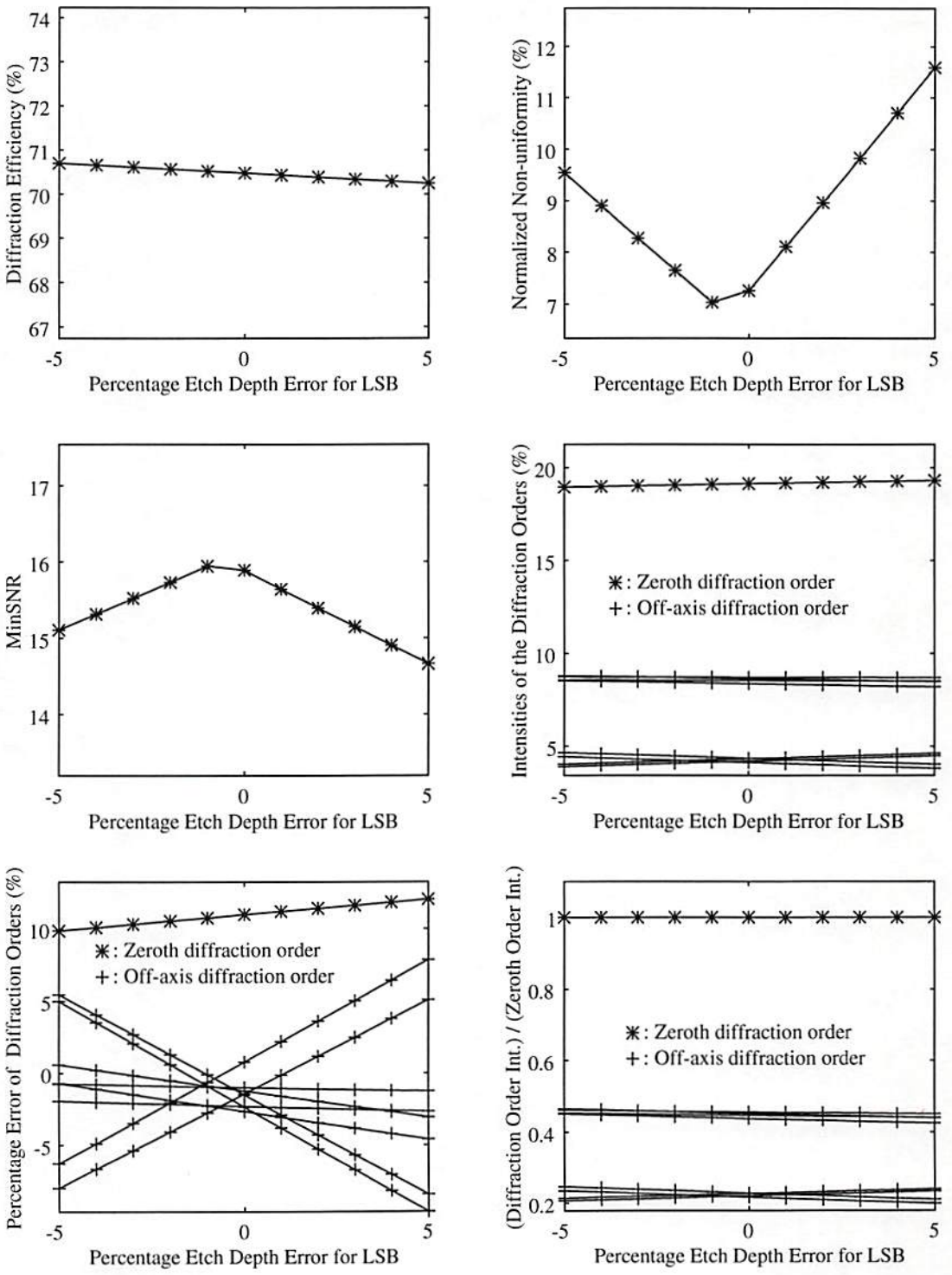

Figure 7-12: The effect of independent etch depth error in etch steps (-5% MSB) on DOE reconstruction for the four-phase-level DOE designed for the triangular pattern.

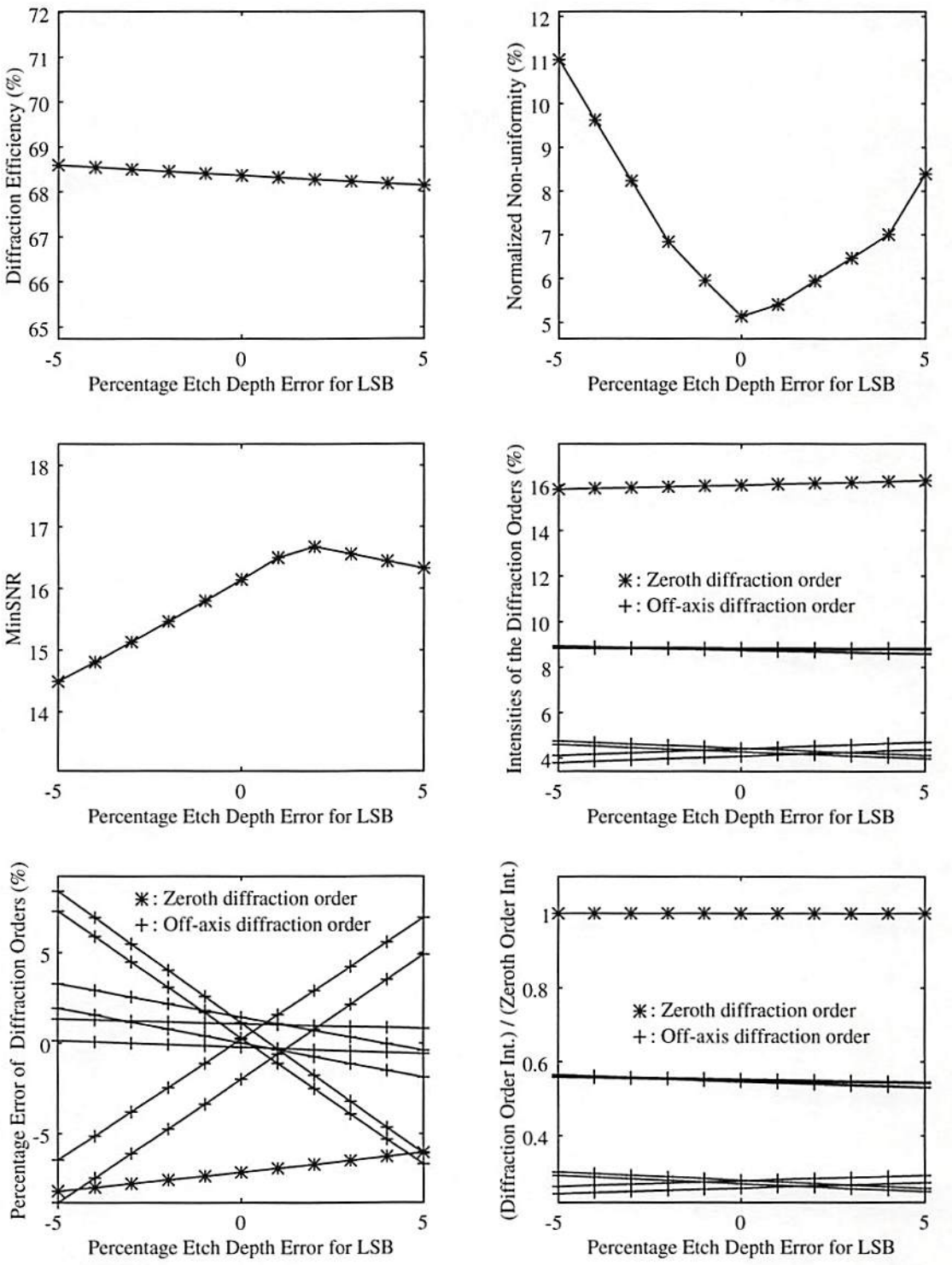

Figure 7-13: The effect of independent etch depth error in etch steps (5% MSB) on DOE reconstruction for the four-phase-level DOE designed for the triangular pattern.

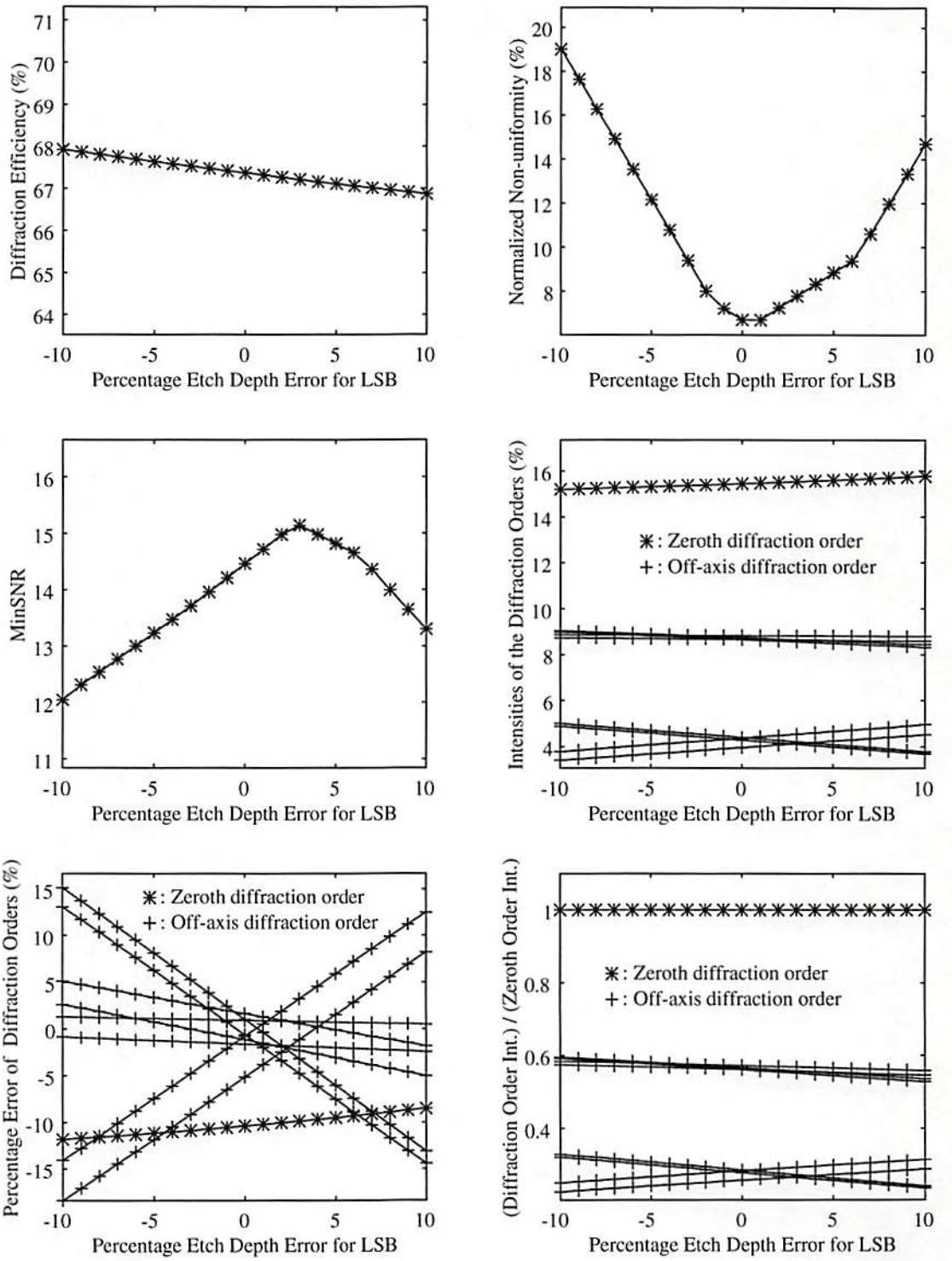

Figure 7-14: The effect of independent etch depth error in etch steps (10% MSB) on DOE reconstruction for the four-phase-level DOE designed for the triangular pattern.

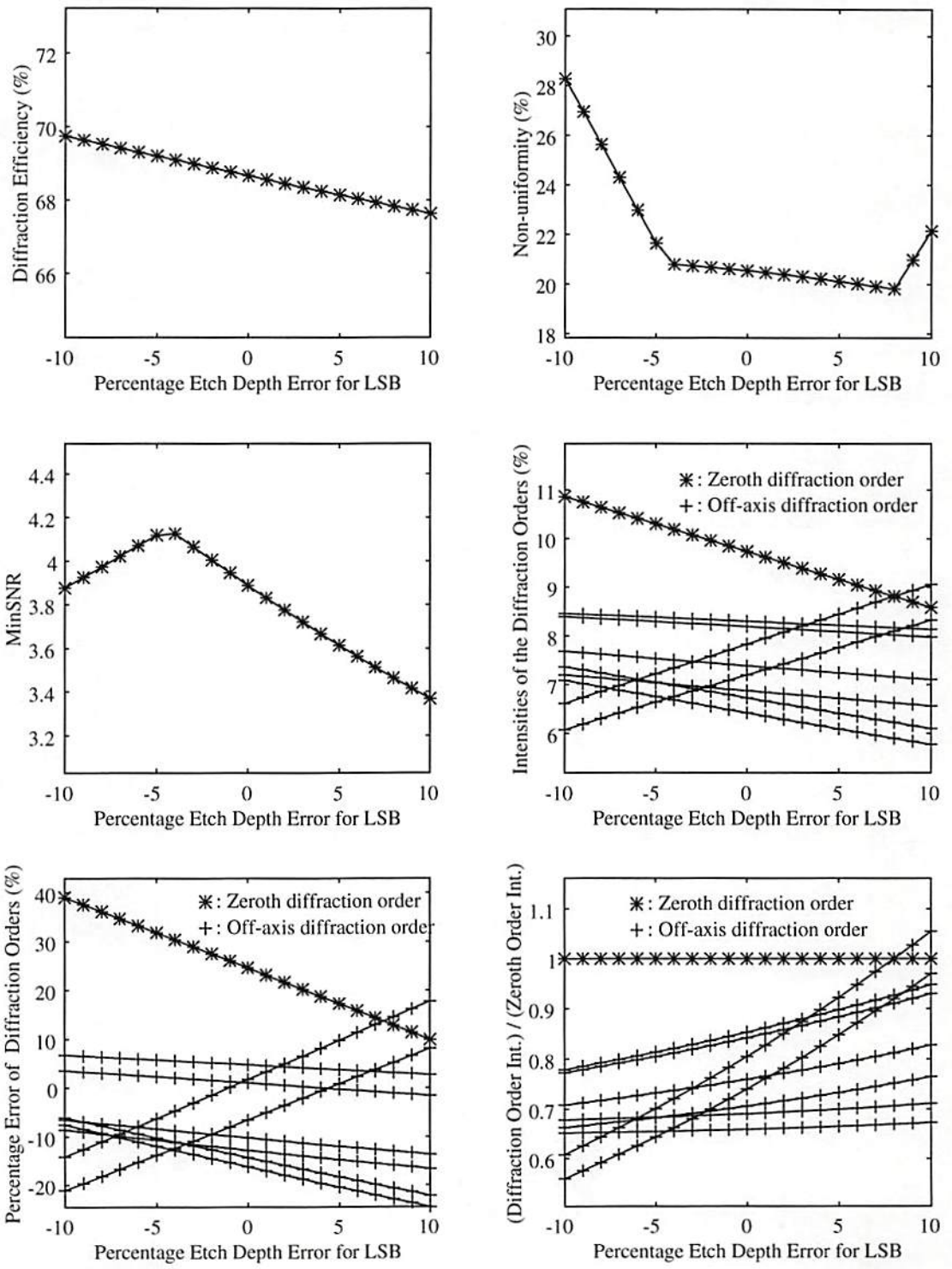

Figure 7-15: The effect of independent etch depth error in etch steps (-10% MSB) on DOE reconstruction for the four-phase-level DOE designed for the spot array pattern.

When both the wavelength and the etch depth are incorrect, the reconstruction of the DOE will be affected by both factors. The phase delay for this case can be shown to be

$$PD(p,q) = \left(\frac{2\pi}{Z}\right) \cdot \left(\frac{\lambda_0}{\lambda_1}\right) \cdot \sum_{i=1}^{\log_2 Z} [1 - e_i(p,q)] \cdot [1 + r_i] \cdot 2^{i-1}, \qquad (7\text{-}27)$$

and the intensity of the reconstructed diffraction order can be shown to be

$$II(r,s) = \left| J^{-2} \operatorname{sinc}\left(\frac{r}{J}\right) \operatorname{sinc}\left(\frac{s}{J}\right) DFT\{\exp[j \cdot PD(p,q)]\} \right|^2. \qquad (7\text{-}28)$$

Chapter 8

New Design Approaches for Diffractive Optical Elements

In the 3-D computation structure described in the previous chapters, the DOE has to reconstruct an interconnection pattern that is as close as possible to the desired one. Hence, in this chapter, we will compare the DOE performance of different design algorithms and develop new design approaches that can increase the performance of DOEs. First, we discuss a modified Gerchberg-Saxton algorithm with quantization scheduling and noise-adjusting scheduling in Section 8.1. Using this modified approach, we designed the spot array generator DOE and the triangular interconnection pattern DOE, then compared the results with the ones designed using the original Gerchberg-Saxton algorithm. In Section 8.2, we discuss the simulated annealing algorithm, effects of cost functions in the algorithm, and a dual-cost-function simulated annealing algorithm. We compare the performance of the DOEs designed using this dual-cost-function simulated annealing algorithm with the performance of the DOEs designed using the modified Gerchberg-Saxton algorithm. In Section 8.3, we describe a new design approach that can further increase the performance of DOEs by optimizing the phase values without increasing the binary-optics fabrication complexity, and compare the performance of the DOEs designed

using this new approach with the performance of the DOEs designed using the modified Gerchberg-Saxton algorithm.

8.1 Original and Modified Gerchberg-Saxton Algorithms

The Gerchberg-Saxton algorithm [45] is an iterative Fourier transform phase-retrieval procedure that tries to satisfy the required constraints on both the DOE plane and reconstruction plane. These constraints are: specification of intensity transmittance in the DOE plane (as unity) and specification of signal intensity pattern in the diffraction (or reconstruction) plane (as the desired interconnection weights). Figure 8-1 shows the

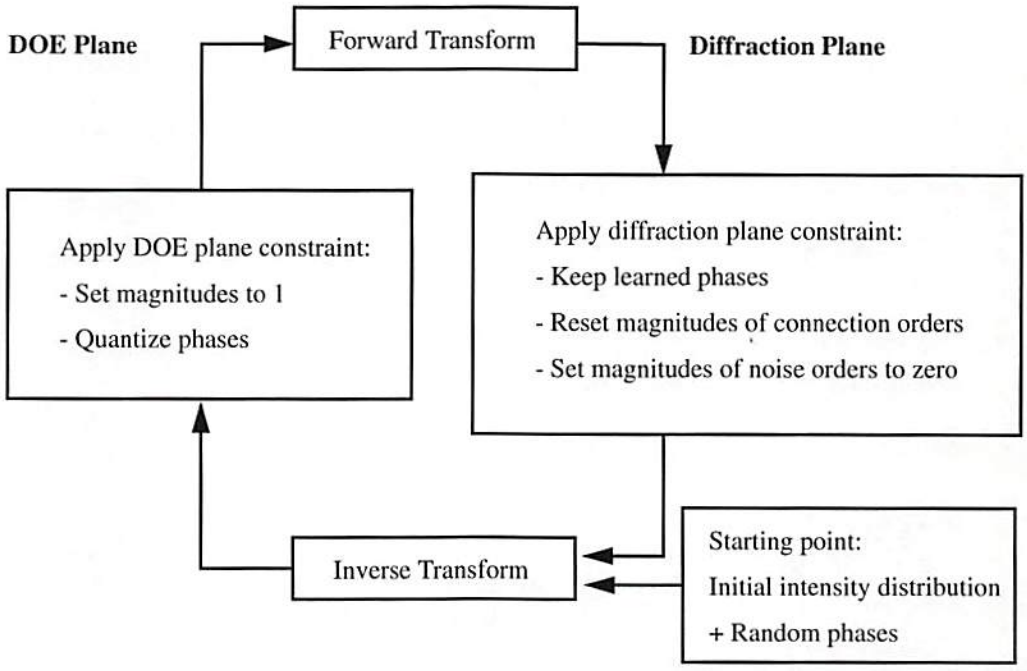

Figure 8-1: Schematic diagram of the original Gerchberg-Saxton algorithm.

schematic flow diagram of the original Gerchberg-Saxton algorithm. Recall from Chapter 3 that the DOE plane and the diffraction plane are related by

$$\underline{A}(r,s) = J^{-2}\operatorname{sinc}\left(\frac{r}{J}\right)\operatorname{sinc}\left(\frac{s}{J}\right)DFT\{\underline{h}(p,q)\}, \tag{8-1}$$

and

$$\underline{h}(p,q) = DFT^{-1}\left\{\frac{J^2\underline{A}(r,s)}{\operatorname{sinc}\left(\frac{r}{J}\right)\operatorname{sinc}\left(\frac{s}{J}\right)}\right\}, \tag{8-2}$$

where J is the number of phase elements in one period of the DOE in each dimension (it is also the number of independent diffraction orders in each dimension), $\underline{A}(r,s)$ is the complex amplitude of each diffraction order in the diffraction plane (r and s are the indices of the diffraction orders with $r, s \in \{-\frac{J}{2}, \ldots, 0, \ldots, \frac{J}{2}-1\}$) and $\underline{h}(p,q)$ is the complex amplitude transmittance distribution in one period of the DOE (p and q are the indices of the phase elements with $p, q \in \{0, 1, \ldots, J-1\}$). The design process begins with the desired intensity pattern and a set of random phases (one for each diffraction order) in the reconstruction plane. It then uses Eq. (8-2) to compute the complex amplitude transmittance of the DOE grating. Since a DOE is a phase-only element, the magnitude of the obtained complex amplitude transmittance is set to one. Also, since the phase of a DOE is discrete, the phase of the obtained complex amplitude transmittance is quantized to the set of pre-defined discrete levels. Afterward, we then use Eq. (8-1) to compute the complex amplitude of each diffraction order and the corresponding intensities. In the reconstruction plane, the learned phases of each diffraction order from the iteration are kept, the magnitudes of the desired connection orders are reset to the desired values, and the

magnitudes of all other orders are reset to zero. This process then continues iterating until a stable DOE phase pattern has been obtained. Usually, the Gerchberg-Saxton algorithm can converge to a solution in few hundred iterations.

One common known problem of using this original Gerchberg-Saxton algorithm to design DOEs is the stagnation problem [46]. Because the original Gerchberg-Saxton algorithm quantizes the continuous phase values into the set of discrete levels in each iteration, a given phase element may not have a large enough phase value change to go from one discrete level to another discrete level after quantization. Therefore, the stagnation problem usually results into poor DOE performance. The stagnation problem can be resolved by using an appropriate quantization scheduling that gradually quantizes the phase of the DOE [46].

We have also incorporated another modification on the algorithm that can improve the DOE performance slightly. In the original Gerchberg-Saxton algorithm, the magnitudes of the noise diffraction orders are forced to be zero in each iteration. In general, the magnitudes of the noise diffraction orders will never be zero. At the early iterations of the algorithm, the noise diffraction orders usually have high magnitudes. Rather than forcing the noise orders to be zero in magnitude, we can attenuate them with a large attenuation factor. This attenuation factor is then gradually decreased as the algorithm iterates. At the end of the iteration, the noise diffraction order should have much lower magnitude than at the early stage of the iteration. Hence, we accept the noise diffraction order without further attenuation.

By using this noise-adjustment scheduling and the phase-quantization scheduling, the algorithm should converge to better DOE solutions. Note that, because of the noise-adjustment scheduling and the phase-quantization scheduling, this proposed algorithm will take more iterations to converge. Figure 8-2 shows the schematic flow diagram of the modified Gerchberg-Saxton algorithm with the noise-adjusting scheduling and the phase-quantization scheduling.

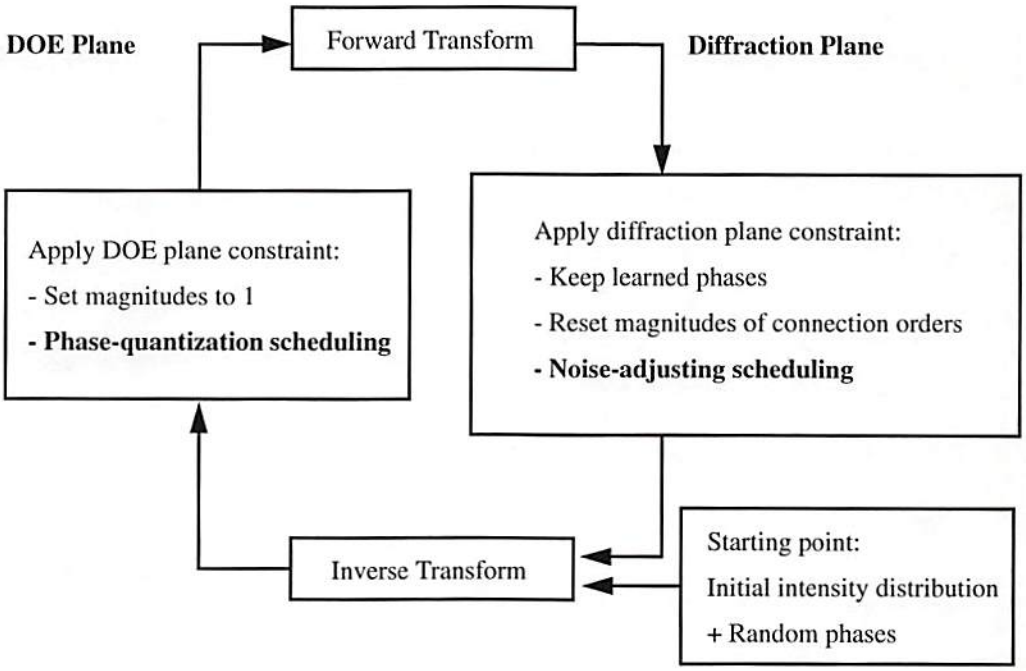

Figure 8-2: Schematic diagram of the modified Gerchberg-Saxton algorithm.

We have designed the 3×3 spot array generator DOE using both the original and modified Gerchberg-Saxton algorithms and compared their performance. The spot array generator has a target pattern of

$$\text{Sopt Array} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{8-3}$$

To measure the performance of DOEs, we have to define a set of appropriate performance indicators. We will use non-uniformity (*NU*), minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency (η), and relative diffraction efficiency (η_w), to measure the performance of the DOEs. They are defined as

$$NU = \frac{\max(\text{signal intensity}) - \min(\text{signal intensity})}{\max(\text{signal intensity}) + \min(\text{signal intensity})} \times 100\%, \tag{8-4}$$

$$MinSNR = \frac{\min(\text{signal intensity})}{\max(\text{noise intensity})}, \tag{8-5}$$

$$\eta = \frac{\text{Light Energy in Signal Orders}}{\text{Total Light Energy}} \times 100\%, \tag{8-6}$$

and

$$\eta_w = \frac{\text{Light Energy in Signal Orders}}{\text{Light Energy in J} \times \text{J Orders}} \times 100\%. \tag{8-7}$$

Note that the denominator of Eq. (8-7) is the total light energy in the central $J \times J$ diffraction orders. Because of the randomness of the starting phase distribution, the algorithms will converge to different solutions for different runs. Therefore, we run the algorithms fifty times and pick the DOE with good non-uniformity, minimum signal-to-noise ratio, and absolute diffraction efficiency for each case. Table 8-1 lists the performance of the DOE designed for the 3×3 spot array generator using both original and modified Gerchberg-Saxton algorithms. Figure 8-3 shows the designed DOE

Table 8-1: Performance of the DOE designed using the original and modified Gerchberg-Saxton algorithms for 3×3 spot array. J is the number of phase element in one period of the DOE in each dimension and Z is the number of phase levels

J	Z	Algorithm	NU (%)	MinSNR	η (%)	η_w (%)
8	2	original	55.9	2.5	69.9	81.5
8	2	modified	24.0	2.6	60.1	74.0
8	4	original	18.6	7.8	75.5	87.6
8	4	modified	1.5	11.1	72.2	86.4
8	8	original	8.7	38.3	85.8	95.0
8	8	modified	3.4	44.1	84.5	94.4
8	16	original	9.2	35.3	87.2	95.9
8	16	modified	1.5	46.1	87.0	95.5
16	2	original	57.2	3.2	72.4	78.4
16	2	modified	32.3	5.8	69.6	76.1
16	4	original	16.9	10.9	77.7	83.5
16	4	modified	1.7	11.1	76.5	82.5
16	8	original	6.1	27.9	88.2	92.5
16	8	modified	2.0	43.7	87.8	92.1
16	16	original	7.8	26.0	91.4	94.5
16	16	modified	0.7	40.2	91.2	94.2

performance using 8×8 phase elements in each period of the DOE with two, four, eight, and sixteen phase levels. Figure 8-4 shows the designed DOE performance using 16×16 phase elements in each period of the DOE with two, four, eight, and sixteen phase levels. From Figure 8-3 and 8-4, we can see that the modified Gerchberg-Saxton algorithm with the noise-adjusting scheduling and the phase-quantization scheduling can effectively converge to DOEs with lower non-uniformity and higher signal-to-noise ratio than the original Gerchberg-Saxton algorithm for this spot array target pattern although it takes longer to converge. Figure 8-5 shows the DOE performance as functions of phase levels

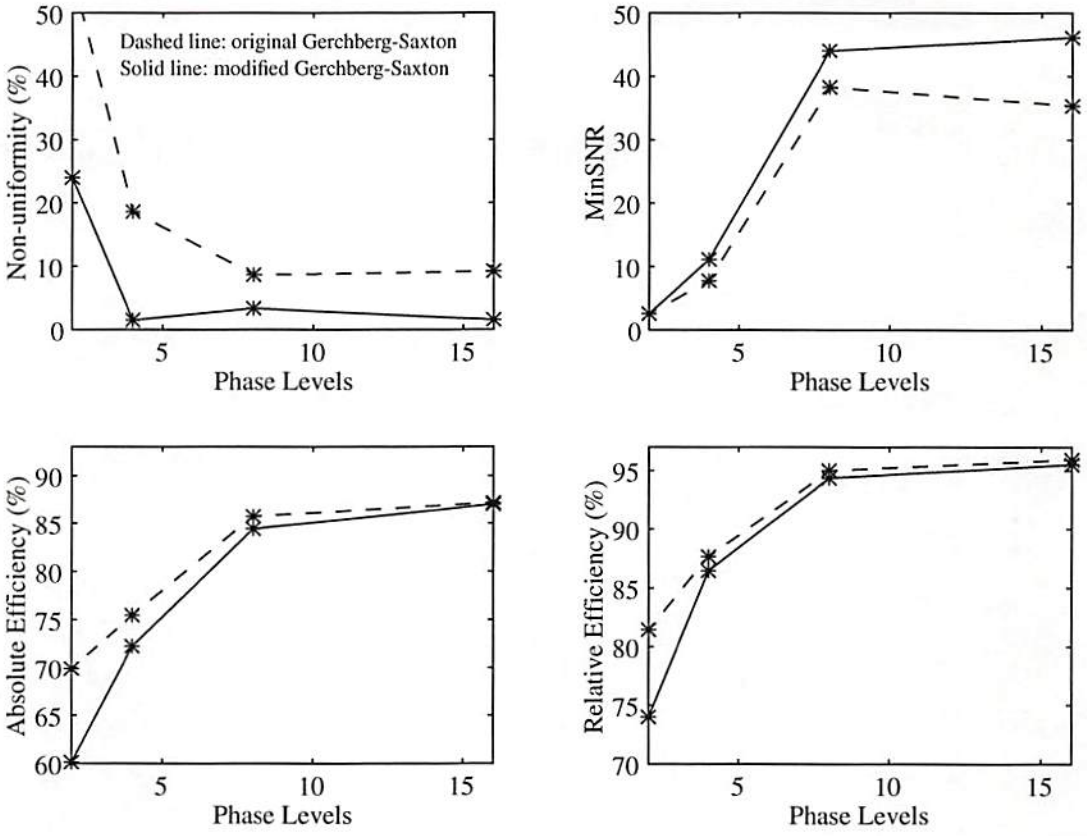

Figure 8-3: DOE performance as a function of phase levels for the 3×3 spot array generator using the original and modified Gerchberg-Saxton algorithms with 8×8 phase elements in each period of the DOEs.

and number of phase elements in each period of the DOE using the modified Gerchberg-Saxton algorithm. From Figure 8-5, it is clear that the DOE will perform better if more phase elements in each period of the DOE and/or phase levels are used.

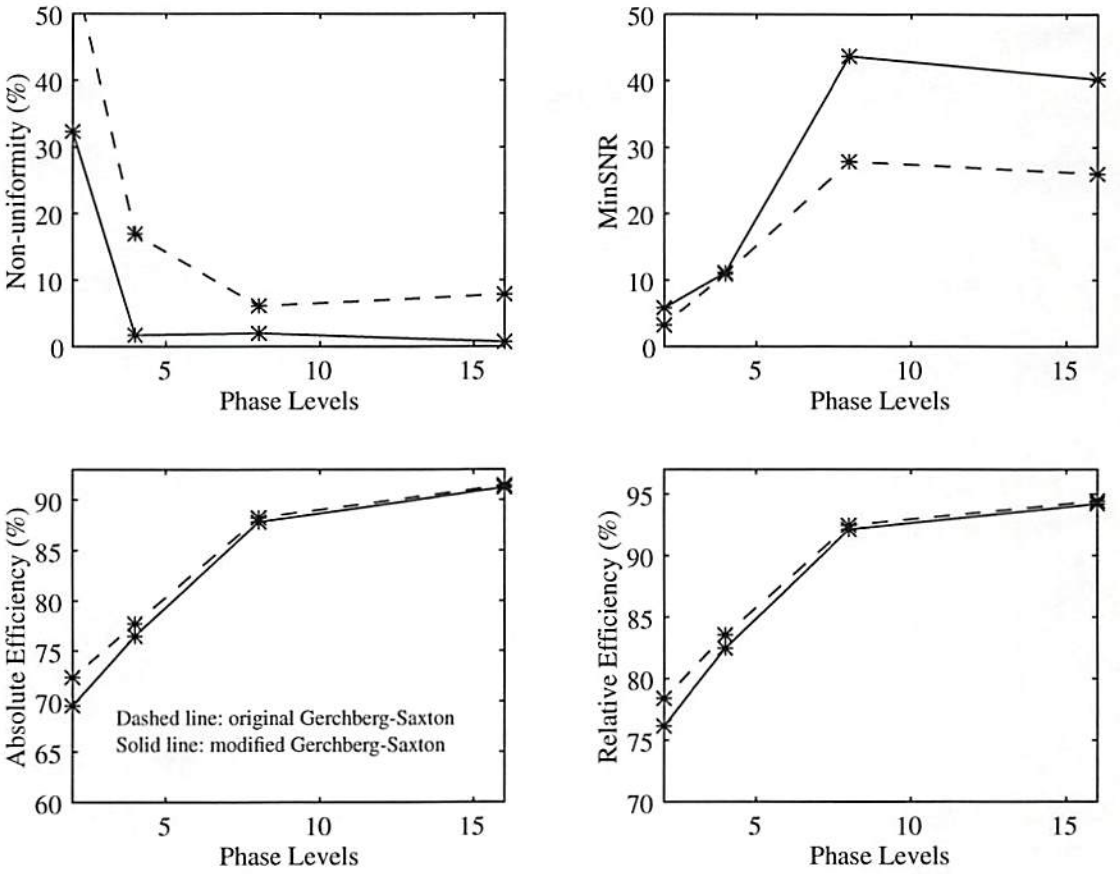

Figure 8-4: DOE performance as a function of phase levels for the 3×3 spot array using the original and modified Gerchberg-Saxton algorithms with 16×16 phase elements in each period of the DOEs.

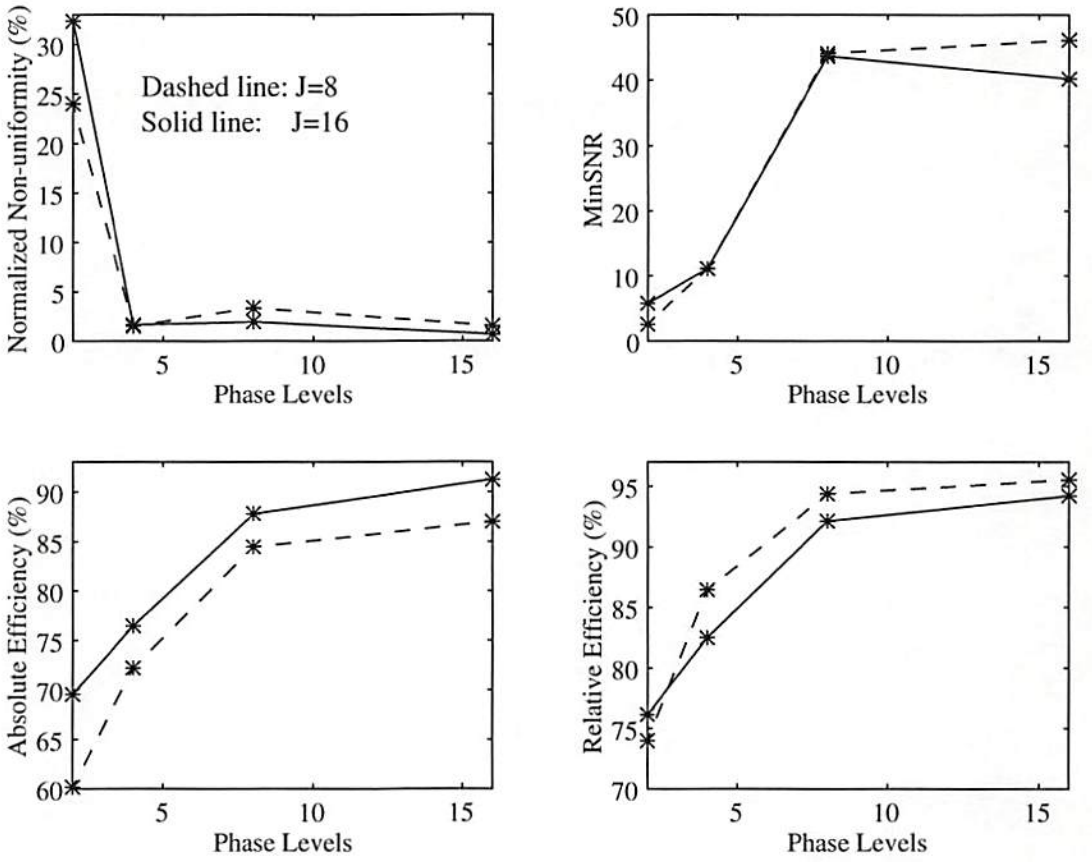

Figure 8-5: DOE performance as a function of phase levels for the 3 × 3 spot array using the modified Gerchberg-Saxton algorithms with 8 × 8 or 16 × 16 phase elements in each period of the DOEs.

We have also designed DOEs using 8×8 or 16×16 phase elements in each period of the DOE with two, four, eight, and sixteen phase levels for the 3×3 triangular pattern using the original and modified algorithms. The triangular connection pattern has a target pattern of

$$Triangular = \begin{bmatrix} 0.25 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{bmatrix}. \tag{8-8}$$

Note that because the desired connection weights in this pattern are not constant as in the case of the spot array generator, the non-uniformity of the DOE is calculated after the reconstructed signal diffraction orders are normalized by their corresponding desired connection weights. Therefore, the non-uniformity (NU) is redefined as

$$NU = \frac{\max\left[\frac{\text{signal intensity}}{\text{target weight}}\right] - \min\left[\frac{\text{signal intensity}}{\text{target weight}}\right]}{\max\left[\frac{\text{signal intensity}}{\text{target weight}}\right] + \min\left[\frac{\text{signal intensity}}{\text{target weight}}\right]} \times 100\%. \tag{8-9}$$

Table 8-2 lists the performance of the DOE designed for the 3×3 triangular pattern using both the original and modified Gerchberg-Saxton algorithms. The results are plotted

Table 8-2: Performance of the DOE designed using the original and modified Gerchberg-Saxton algorithm for 3×3 triangular pattern

J	Z	Algorithm	NU (%)	MinSNR	η (%)	η_w (%)
8	2	original	20.	3.6	82.0	90.7
8	2	modified	28.5	3.4	80.0	89.8
8	4	original	15.8	6.5	78.3	88.6
8	4	modified	8.3	6.4	73.6	85.7
8	8	original	28.6	5.1	84.8	92.8
8	8	modified	3.2	7.9	83.9	92.2
8	16	original	20.7	17.0	87.6	95.2
8	16	modified	1.9	11.9	86.8	94.3
16	2	original	22.9	9.2	82.7	87.4
16	2	modified	19.6	5.3	82.0	86.7
16	4	original	18.1	8.0	82.3	87.0
16	4	modified	5.2	11.5	80.2	85.5
16	8	original	28.0	11.3	87.4	91.4
16	8	modified	1.9	12.9	86.4	90.1
16	16	original	19.8	4.0	87.1	90.1
16	16	modified	1.1	17.8	90.3	92.7

in Figure 8-6 and Figure 8-7. From Figure 8-6 and 8-7, we can see that the modified Gerchberg-Saxton algorithm with the noise-adjusting scheduling and the phase-quantization scheduling can effectively converge to DOEs with much lower non-uniformity and higher signal-to-noise ratio than the original Gerchberg-Saxton algorithm for this triangular connection pattern. We have also plotted in Figure 8-8 the DOE performance as functions of phase levels and number of phase elements in each period of the DOE using the modified algorithm for the triangular pattern. Figure 8-8 still shows that the DOE will perform better if more phase elements in each period of the DOE and/or

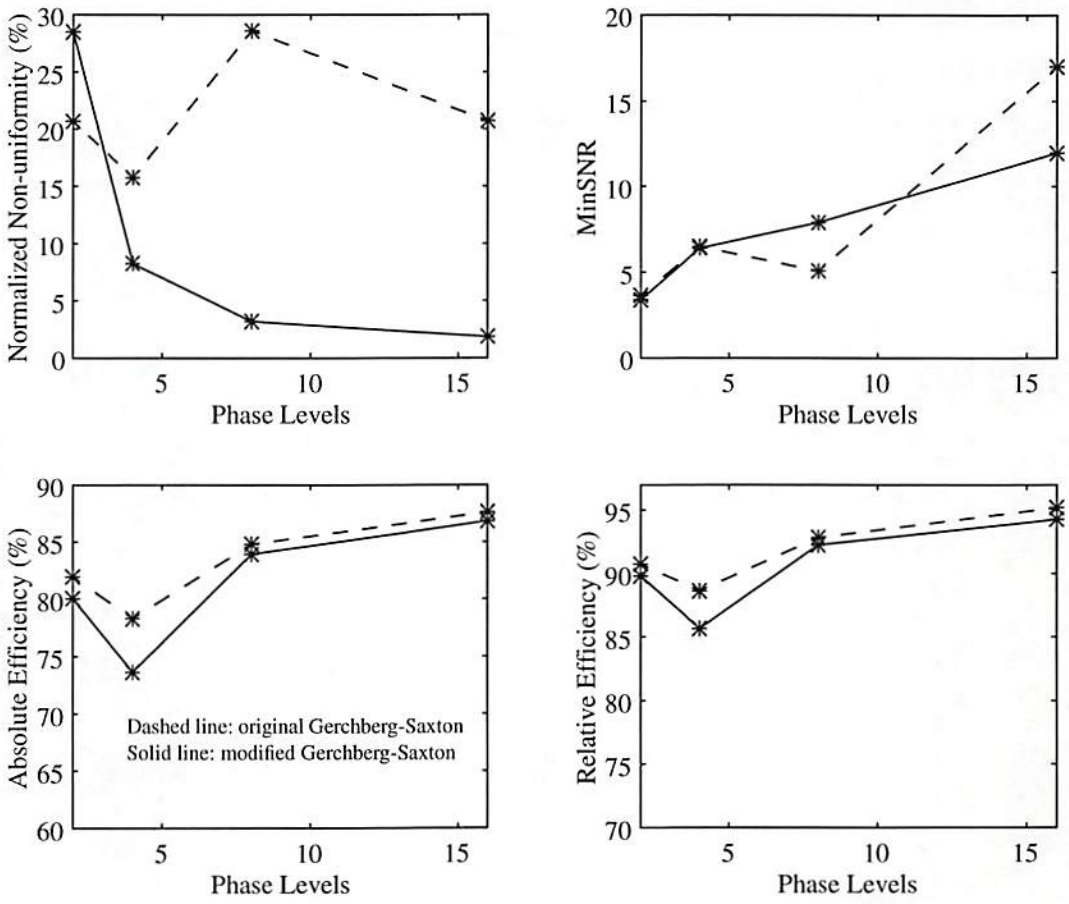

Figure 8-6: DOE performance as a function of phase levels for the 3×3 triangular pattern using the original and modified Gerchberg-Saxton algorithms with 8×8 phase elements in each period of the DOEs.

phase levels are used. Note that both the original and modified Gerchberg-Saxton algorithms give rather poor DOE solutions when the phase level is two (binary-phase-level DOEs). In the next section, we will discuss the use of simulated annealing algorithm that will be proven useful in these cases. In the rest of the chapter, we will develope other design approaches and compare their performance against the performance of this modified Gerchberg-Saxton.

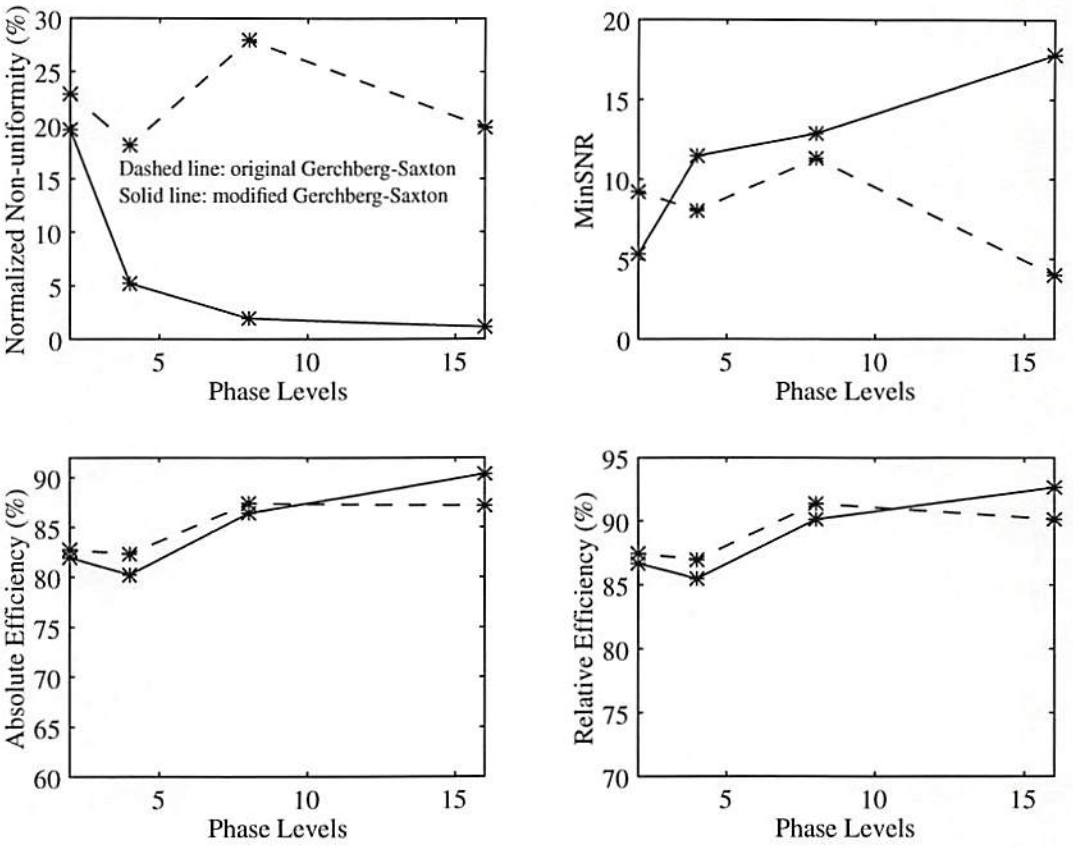

Figure 8-7: DOE performance as a function of phase levels for the 3×3 triangular pattern using the original and modified Gerchberg-Saxton algorithms with 16×16 phase elements in each period of the DOEs.

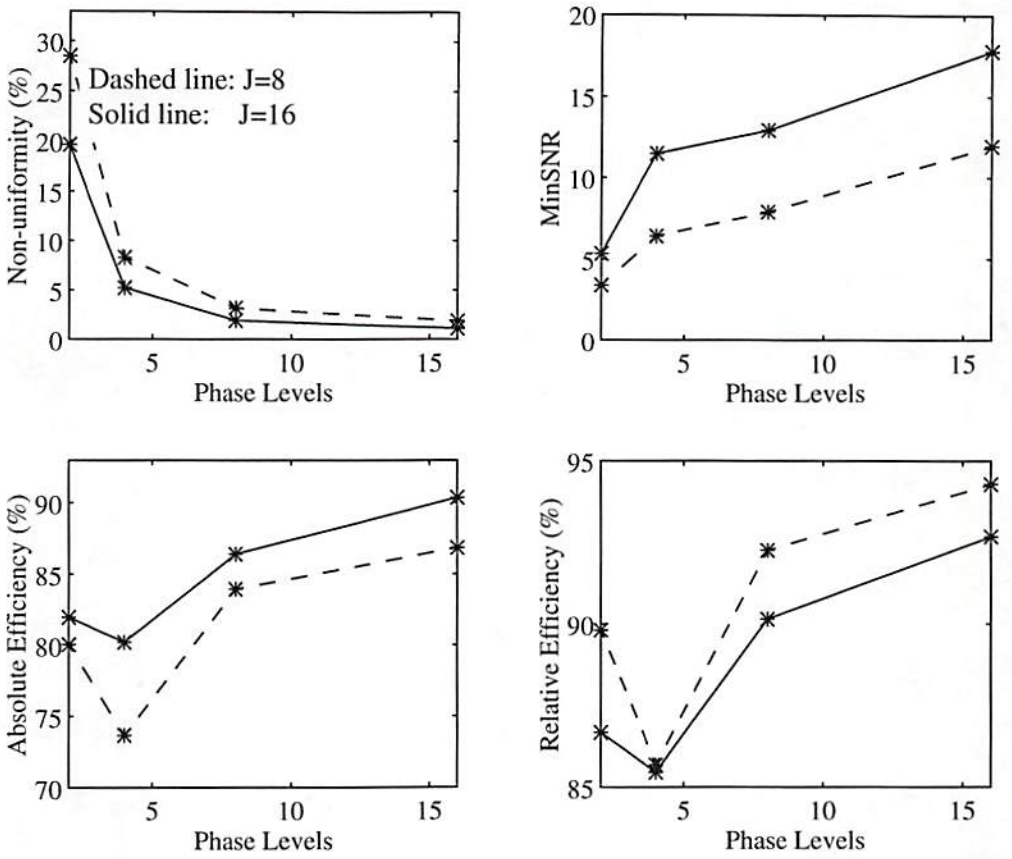

Figure 8-8: DOE performance as a function of phase levels for the 3×3 triangular pattern using the modified Gerchberg-Saxton algorithms with 8×8 or 16×16 phase elements in each period of the DOEs.

8.2 Dual-Cost-Function Simulated Annealing Algorithm

Simulated annealing is a time-consuming process that can theoretically obtain an optimum solution by using a statistical searching procedure with an appropriate annealing schedule and unlimited computation time [42]. Figure 8-9 shows the basic conceptual flow diagram of the simulated annealing algorithm. The algorithm begins with a randomly selected phase distribution and an initial annealing temperature. The corresponding diffraction pattern is then calculated from Eq. (8-1) and the cost function of the DOE is calculated. The cost function is the figure of merit of the DOE, and a commonly used cost function is the root-mean-squared-error (*RMSE*) [42, 50, 52] that is defined by

$$RMSE \equiv \left[\frac{\sum_r \sum_s \left[W(r,s) - \eta(r,s) \right]^2}{J^2} \right]^{\frac{1}{2}}, \tag{8-10}$$

where $W(r,s)$ is the desired connection weight and $\eta(r,s)$ is the intensity of the diffraction order. The algorithm then randomly selects a phase element from one period of the DOE and randomly changes its phase to one of the Z pre-defined phase levels. The corresponding diffraction pattern and cost are then calculated for this trial DOE phase distribution. If the cost of the trial is less than the original cost, the change is unconditionally accepted. Otherwise, the change is accepted only if the probability of accept, which is usually some function of the annealing temperature and the difference between the trial cost and the original cost, is greater than some randomly selected minimum probability (a uniformly distributed random number

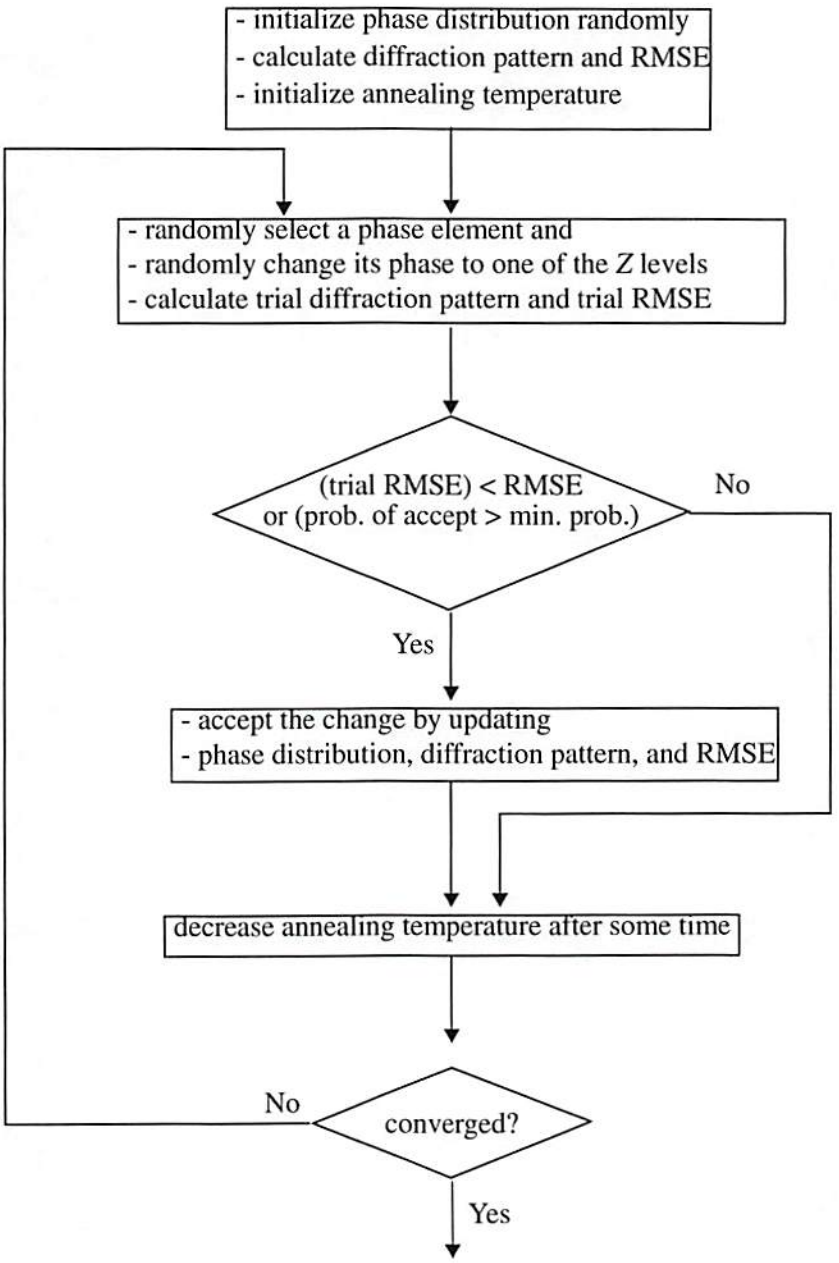

Figure 8-9: Schematic diagram of the simulated annealing algorithm

between 0 and 1). In this thesis, we use the following form of the probability of accepting a bad change [50] as

$$\text{Probability of accepting a bad change} \equiv \frac{1}{1+\exp\left[\frac{(RMSE \text{ of the trial}) - RMSE}{\text{annealing temperature}}\right]}. \quad (8\text{-}11)$$

This mechanism of accepting a bad change is very important because it is used to move the algorithm out of a local minimum. After some time, the annealing temperature is decreased to reduce the probability of accepting a bad change. In this thesis, we decrease the annealing temperature by a factor of 0.9 ~0.95 when the number of accepted changes reaches a pre-defined threshold. This process is then repeated until it converges to the final solution. In each iteration, the simulated annealing algorithm operates on only one of the J^2 phase elements rather than on the entire J^2 phase elements simultaneously as in the case of the Gerchberg-Saxton algorithm. Therefore, the simulated annealing algorithm will, in general, take on the order of J^2 times more iterations to converge than the Gerchberg-Saxton algorithm. Further, the total number of iterations will also increase as the number of phase levels increases. It is important to use an appropriate initial value and annealing scheduling for the annealing temperature to ensure that the algorithm can move away from the local minimum. If the initial annealing temperature is too high or if the annealing temperature is decreased at a very low rate in the iterations, then it will take the annealing algorithm a very long time to converge. On the other hand, if the initial annealing temperature is too low or if the annealing temperature is decreased too fast in the iterations, then the annealing algorithm will be more likely to get trapped at one of the local minima. It is also important to choose a suitable cost function in order to get good performance from

the DOE. In this section, we will introduce a new cost function other than the root-mean-squared-error and compare their effect on the results of the simulated annealing algorithm.

We can form a weight vector, $\vec{W}$, from the desired connection weights and a diffraction vector, $\vec{\eta}$, from the diffraction order intensities

$$\vec{W} \equiv [\, W(r,s) \,] \ , \tag{8-12}$$

and

$$\vec{\eta} \equiv [\, \eta(r,s) \,] \ , \tag{8-13}$$

where $r, s \in \{-\frac{J}{2}, \ldots, 0, \ldots, \left(\frac{J}{2}-1\right)\}$. Note that both the weight vector and the diffraction vector have J^2 components and only ${J_0}^2$ components in each of the two vectors are the desired ones. In this way, we can view the set of desired connection weights and the set of diffraction order intensities as vectors in the J^2-dimensional hyper-space. The absolute diffraction efficiency is then proportional to the sum of the desired components of the diffraction vector. The root-mean-squared-error can be expressed in terms of the length of the difference of the weight and diffraction vectors as

$$RMSE \equiv \left[\frac{\sum_r \sum_s [\, W(r,s) - \eta(r,s) \,]^2}{J^2} \right]^{\frac{1}{2}} = \left[\frac{|\vec{W} - \vec{\eta}|^2}{J^2} \right]^{\frac{1}{2}} = \frac{|\vec{W} - \vec{\eta}|}{J} . \tag{8-14}$$

Therefore, the simulated annealing algorithm will try to find the DOE with the shortest distance between the diffraction vector and the weight vector. This sometimes can lead to a DOE with high diffraction efficiency but poor non-uniformity. In most applications, it is important to maintain the relative strength among the diffraction orders while looking for

high diffraction efficiency, but less important for the overall diffraction efficiency to be the same as the original target value. One way to emphasize the relative strength among desired connection orders and undesired noise orders in the weight vector is to normalize the vectors before measuring the distance between them. Let's now define a new cost function, normalized-root-mean-squared-error (*NRMSE*), as

$$NRMSE \equiv \frac{\sum_r \sum_s \left[\frac{W(r,s)}{|\vec{W}|} - \frac{\eta(r,s)}{|\vec{\eta}|} \right]^2}{J^2} = \frac{\left| \frac{\vec{W}}{|\vec{W}|} - \frac{\vec{\eta}}{|\vec{\eta}|} \right|^2}{J}. \qquad (8\text{-}15)$$

This cost function will force the annealing algorithm to look for a DOE to best match the relative strength in the weight vector. This concept is demonstrated in Figure 8-10. In

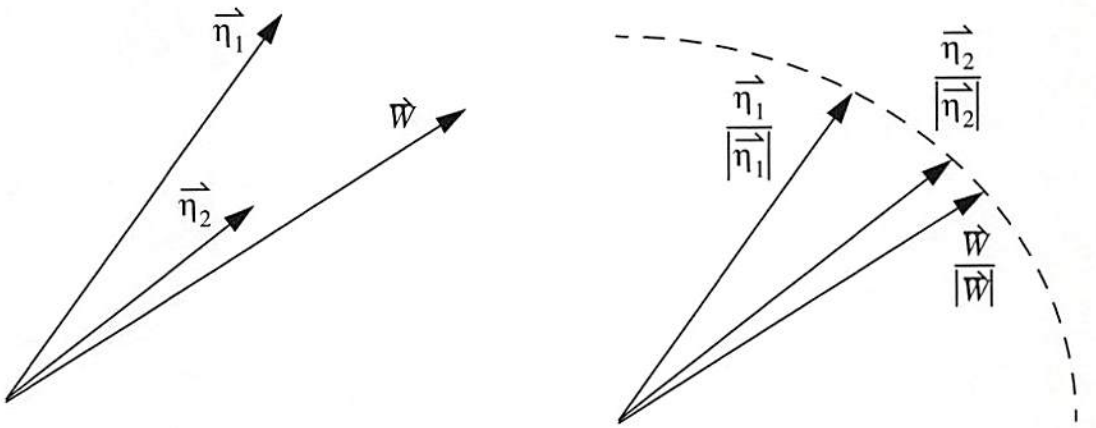

Figure 8-10: Conceptual diagram for the effect of the cost function on simulated annealing algorithm

Figure 8-10, the original simulated annealing algorithm will choose $\vec{\eta}_1$ over $\vec{\eta}_2$ because the distance between $\vec{\eta}_1$ and $\vec{W}$ is smaller than the distance between $\vec{\eta}_2$ and $\vec{W}$. However,

$\vec{\eta}_2$ can offer a better non-uniformity among the desired connection orders for the case. If we normalize the vectors and use the new cost function (*NRMSE*), then the algorithm will now select $\vec{\eta}_2$ instead of $\vec{\eta}_1$. On the other hand, from our design experience, the simulated annealing algorithm with NRMSE as the cost function usually converges to DOE solutions with good uniformities but low diffraction efficiencies. This is because the NRMSE cost function forces the simulated annealing algorithm to look for a solution that best matches the relative strength of desired connection orders and undesired noise orders in the weight vector. What we really want is DOEs with good uniformities and high diffraction efficiencies. The normalized-root-mean-squared-error alone can not achieve the goal. Therefore, we introduce a second companion cost function, total noise energy (*TN*), as the total noise energy in the mainlobe and all the sidelobes. Basically, we want to have DOEs with good uniformities and low noise energy. By forcing the simulated annealing algorithm look for a DOE solution with low normalized-root-mean-squared-error and low total noise energy, we should have a better solution than the *RMSE*-annealing approach.

In order to incorporate the two cost functions, *NRMSE* and *TN*, we have developed a dual-cost-function simulated annealing algorithm (*DCFSA*) as shown in Figure 8-11. This new algorithm accepts the change if one of the four conditions is true: the trial costs for *NRMSE* and *TN* are both decreased, or the trial cost for *NRMSE* is decreased and the probability of accepting a bad change for *TN* is greater than some randomly selected probability, or the probability of accepting a bad change for *NRMSE* is greater than some randomly selected probability and the trial cost for *TN* is decreased, or the probability of accepting a bad change for *NRMSE* is greater than some randomly selected probability

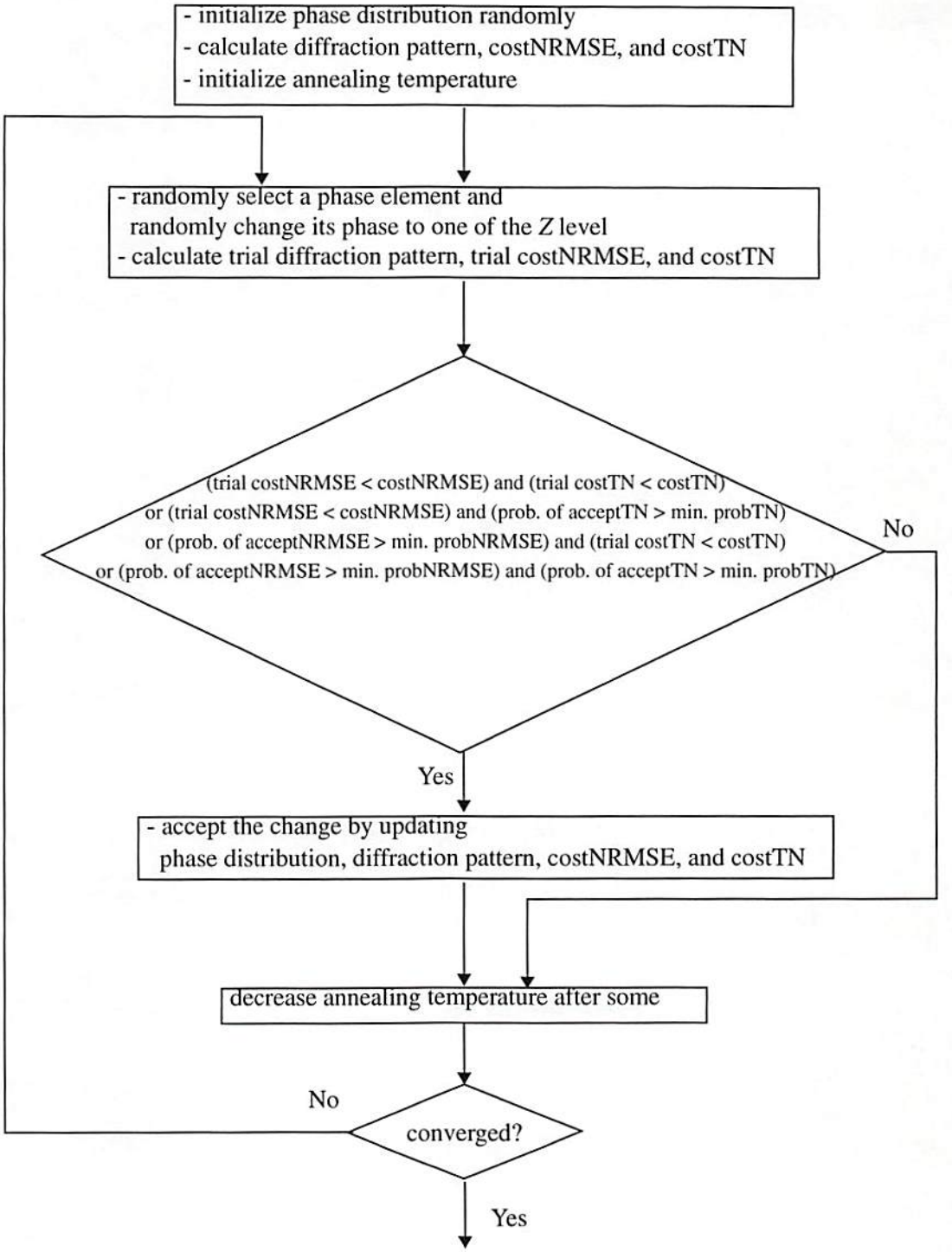

Figure 8-11: Schematic diagram of the dual-cost-function simulated annealing algorithm

and the probability of accepting a bad change for *TN* is greater than some randomly selected probability. It is also possible to incorporate an additional weight factor to control the relative importance of the two cost functions to further tune the algorithm.

Using this new dual-cost-function simulated annealing algorithm, we have redesigned the 3×3 spot array and the 3×3 triangular connection pattern using 8×8 phase elements in each period of the DOE with two, four, and eight phase levels. For comparison, we have also designed DOEs using the original simulated annealing algorithm with the cost function of the sum-of-squared-error. Because of randomness in the simulated annealing algorithm, we have run the design algorithms 30~50 times for each case and picked the one with best NRMSE. Table 8-3 lists the performance of the DOEs designed using the original and dual-cost-function simulated annealing algorithms (*RMSESA* and *DCFSA*, respectively) in terms of non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency (η), and normalized-root-mean-squared-error (*NRMSE*) for the target pattern of the 3×3 spot array generator. For comparison, we also list the data of the DOEs designed in the previous section using the modified Gerchberg-Saxton (*MGS*) algorithm. Table 8-4 lists the performance of the DOE in terms of normalized non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*) for the 3×3 triangular weighted connection pattern. From Table 8-3 and 8-4, it is clear that dual-cost-function simulated annealing algorithm performs better than the original simulated annealing algorithm in almost all the figure of merits.

Table 8-3: Performance of the DOE designed using modified Gerchberg-Saxton (MGS), original simulated annealing (RMSESA), and dual-cost-function simulated annealing (DCFSA) algorithms for 3 × 3 spot array generator. NRMSE is the normalized-root-mean-squared-error.

J	Z	Algorithm	NU (%)	MinSNR	η (%)	NRMSE
8	2	MGS	24.0	2.6	60.1	3.6495
8	2	RMSESA	18.2	3.1	65.1	2.6643
8	2	DCFSA	4.3	6.4	58.5	1.9104
8	4	MGS	1.5	11.1	72.2	1.1328
8	4	RMSESA	10.0	7.4	75.1	1.4328
8	4	DCFSA	3.8	15.0	72.2	1.0028
8	8	MGS	3.4	44.1	84.5	0.5529
8	8	RMSESA	4.5	31.9	85.4	0.5593
8	8	DCFSA	3.4	44.1	84.5	0.5529

Table 8-4: Performance of the DOE designed using modified Gerchberg-Saxton (MGS), original simulated annealing (RMSESA), and dual-cost-function simulated annealing (DCFSA) algorithms for 3 × 3 triangular pattern. NRMSE is the normalized-root-mean-squared-error.

J	Z	Algorithm	NU (%)	MinSNR	η (%)	NRMSE
8	2	MGS	20.7	3.6	82.0	1.9381
8	2	RMSESA	20.7	3.6	82.0	1.9381
8	2	DCFSA	12.2	5.9	73.2	1.1364
8	4	MGS	8.3	6.4	73.6	1.1024
8	4	RMSESA	20.7	3.6	82.0	1.9381
8	4	DCFSA	9.2	7.4	73.9	0.9914
8	8	MGS	3.2	7.9	83.9	0.6129
8	8	RMSESA	9.4	10.8	85.6	0.7692
8	8	DCFSA	3.2	8.1	84.4	0.5963

Let's now compare the DOE performance of the dual-cost-function simulated annealing algorithm with the DOE performance of the modified Gerchberg-Saxton

algorithm developed in the previous section. Figure 8-12 shows the non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*) of the DOEs designed for the 3×3 spot array generator using the modified Gerchberg-Saxton algorithm developed in the previous section and the dual-cost-function simulated annealing algorithm. Figure 8-13 shows the

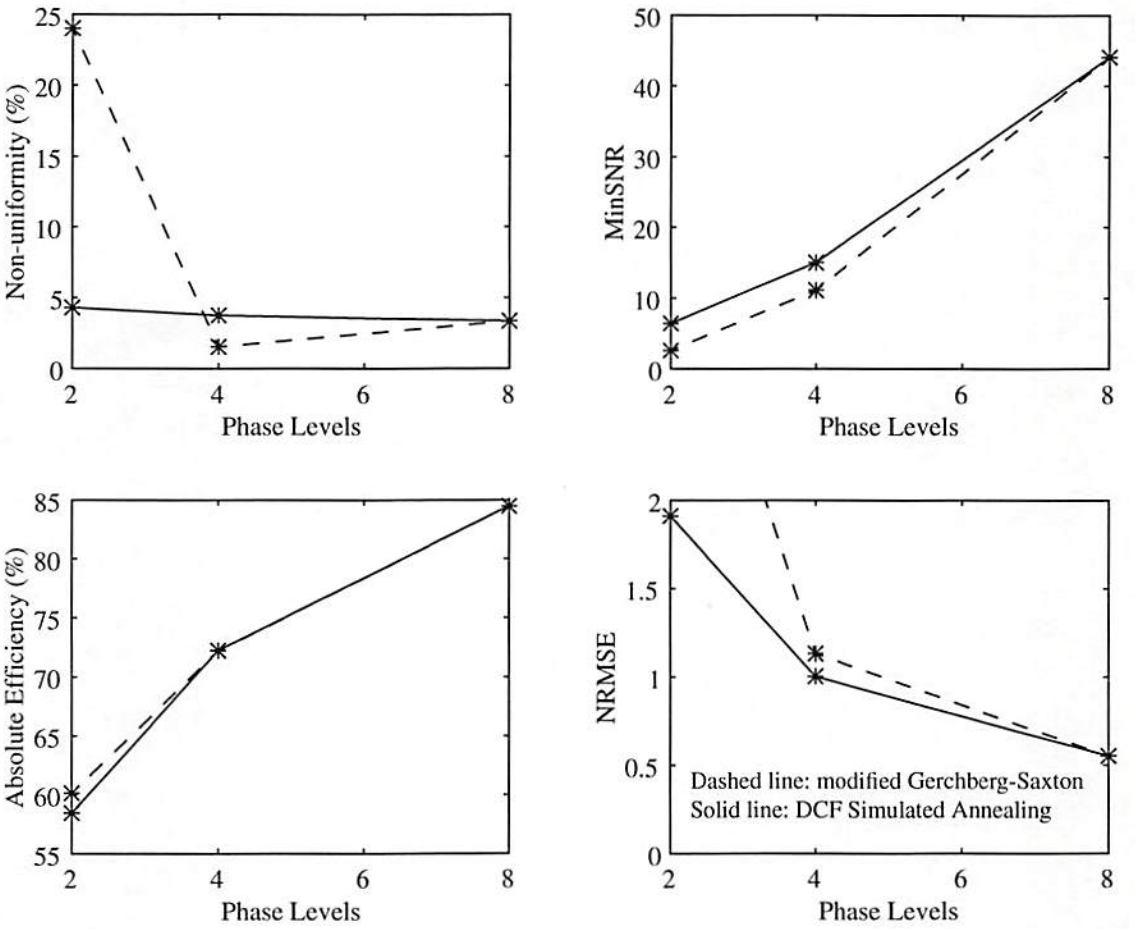

Figure 8-12: DOE performance comparison as a function of phase levels for the 3×3 spot array generator between the modified Gerchberg-Saxton and dual-cost-function simulated annealing algorithms with 8×8 phase elements in each period of the DOEs.

normalized non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*) of the DOEs designed for the 3×3 triangular connection pattern using the modified Gerchberg-Saxton algorithm developed in the previous section and the dual-cost-function simulated annealing algorithm. For the case of two-phase-level DOE, as can be seen from Table 8-4 and

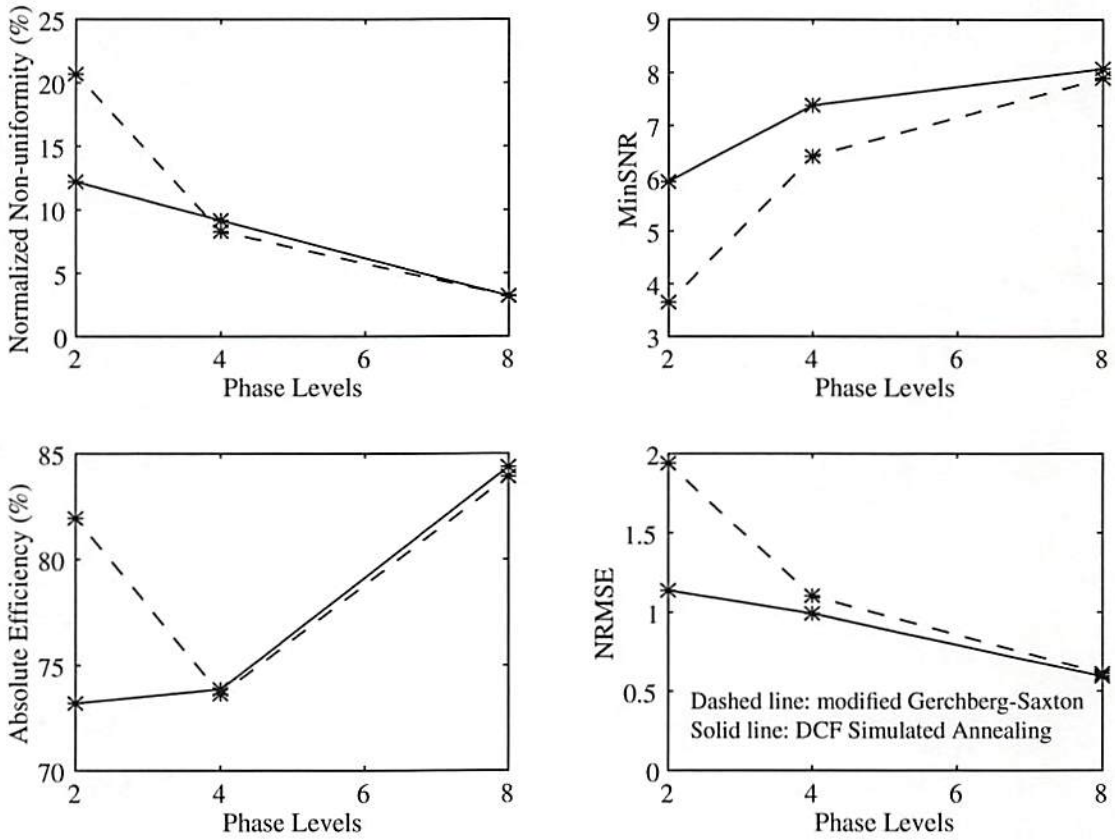

Figure 8-13: DOE performance comparison as a function of phase levels for the 3×3 triangular connection pattern between the modified Gerchberg-Saxton and dual-cost-function simulated annealing algorithms with 8×8 phase elements in each period of the DOEs.

Figure 8-13, although the modified Gerchberg-Saxton algorithm provides higher diffraction efficiency, the normalized non-uniformity, minimum signal-to-noise ratio,

absolute diffraction efficiency, and normalized-root-mean-squared-error are all much worse than the dual-cost-function simulated annealing algorithm for the triangular connection pattern. From Figure 8-12 and 8-13, it is also clear that the dual-cost-function simulated annealing algorithm can provide a substantial amount of performance improvement over the already modified Gerchberg-Saxton algorithm. This performance improvement is most significant for the case of binary-phase-level DOEs and generally decreases as the number of the phase levels in the DOEs increases.

8.3 Using Unequally Spaced Phase Levels in DOE Design

Each phase element of a DOE can only have one of the pre-defined phase level indices. Each of the phase level indices is assigned a phase delay relative to the zero-phase-level-index. Usually, those phase delays are equally distributed between 0 and 2π. In this section, we will point out that the use of equally spaced phase levels is really unnecessary and can be relieved to increase the design freedom of the optimization algorithm. For a Z-phase-level DOE, the phase level index, $PI(p, q)$, of a phase element can have any index value from 0 to (Z-1). Each phase level index corresponds to one of the Z possible phase delays. To fabricate this Z-phase-level DOE, $\log_2 Z$ etch steps are usually used as suggested by Veldkamp [10]. Note that it is also possible to use deposition steps (adding material onto the DOE substrate) instead of etch steps in DOE fabrication. By changing the sign of the phase levels, the work in this section will still be applicable for the case of deposition steps. In this thesis, DOEs are assumed to be fabricated by etching.

In each etch step, the relative phase delay, φ_i, (relative to $PI(p, q) = 0$) created by the etch is usually

$$\varphi_i = \left(\frac{2\pi}{Z}\right) \cdot 2^{i-1}, \tag{8-16}$$

where $i = 1, 2, \ldots\ldots, \log_2 Z$. In this thesis, we refer to these $\log_2 Z$ phase delays as the independent phase delays and the remaining $Z - \log_2 Z$ phase delays as the dependent phase delays since each of the dependent phase delays is just a linear combination of the independent phase delays. If we express the phase level index, $PI(p, q)$, by its binary representation as

$$PI(p, q) = \sum_{i=1}^{\log_2 Z} u_i \cdot 2^{i-1}, \tag{8-17}$$

where u_i is the i'th bit of the binary representation of $PI(p, q)$ and it can take on either one or zero. Then the relative phase delay, $PD(p, q)$, can be expressed as

$$PD(p, q) = \sum_{i=1}^{\log_2 Z} u_i \cdot \varphi_i. \tag{8-18}$$

At the i'th etch step, if the i'th binary bit of the phase level index, $PI(p, q)$, of a phase element is one, then a phase delay of φ_i is needed and this phase delay is generated by performing an etch into the DOE substrate for the phase elements with a zero in their i'th bit. Otherwise, there is no etch and no phase delay generated. The correct etch depth, d_i, for generating a phase delay of φ_i can be shown to be

$$d_i = \frac{\varphi_i \cdot \lambda}{2\pi \cdot (n-1)}. \tag{8-19}$$

where n is the index of refraction and λ is the optical wavelength in the air. Note that, in Eq. (8-19), we have assumed an air interface for the DOE. For a more general case, the one in Eq. (8-19) should be replaced by the refraction index of the surrounding material. The total phase delay of a phase element is the sum of the phase delay generated from all the etch steps. Combining Eq. (8-16), (8-17), and (8-18), we can immediately see that the phase delay, $PD(p, q)$, of a phase element can be simply expressed as

$$PD(p, q) = \sum_{i=1}^{\log_2 Z} u_i \cdot \left(\frac{2\pi}{Z}\right) \cdot 2^{i-1} = \frac{2\pi}{Z} \cdot \sum_{i=1}^{\log_2 Z} u_i \cdot 2^{i-1} = \frac{2\pi}{Z} \cdot PI(p, q). \tag{8-20}$$

The phase delays are therefore equally spaced with a spacing of $2\pi/Z$.

The phase delays are equally spaced between 0 and $\frac{2\pi \cdot (Z-1)}{Z}$ because we restrict ourselves to etch to create a phase delay of $\left(\frac{2\pi}{Z}\right) \bullet 2^{i-1}$ at the i'th etch step. This restriction is really unnecessary and can be relieved. We can use these independent phase delays as additional free parameters in DOE design. Therefore, we will have an additional set of $\log_2 Z$ free parameters for optimizing the DOE. Each of the independent phase delays can have any value between 0 and 2π. Table 8-5 lists the phase delay distributions of the conventional equal-spaced independent phase delay scheme and the optimized independent phase delay scheme for two, four, and eight phase levels. From Table 8-5, we can see that the conventional equal-spaced independent phase delay scheme is just a special case of the more general optimized independent phase delay scheme. With this optimized independent phase delay scheme, the maximum phase delay is just the sum of all the independent phase delays and this maximum phase delay can be greater than 2π. In the conventional equal-spaced independent phase delay scheme, the maximum phase delay is always less than 2π.

Table 8-5: Equally spaced phase delays and optimized phase delays for two, four, and eight phase levels

Z	Phase Delay Type	Independent Phase Delays	All Possible Phase Delays
2	Equally Spaced	$\varphi_1 = \pi$	$0, \pi$
2	Optimized	φ_1	$0, \varphi_1$
4	Equally Spaced	$\varphi_1 = \pi/2$ $\varphi_2 = \pi$	$0, \pi/2, \pi, 3\pi/2$
4	Optimized	φ_1, φ_2	$0, \varphi_1, \varphi_2, \varphi_1 + \varphi_2$
8	Equally Spaced	$\varphi_1 = \pi/4$ $\varphi_2 = \pi/2$ $\varphi_3 = \pi$	$0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4$
8	Optimized	$\varphi_1, \varphi_2, \varphi_3$	$0, \varphi_1, \varphi_2, \varphi_1 + \varphi_2, \varphi_3, \varphi_1 + \varphi_3$ $\varphi_2 + \varphi_3, \varphi_1 + \varphi_2 + \varphi_3$

Note that the fabrication complexity remains the same for this proposed new scheme since we are not trying to optimize the phase delays of all the Z phase levels. We are only optimizing the $\log_2 Z$ independent phase delays and each of the other phase delays is still

the linear combination of the independent phase delays. We only need to change the etch depth in each etch step according to the optimized values for the independent phase delays from the design algorithm. In general, the resultant DOE will not have equally spaced phase delays between 0 and 2π. On the other hand, this proposed concept does increase the complexity of the design algorithm. Although this new concept should be applicable using other design algorithms, we will implement this idea using the dual-cost-function simulated annealing algorithm developed in the previous section.

Figure 8-14 shows the conceptual flow diagram for the dual-cost-function simulated annealing algorithm with phase delay optimization. The algorithm begins with a randomly selected phase distribution, an initial annealing temperature, and a randomly selected set of independent phase delays (each between 0 and 2π). The corresponding diffraction pattern is then calculated from Eq. (8-1) and the two cost functions of the DOE are calculated. The algorithm then randomly selects a phase element from one period of the DOE and randomly selects a new phase index for it. At the same time, a set of phase delay changes, $\{\Delta\varphi_i, i = 1, \ldots, \log_2 Z\}$, is randomly selected and the trial set of independent phase delays is formed by $\varphi_i^{trial} = \varphi_i + \Delta\varphi_i$. The maximum allowable value for the phase delay changes is limited to a fraction of π radian and this maximum allowable value is decreased with the annealing temperature. The corresponding diffraction pattern and costs are then calculated for this trial DOE phase distribution and independent phase delays. This new algorithm will accept the change on the phase index and the independent phase delays only if one of the four following conditions is true: the trial costs for *NRMSE* and *TN* are both decreased, or the trial cost for *NRMSE* is decreased

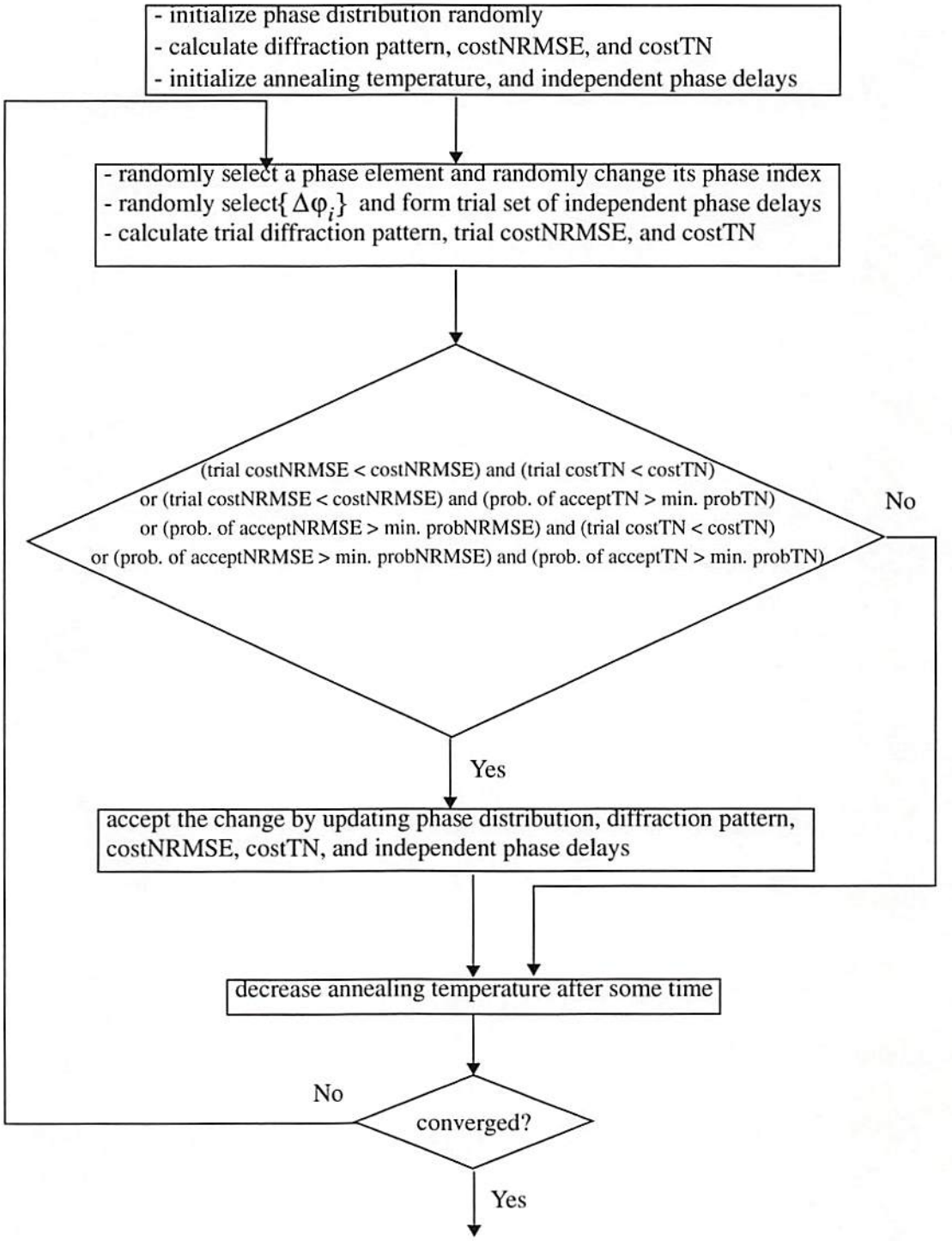

Figure 8-14: Schematic diagram of the dual-cost-function simulated annealing algorithm with phase delay optimization.

and the probability of accepting a bad change for *TN* is greater than some randomly selected probability, or the probability of accepting a bad change for *NRMSE* is greater than some randomly selected probability and the trial cost for *TN* is decreased, or the probability of accepting a bad change for *NRMSE* is greater than some randomly selected probability and the probability of accepting a bad change for *TN* is greater than some randomly selected probability. After some time, the annealing temperature is decreased to reduce the probability of accepting a set of bad changes. This process is then repeated until it converges to the final solution with a phase index distribution and set of independent phase delays.

Using this new dual-cost-function simulated annealing algorithm with phase delay optimization, we have redesigned the 3×3 spot array with 8×8 or 16×16 phase elements in each period of the DOE and two, four, and eight phase levels. Because of randomness in the simulated annealing algorithm, we have run the design algorithms 30~50 times for each case and picked the one with best NRMSE. Table 8-6 lists the performance of the DOE designed using the dual-cost-function simulated annealing algorithm without phase delay optimization (referred as DCFSA) and with phase delay optimization (referred as DCFSAOP) in terms of non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*). Table 8-6 also lists the optimized independent phase delays at the end of each row. For comparison purposes, the performance of the DOEs designed using modified Gerchberg-Saxton algorithm (referred as MGS) is also listed in Table 8-6. By comparing the data for dual-cost-function simulated annealing approaches in Table 8-6,

Table 8-6: Performance of the DOE designed using the modified Gerchberg-Saxton (MGS) algorithm and dual-cost-function simulated annealing algorithm with and without phase delay optimization (DCFSA and DCFSAOP, respectively) for 3×3 spot array generator.

J	Z	Algorithm	NU (%)	MinSNR	η (%)	NRMSE	Independent Phase Delays
8	2	MGS	24.0	2.6	60.1	3.6495	π
8	2	DCFSA	4.3	6.4	58.5	1.9104	π
8	2	DCFSAOP	4.4	7.1	67.7	1.5195	1.16π
8	4	MGS	1.5	11.1	72.2	1.1328	$\pi/2, \pi$
8	4	DCFSA	3.8	15.0	72.2	1.0028	$\pi/2, \pi$
8	4	DCFSAOP	2.4	21.9	79.0	0.7119	$0.38\pi, 0.76\pi$
8	8	MGS	3.4	44.1	84.5	0.5529	$\pi/4, \pi/2, \pi$
8	8	DCFSA	3.4	44.1	84.5	0.5529	$\pi/4, \pi/2, \pi$
8	8	DCFSAOP	0.7	35.6	84.5	0.3909	$0.25\pi, 0.69\pi, 1.2\pi$
16	2	MGS	32.3	5.8	69.6	1.8229	π
16	2	DCFSA	11.1	7.9	65.7	0.8835	π
16	2	DCFSAOP	0.9	7.1	72.1	0.7195	0.82π
16	4	MGS	1.7	11.1	76.5	0.5109	$\pi/2, \pi$
16	4	DCFSA	1.9	16.3	76.8	0.4710	$\pi/2, \pi$
16	4	DCFSAOP	0.9	19.8	81.8	0.3404	$0.34\pi, 0.80\pi$
16	8	MGS	2.0	43.7	87.8	0.1933	$\pi/4, \pi/2, \pi$
16	8	DCFSA	0.7	30.5	88.0	0.1874	$\pi/4, \pi/2, \pi$
16	8	DCFSAOP	1.4	40.0	88.1	0.1861	$0.24\pi, 0.48\pi, 1.06\pi$

we can see that the phase delay optimization approach generally provides better non-uniformity, better (or comparable) minimum signal-to-noise ratio, higher absolute diffraction efficiency, and lower normalized-root-mean-squared-error than the non-optimized approach.

Let's now compare the DOE performance between the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithm with phase delay optimization. In Figure 8-15, we have plotted non-uniformity, minimum signal-to-noise

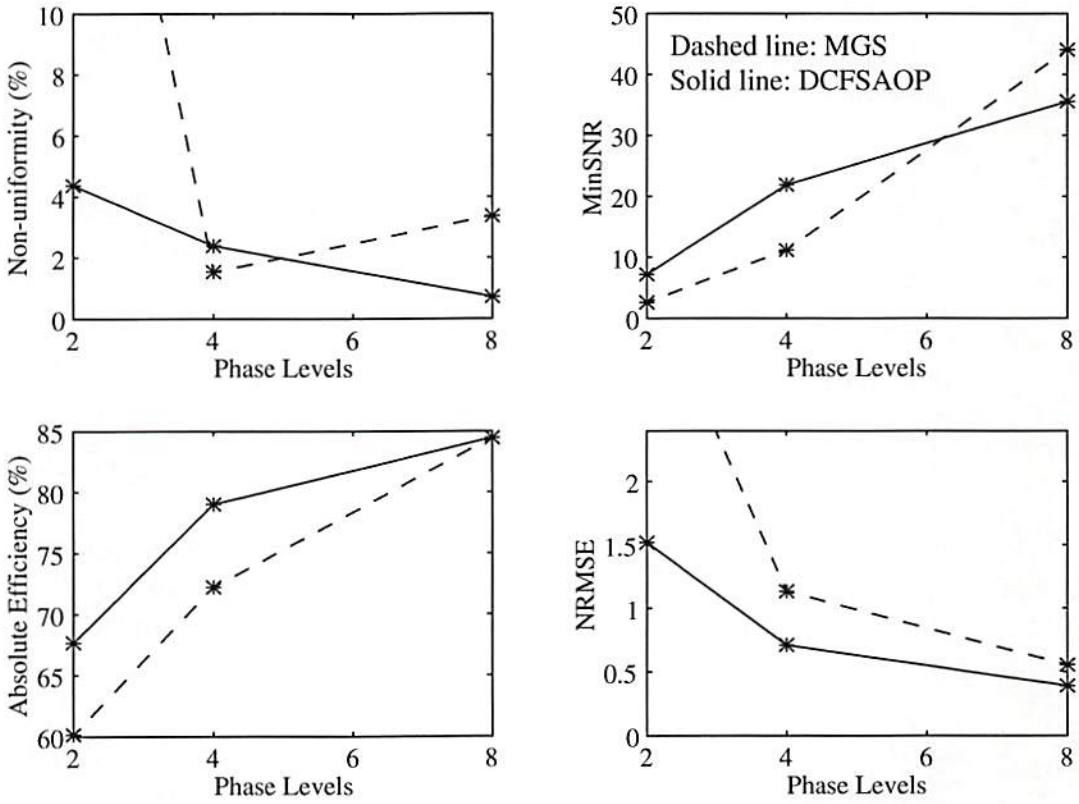

Figure 8-15: DOE performance comparison as a function of phase levels for the 3 × 3 spot array generator with 8 × 8 phase elements in each period of the DOEs using the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithms with phase optimization.

ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*) as a function of phase levels for the case of 8×8 phase elements in each period of the DOE for the modified Gerchberg-Saxton algorithm and dual-cost-function simulated annealing algorithm with phase delay optimization. Figure 8-16 shows non-

uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*) as a function of phase levels for the case of 16×16 phase elements in each period of the DOE for the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithm with phase delay optimization. From Figure 8-15 and Figure 8-16, we can see that the dual-cost-function simulated annealing algorithm with phase delay optimization can provide substantial performance improvement over the modified Gerchberg-Saxton algorithm. For a given

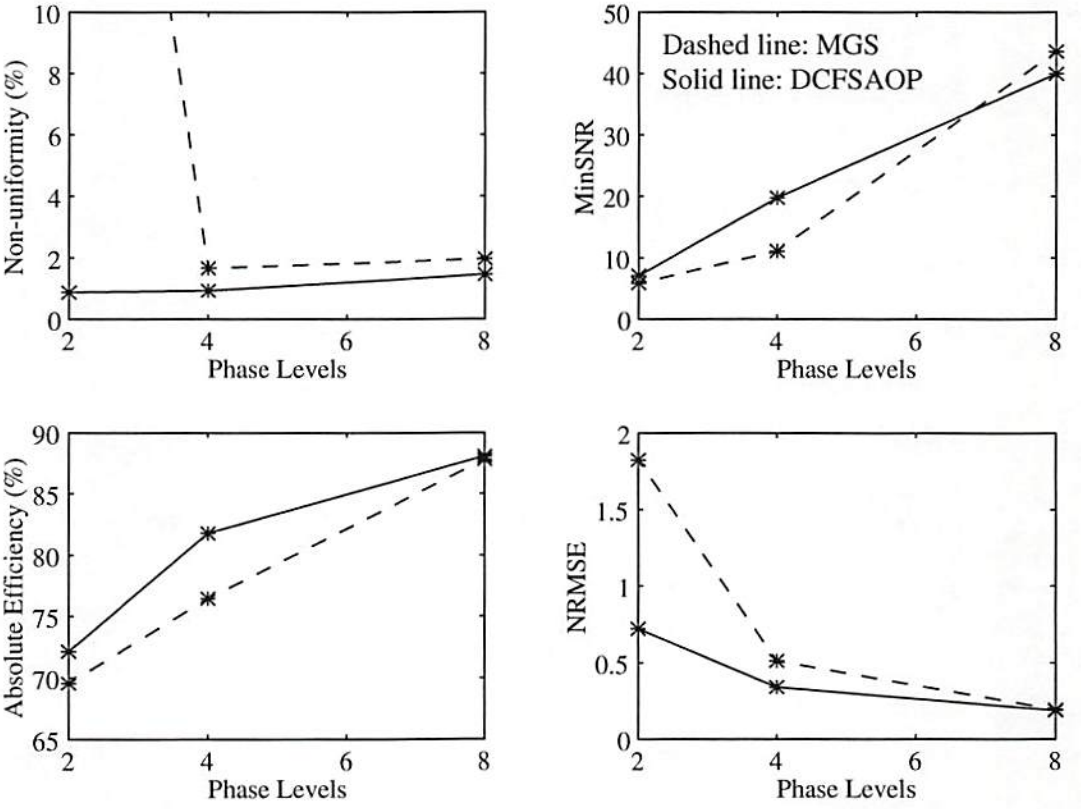

Figure 8-16: DOE performance comparison as a function of phase levels for the 3×3 spot array generator with 16×16 phase elements in each period of the DOEs using the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithms with phase optimization.

space-bandwidth-product of the DOE, the improvement on NRMSE is most significant for the two-phase-level case and least significant for the eight-phase-level case. In the cases that the substantial improvement are seen, the optimized independent phase delays for the DOEs are quite different from the non-optimized independent phase delays (e. g., the third or sixth row in Table 8-6). On the other hand, when the optimized independent phase delays for the DOEs are very close to the non-optimized independent phase delays, the performance improvement becomes marginal (e. g., the last row in Table 8-6). Note that, as can be seen from the sixth and thirteen rows of Table 8-6, the four-phase-level DOE with a space-bandwidth-product of 64 (8×8 phase elements in each period of the DOE) designed using the dual-cost-function simulated annealing algorithm with phase delay optimization has much higher minimum signal-to-noise ratio, and higher diffraction efficiency (2.5% more) than the four-phase-level DOE with a space-bandwidth-product of 256 (16×16 phase elements in each period of the DOE) designed using the modified Gerchberg-Saxton algorithm. Therefore, for this particular case, we can use less space-bandwidth-product in each period of the DOE and still obtain the DOE with better minimum signal-to-noise ratio and diffraction efficiency.

In Figure 8-17, we plot the optimized phase delays for the DOEs listed in Table 8-6. Note that the maximum phase delay for the DOE in the fourth and tenth rows of Table 8-6 are much less than the maximum phase delay of the non-optimized approach (which is $3\pi/2$). Furthermore, the optimized phase delays for DOE in the sixth row of Table 8-6 are approximately equally distributed between 0 and 1.14π with a phase delay step of roughly 0.38π. This can also be seen clearly in Figure 8-17. In this case, we have

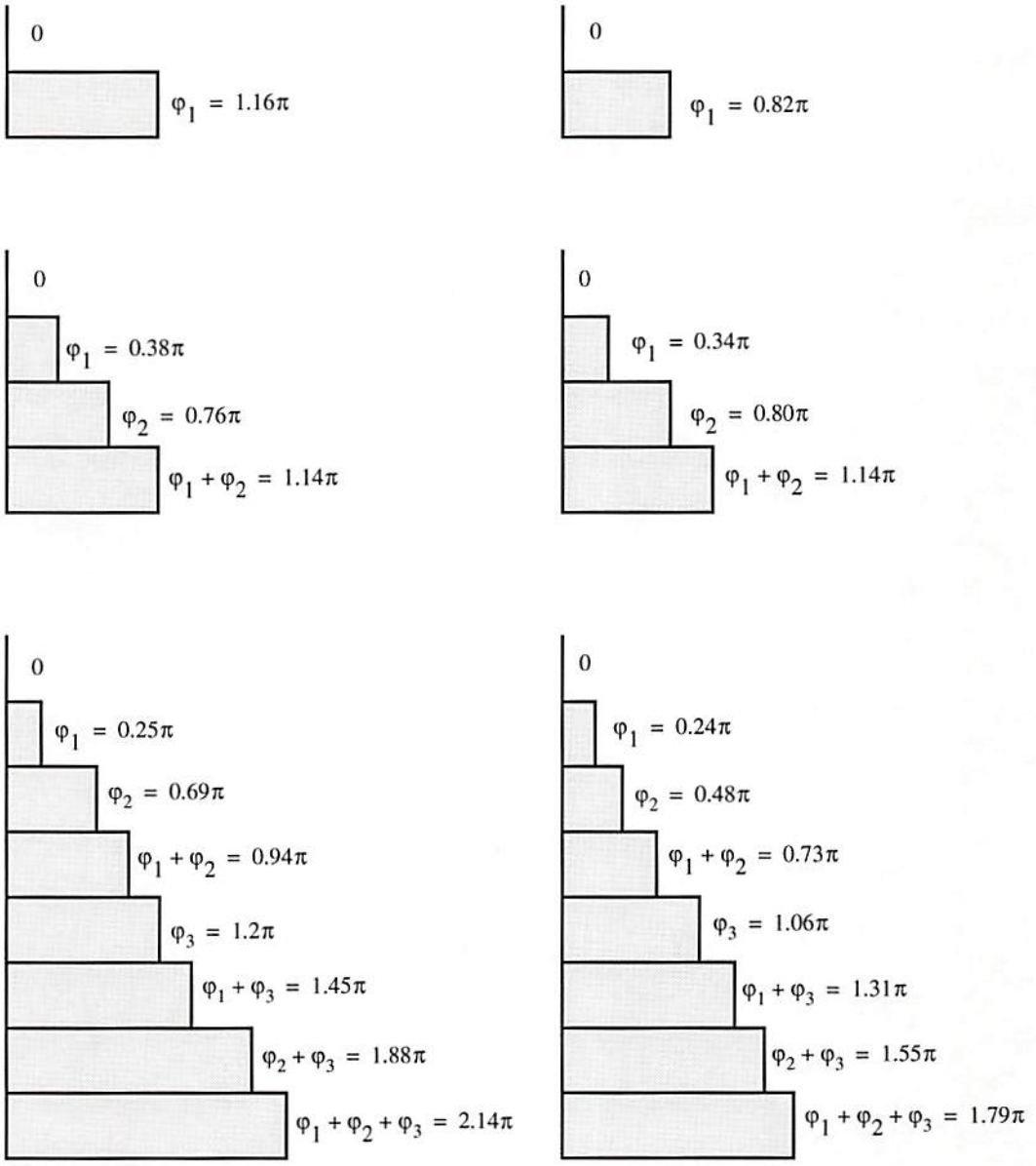

Figure 8-17: Phase delays for the DOEs in Table 8-6. On the left-hand-side are the optimized phase delays for the case of 8×8 phase elements in each period of the DOEs and the optimized phase delays for the case of 16×16 phase elements in each period of the DOEs are on the right-hand-side.

basically found a solution with the same fixed number of phase levels but finer phase steps that can better approximate the ideal DOE phase profile and improve the DOE performance. On the other hand, the maximum phase delay for the DOE in the ninth row of Table 8-6 has a value of 2.14π and is substantially larger than the maximum phase delay of the non-optimized approach (which is $7\pi/4$). This is quite different from the case of conventional DOEs, in which the maximum phase delay is always less than 2π. Recall that the maximum phase delay is just the sum of all the independent phase delays, and because the optimization algorithm optimizes each of the independent phase delays in the range between 0 and 2π, a maximum phase delay value greater than 2π is possible although the effective phase delay is the same as the remainder of that phase delay divided by 2π.

We have also redesigned DOEs for the 3×3 triangular connection pattern with 8×8 or 16×16 phase elements in each period of the DOE and two, four, and eight phase levels. Table 8-7 lists the performance of the DOE designed using the modified Gerchberg-Saxton (MGS) algorithm and the dual-cost-function simulated annealing algorithm without phase delay optimization (DCFSA) and with phase delay optimization (DCFSAOP) in terms of normalized non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*). The optimized independent phase delays are listed at the end of each row of Table 8-7. By comparing the data for dual-cost-function simulated annealing approaches in Table 8-7, we can see that the phase delay optimization approach generally performs better than the non-optimized approach for this target pattern. Figure 8-18 shows

Table 8-7: Performance of the DOE designed using the modified Gerchberg-Saxton (MGS) algorithm and the dual-cost-function simulated annealing algorithm with and without phase delay optimization (DCFSA and DCFSAOP, respectively) for 3×3 triangular pattern

J	Z	Algorithm	NU (%)	MinSNR	η (%)	NRMSE	Independent Phase Delays
8	2	MGS	20.7	3.6	82.0	1.9381	π
8	2	DCFSA	12.2	5.9	73.2	1.1364	π
8	2	DCFSAOP	3.0	7.2	69.6	1.0369	0.73π
8	4	MGS	8.3	6.4	73.6	1.1024	$\pi/2, \pi$
8	4	DCFSA	9.2	7.4	73.9	0.9914	$\pi/2, \pi$
8	4	DCFSAOP	4.9	9.0	83.6	0.5508	0.61π, 1.66π
8	8	MGS	3.2	7.9	83.9	0.6129	$\pi/4, \pi/2, \pi$
8	8	DCFSA	3.2	8.1	84.4	0.5963	$\pi/4, \pi/2, \pi$
8	8	DCFSAOP	3.7	18.3	87.0	0.4106	0.18π, 0.41π, 1.5π
16	2	MGS	19.6	5.3	82.0	0.7157	π
16	2	DCFSA	9.1	5.8	77.3	0.4999	π
16	2	DCFSAOP	2.3	6.5	73.7	0.4788	0.71π
16	4	MGS	5.2	11.5	80.2	0.3583	$\pi/2, \pi$
16	4	DCFSA	3.2	11.8	80.5	0.3752	$\pi/2, \pi$
16	4	DCFSAOP	1.2	13.3	85.5	0.2357	0.69π, 1.7π
16	8	MGS	1.9	12.9	86.4	0.2516	$\pi/4, \pi/2, \pi$
16	8	DCFSA	1.9	12.1	86.8	0.2444	$\pi/4, \pi/2, \pi$
16	8	DCFSAOP	1.6	20.0	89.9	0.2117	0.33π, 0.61π, 1.84π

normalized non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*) as a function of phase levels for the case of 8×8 phase elements in each period of the DOE for the modified Gerchberg-Saxton algorithm and dual-cost-function simulated annealing algorithm with

phase delay optimization. It is obvious that, for a given number of phase levels, the dual-cost-function simulated annealing algorithm with phase delay optimization generally performs better than the already modified Gerchberg-Saxton algorithm for DOEs with 8×8 phase elements in each period of the DOE. As can be seen from the sixth and fourth rows of Table 8-7, the optimized-phase-delay DOE can even provide 10% improvement on the diffraction efficiency with better non-uniformity, minimum signal-to-noise ratio, and normalized-root-mean-squared-error than the modified Gerchberg-Saxton algorithm. Furthermore, as can be seen from Table 8-7, the four-phase-level DOE with 8×8 phase elements in each period of the DOE designed using the dual-cost-function simulated annealing algorithm with phase delay optimization has only slightly higher non-uniformity, higher minimum signal-to-noise ratio, equivalent diffraction efficiency, and lower normalized-root-mean-squared-error than the eight-phase-level DOE with 8×8 phase elements in each period of the DOE designed using the modified Gerchberg-Saxton algorithm. This means that we can reduce the fabrication complexity (use four rather than eight phase levels) while still producing comparable DOE performance.

Figure 8-19 shows normalized non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*) as a function of phase levels for the case of 16×16 phase elements in each period of the DOE for the modified Gerchberg-Saxton algorithm and dual-cost-function simulated annealing algorithm with phase delay optimization. From Figure 8-18 and Figure 8-19, we can see that the dual-cost-function simulated annealing algorithm with phase delay optimization still provides substantial performance improvement over the

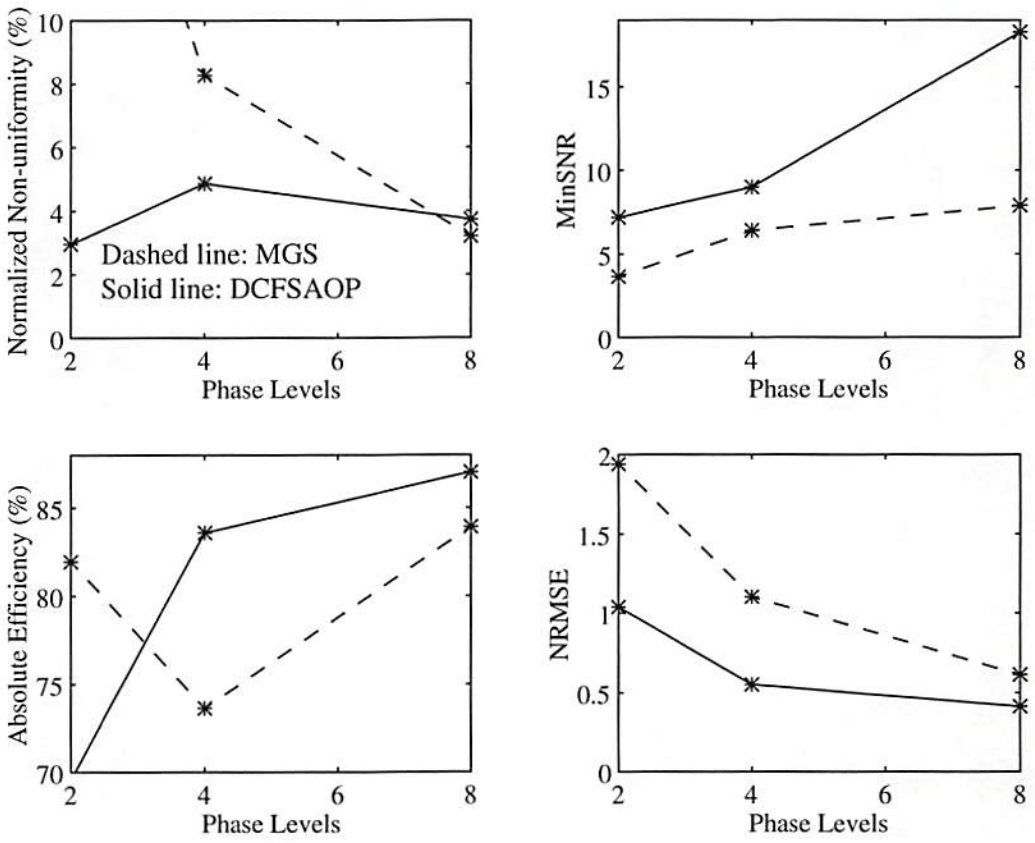

Figure 8-18: DOE performance comparison as a function of phase levels for the 3 × 3 triangular pattern with 8 × 8 phase elements in each period of the DOEs using the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithms with phase optimization.

modified Gerchberg-Saxton algorithm for the case of the 3×3 triangular connection pattern. Note that, for the two-phase-level case, the dual-cost-function simulated annealing algorithm with phase delay optimization found the DOE with much lower normalized non-uniformity, higher minimum-signal-to-noise-ratio, and lower normalized-root-mean-squared-error but lower absolute diffraction efficiency than the DOE that the modified Gerchberg-Saxton obtained. For the cases of four and eight phase levels, the dual-cost-function simulated annealing algorithm with phase delay optimization always

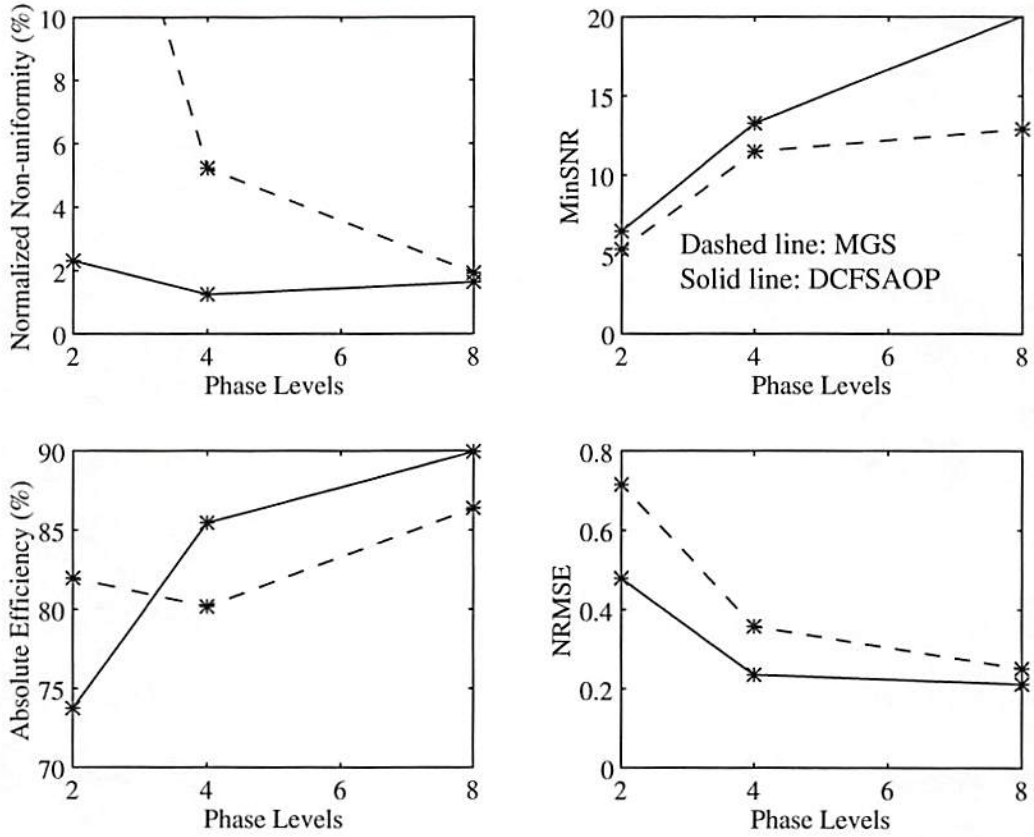

Figure 8-19: DOE performance comparison as a function of phase levels for the 3 × 3 triangular pattern with 16 × 16 phase elements in each period of the DOEs using the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithms with phase optimization.

found the DOE with better non-uniformity, minimum signal-to-noise ratio, and normalized-root-mean-squared-error, and absolute diffraction efficiency than the DOE that the modified Gerchberg-Saxton obtained. Note that, as can be seen from Table 8-7, the four-phase-level DOE with a space-bandwidth-product of 64 (8×8 phase elements in each period of the DOE) designed using the dual-cost-function simulated annealing algorithm with phase delay optimization has lower non-uniformity, comparable minimum signal-to-noise ratio, higher diffraction efficiency (3.4% more), and slightly worse normalized-root-mean-squared-error than the four-phase-level DOE with a space-

bandwidth-product of 256 (16×16 phase elements in each period of the DOE) designed using the modified Gerchberg-Saxton algorithm. Therefore, we can use one fourth of the space-bandwidth-product and still get comparable results by using the dual-cost-function simulated annealing algorithm with phase delay optimization.

In Figure 8-20, we plot the optimized phase delays for the DOEs listed in Table 8-7. Note that, for this 3×3 triangular connection pattern, the maximum phase delays for the binary-phase-level DOEs (either 8×8 or 16×16 phase elements in each period of the DOE) are smaller than the maximum phase delays of the non-optimized approach. On the other hand, the maximum phase delays for the four (or eight)-phase-level DOEs (either 8×8 or 16×16 phase elements in each period of the DOE) are larger than the maximum phase delays of the non-optimized approach.

In this thesis, we define dynamic range of a target pattern as the ratio of the maximum weight value over the minimum weight value in the target pattern. Using this definition, the triangular connection pattern has a dynamic range of four. It is common to see dynamic range of ten or more for connection patterns in neural networks. Next, we designed DOEs for target patterns with a high dynamic range using the dual-cost-function simulated annealing algorithm with phase delay optimization. For this purpose, we have formed a target pattern with a dynamic range of 10 (referred to as triangular2) as

$$Triangular2 = \begin{bmatrix} 0.1 & 0.5 & 0.1 \\ 0.5 & 1 & 0.5 \\ 0.1 & 0.5 & 0.1 \end{bmatrix}. \tag{8-21}$$

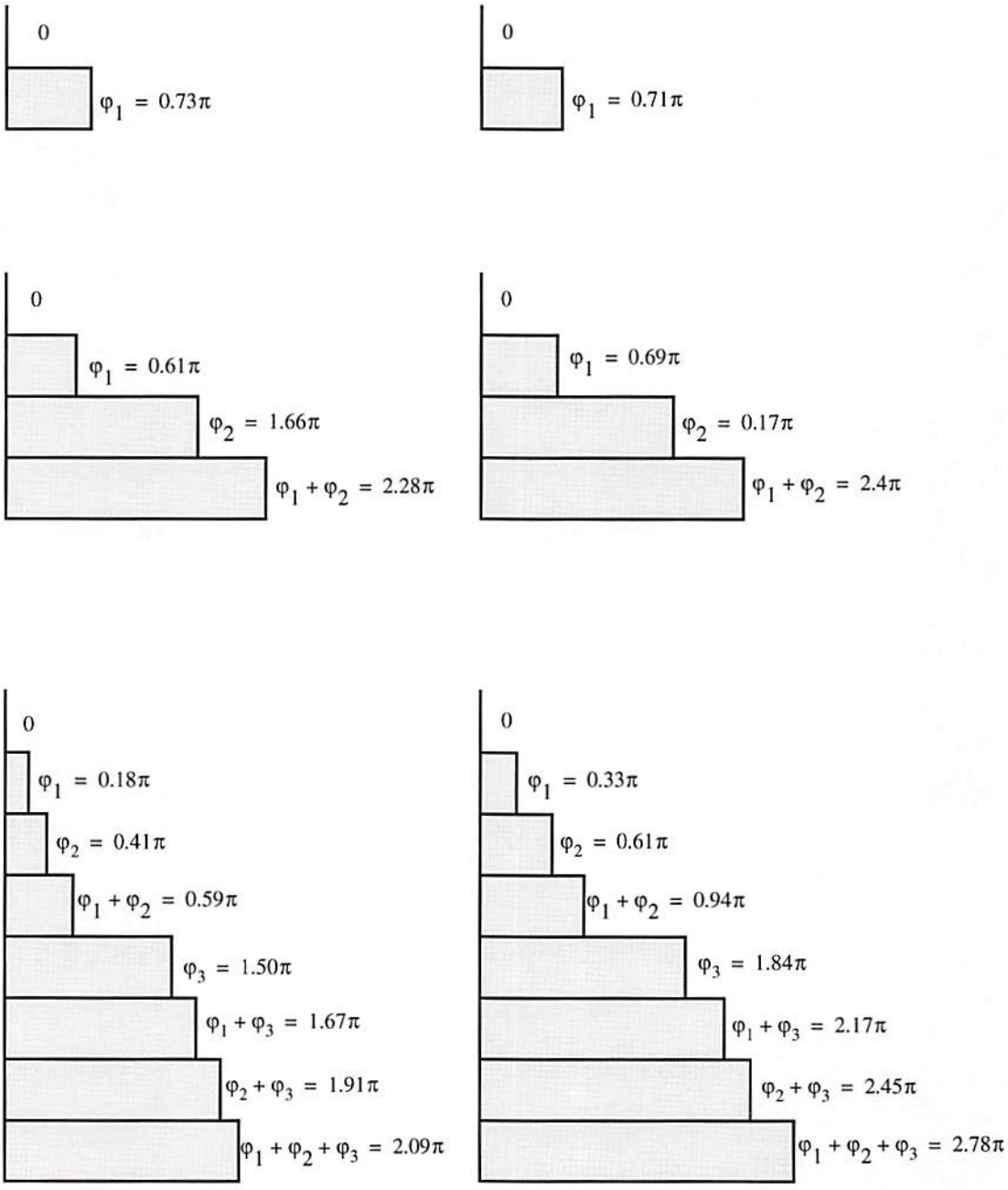

Figure 8-20: Phase delays for the DOEs in Table 8-7. On the left-hand-side are the optimized phase delays for the case of 8×8 phase elements in each period of the DOEs and the optimized phase delays for the case of 16×16 phase elements in each period of the DOEs are on the right-hand-side.

We have designed the DOEs for this high dynamic range weight pattern with 8×8 (or 16×16) phase elements in each period of the DOEs and two, four, and eight phase levels using the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithm with phase delay optimization. Table 8-8 lists the performance of the

Table 8-8: Performance of the DOE designed using the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithm with phase delay optimization for 3×3 triangular2 pattern

J	Z	Algorithm	NU (%)	MinSNR	η (%)	NRMSE	Independent Phase Delays
8	2	MGS	17.1	2.3	76.7	1.8359	π
8	2	DCFSAOP	5.1	5.2	80.8	0.6484	0.76π
8	4	MGS	4.4	3.8	76.6	0.9394	$\pi/2, \pi$
8	4	DCFSAOP	2.0	6.8	84.5	0.4975	$0.18\pi, 1.26\pi$
8	8	MGS	5.4	6.4	86.9	0.7059	$\pi/4, \pi/2, \pi$
8	8	DCFSAOP	5.9	7.4	87.9	0.4435	$0.15\pi, 0.42\pi, 1.58\pi$
16	2	MGS	19.9	3.6	80.3	0.5761	π
16	2	DCFSAOP	7.6	7.0	81.7	0.3358	1.24π
16	4	MGS	18.7	5.5	79.7	0.4792	$\pi/2, \pi$
16	4	DCFSAOP	1.9	6.8	86.9	0.2591	$0.31\pi, 0.6\pi$
16	8	MGS	6.0	4.9	87.9	0.2753	$\pi/4, \pi/2, \pi$
16	8	DCFSAOP	2.8	4.9	88.7	0.2419	$0.13\pi, 0.32\pi, 0.52\pi$

DOEs designed using the modified Gerchberg-Saxton algorithm and dual-cost-function simulated annealing algorithms with phase delay optimization in terms of normalized non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and

normalized-root-mean-squared-error (*NRMSE*). Table 8-8 also lists the optimized independent phase delays at the end of each row. In Figure 8-21, we have plotted the normalized non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*) as a function of phase levels for the case of 8×8 phase elements in each period of the DOE for the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithm with phase delay optimization for this 3×3 high dynamic range target pattern. From Table 8-8 and Figure 8-21, it is apparent that, for a given number of phase levels, the dual-cost-function simulated annealing algorithm with phase delay optimization generally performs much better than the already modified Gerchberg-Saxton algorithm for DOEs with 8×8 phase elements in each period of the DOE. Furthermore, in this particular case, the two-phase-level DOE designed using the dual-cost-function simulated annealing algorithm with phase delay optimization has only slightly higher non-uniformity, higher minimum signal-to-noise ratio, higher diffraction efficiency (4.2% more), and lower normalized-root-mean-squared-error than the four-phase-level DOE designed using the modified Gerchberg-Saxton algorithm. This means that we can reduce the fabrication complexity (use two rather than four phase levels) while still generally outperforming the MGS-designed DOE with higher fabrication complexity.

Figure 8-22 shows the normalized non-uniformity, minimum signal-to-noise ratio (*MinSNR*), absolute diffraction efficiency, and normalized-root-mean-squared-error (*NRMSE*) as a function of phase levels for the case of 16×16 phase elements in each period of the DOE for the modified Gerchberg-Saxton algorithm and the dual-cost-

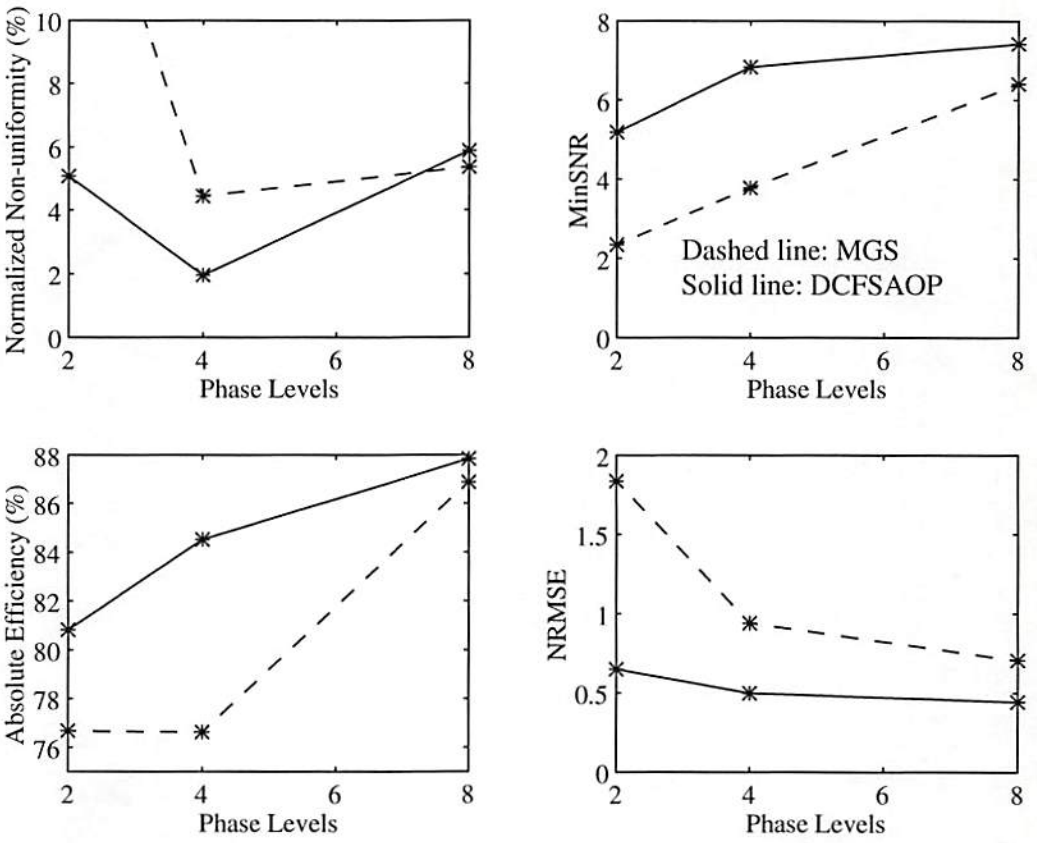

Figure 8-21: DOE performance comparison as a function of phase levels for the 3×3 triangular2 connection pattern (dynamic range of 10) with 8×8 phase elements in each period of the DOEs using the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithms with phase delay optimization.

function simulated annealing algorithm with phase delay optimization. From Table 8-8, Figure 8-21, and Figure 8-22, it is apparent that the dual-cost-function simulated annealing algorithm with phase delay optimization performs much better overall than the already modified Gerchberg-Saxton algorithm for a given value of space-bandwidth-product and number of phase levels. Furthermore, for a fixed space-bandwidth-product in each period of the DOE, the DOE designed using the dual-cost-function simulated

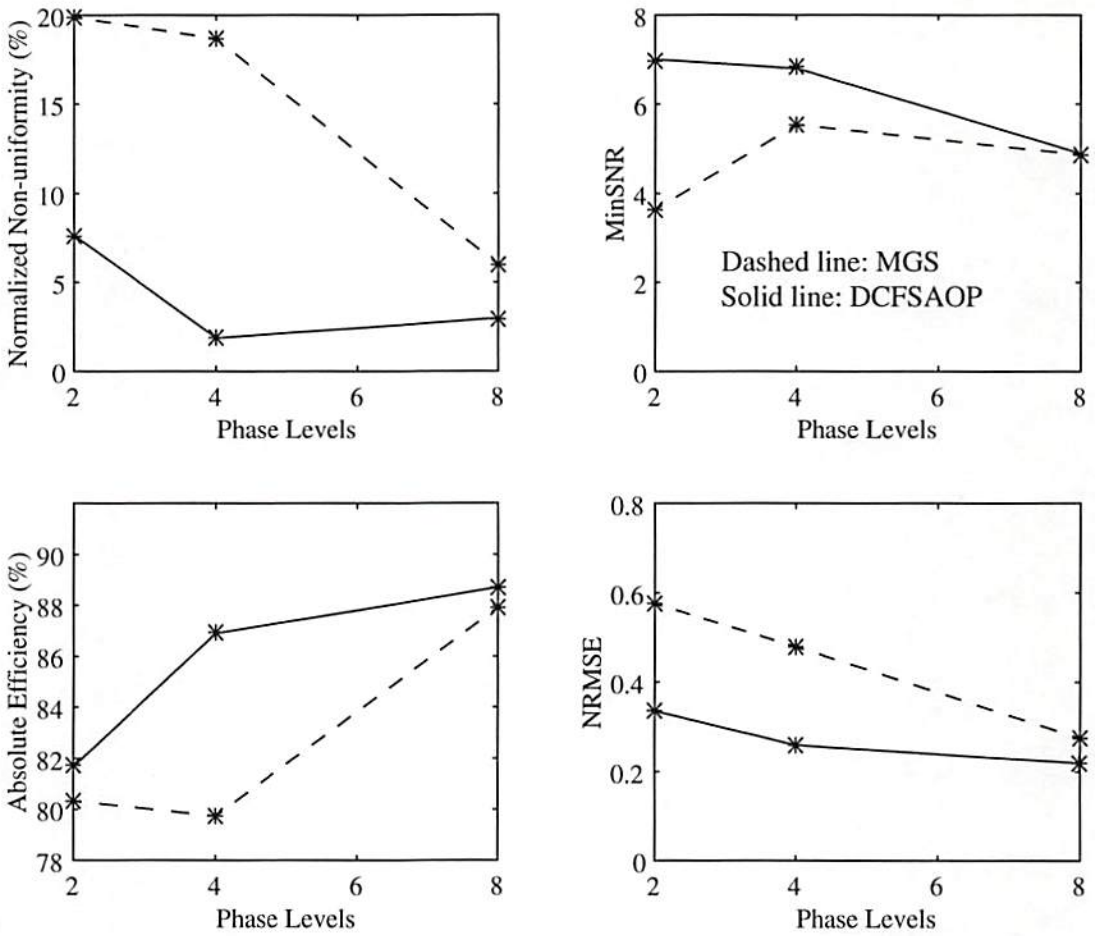

Figure 8-22: DOE performance comparison as a function of phase levels for the 3×3 triangular2 connection pattern (dynamic range of 10) with 16×16 phase elements in each period of the DOEs using the modified Gerchberg-Saxton algorithm and the dual-cost-function simulated annealing algorithms with phase delay optimization.

annealing algorithm with phase delay optimization generally performs either comparable with or better than the DOE designed using the modified Gerchberg-Saxton algorithm and twice as many phase levels. For example, the two-phase-level DOE designed using the dual-cost-function simulated annealing algorithm with phase delay optimization has lower non-uniformity, higher minimum signal-to-noise ratio, higher diffraction efficiency (2%

more), and lower normalized-root-mean-squared-error than the four-phase-level DOE designed using the modified Gerchberg-Saxton algorithm for the case of 16×16 phase elements in each period of the DOE. Also, for a fixed number of phase levels, the DOE designed using the dual-cost-function simulated annealing algorithm with phase delay optimization generally performs either comparable with or better than the DOE designed using the modified Gerchberg-Saxton algorithm and four times as many space-bandwidth-product in each period of the DOE. For example, the four-phase-level DOE with a space-bandwidth-product of 64 (8×8 phase elements in each period of the DOE) designed using the dual-cost-function simulated annealing algorithm with phase delay optimization has much lower non-uniformity, higher minimum signal-to-noise ratio, higher diffraction efficiency (4.8% more), and comparable normalized-root-mean-squared-error than the four-phase-level DOE with a space-bandwidth-product of 256 (16×16 phase elements in each period of the DOE) designed using the modified Gerchberg-Saxton algorithm. Therefore, by using the dual-cost-function simulated annealing algorithm with phase delay optimization, we can either use fewer phase levels (reducing fabrication complexity) or fewer phase elements in each period of the DOE (reducing space-bandwidth-product) while producing DOEs with comparable or even better performance.

In Figure 8-23, we plot the optimized phase delays for the DOEs listed in Table 8-8. Note that most of the maximum phase delays for the DOE in Figure 8-23 are much less than the maximum phase delay of the non-optimized approach. On the other hand, for this particular pattern, only two of the maximum phase delay in Figure 8-23 are substantially larger than the maximum phase delay of the non-optimized approach. Also note that the

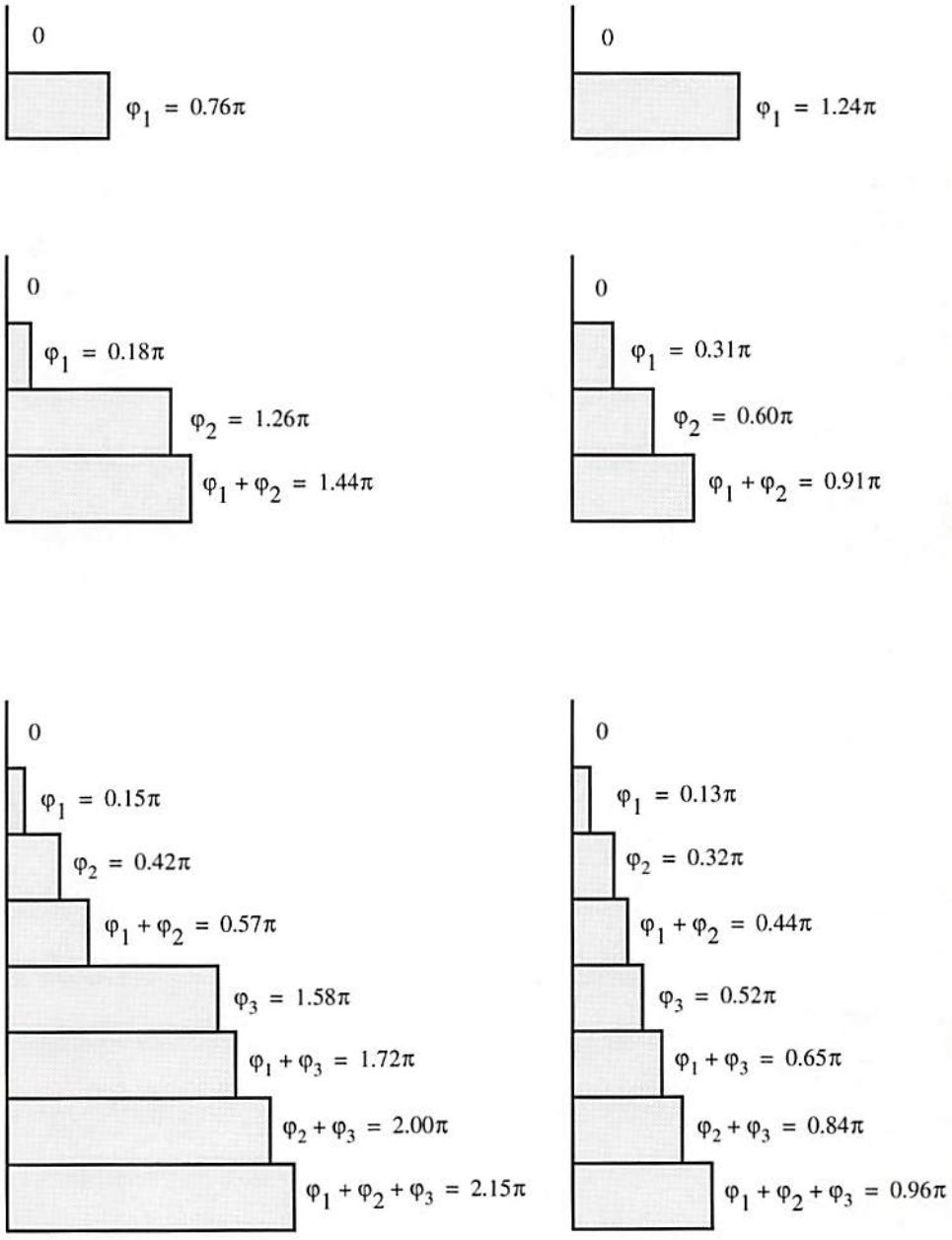

Figure 8-23: Phase delays for the DOEs in Table 8-8. On the left-hand-side are the optimized phase delays for the case of 8×8 phase elements in each period of the DOEs and the optimized phase delays for the case of 16×16 phase elements in each period of the DOEs are on the right-hand-side.

optimized phase delays for DOE in the tenth row of Table 8-8 are approximately equally distributed between 0 and 0.91π with a phase delay step of roughly 0.30π. This can also be seen clearly in Figure 8-23. In this case, we have found a solution with the same fixed number of phase levels but finer phase steps that can better approximate the ideal DOE phase profile and improve the DOE performance.

Although the approach of optimizing the independent phase delays has been proven useful in improving the performance of the DOEs, there are cases that the conventional equally spaced phase delays are the optimum design approach. For example, let's consider the following deflector target pattern:

$$Deflector = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}. \tag{8-22}$$

Basically, this pattern will deflect the input light beam into the +1 diffraction order in the horizontal direction. It can be shown that the optimum phase delays for this target pattern are just the phase delays equally spaced between 0 and $\frac{2\pi \cdot (Z-1)}{Z}$. Also, it can be shown that the optimum diffraction efficiency, η_{opt}, for this deflector can be expressed as

$$\eta_{opt} = \sin c^2(\frac{1}{Z}) \cdot 100\%. \tag{8-23}$$

The minimum signal-to-noise ratio for the optimum DOE will be infinite since all the noise diffraction orders have zero energy and all the light energy goes to the desired diffraction order and its signal sidelobes. Detailed derivation for this stepped, quantized blazed grating

can be found in several reference entries from Taghizadeh and Turunen's DOE review paper [40].

To test whether the phase-delay-optimization algorithm can find the theoretical optimum phase delays, we have designed four-(and eight)-phase-level DOEs for this deflector target pattern using the dual-cost-function simulated annealing algorithm with phase delay optimization. Table 8-9 shows the minimum signal-to-noise ratio, absolute diffraction efficiency, and independent phase delays for DOEs designed using the dual-cost-function simulated annealing algorithm with phase delay optimization and the theoretically calculated values (uniformity and NRMSE are not listed since there is only one non-zero weight value in the target pattern). It can be seen from Table 8-9 that the

Table 8-9: DOE performance comparison between the theoretical optimum and the dual-cost-function simulated annealing algorithm with phase delay optimization for the deflector pattern

J	Z	Algorithm	MinSNR	η (%)	Independent Phase Delays
4	4	Optimum	∞	81.07	$\pi/2, \pi$
4	4	DCFSAOP	4558	81.03	$0.514\pi, 1.012\pi$
8	8	Optimum	∞	94.96	$\pi/4, \pi/2, \pi$
8	8	DCFSAOP	5626	94.93	$0.257\pi, 0.496\pi, 1.001\pi$

dual-cost-function simulated annealing algorithm with phase delay optimization does find DOEs with phase delays that are very close to the optimum ones. Since each of the independent phase delays is allowed to be any value between 0 and 2π in each iteration of the annealing algorithm, therefore, it is unlikely that the algorithm will find the exact optimum phase delays.

In conclusion, the dual-cost-function simulated annealing algorithm with phase delay optimization almost always performed better than the dual-cost-function simulated annealing algorithm without phase delay optimization and modified Gerchberg-Saxton algorithms for a given value of space-bandwidth-product and number of phase levels. However, The improvement on DOE performance does vary for different target patterns, phase levels used, and number of phase elements in one period of the DOE. In some cases, for a fixed space-bandwidth-product in each period of the DOE, the DOE designed using the dual-cost-function simulated annealing algorithm with phase delay optimization performed better than the DOE designed using the other algorithms and twice as many phase levels. In some cases, the DOE with a fixed number of phase levels designed using the dual-cost-function simulated annealing algorithm with phase delay optimization even performed better than the DOE designed using the other algorithms and four times the space-bandwidth-product in each period of the DOE. Therefore, If we keep the space-bandwidth-product in each period of the DOE and number of phase levels unchanged, then we can use the dual-cost-function simulated annealing algorithm with phase delay optimization to obtain DOEs with better performance. On the other hand, we can either use fewer phase levels (reducing fabrication complexity) or fewer phase elements in each period of the DOE (reducing space-bandwidth-product) and use the dual-cost-function simulated annealing algorithm with phase delay optimization to obtain DOEs with generally comparable or, in some cases, even better performance. These conclusions are necessarily based on simulations of specific examples of DOE target patterns. And while it is unknown to what degree the results generalize, the target patterns were chosen to

represent a variety of possibilities. Since similar results were obtained for all target patterns that we tried, we would expect the results to generalize to other target pattern of similar nature.

Chapter 9

Conclusions and Future Research Directions

9.1 Conclusions

This dissertation has discussed DOE-based space-variant interconnections for multilayer feed-forward three-dimensional computation structures. Two DOE-based space-variant interconnection architectures, fully connected and limited-fanout, were investigated and compared. Since the reconstruction of a diffractive optical element is sensitive to illumination wavelength and to etch depth error due to fabrication, the effects of incorrect wavelength and etch depth on DOE reconstruction were studied in this dissertation. Because the performance of diffractive optical elements is critical for the interconnection systems, this dissertation has investigated two design algorithms, Gerchberg-Saxton and simulated annealing, and, based on these two algorithms, developed new design approaches that are substantially better than the original algorithms.

We studied the properties of DOEs and established a reconstruction model of DOEs in Chapter 3. The relations between a diffractive optical element and its reconstruction were derived. We have also categorized the reconstructed diffraction orders as signal orders, signal sidelobes, and spurious diffraction orders then derived the relations between signal orders and signal sidelobes. Based on this categorization, we have established a

reconstruction model, and have incorporated into this model a course approximation to the non-ideal behavior of the DOE-system reconstruction lens. From this reconstruction model, we identified and classified possible crosstalk components for the interconnection crosstalk analysis in the following chapters.

The DOE-based fully connected space-variant interconnection architecture was studied in Chapter 4. The expressions for propagation length, system lateral dimension, system aspect ratio, *f*-number of the reconstruction optics, system volume, interconnection density, and space-bandwidth product requirement were derived for a cascadable fully connected space-variant interconnection system by choosing an appropriate set of independent parameters. We then analyzed and simulated the interconnection crosstalk for a layer of fully connected space-variant interconnections with 100×100 input nodes and 100×100 output nodes. The simulation showed that the interconnection crosstalk level in such a system is likely tolerable for many applications.

The DOE-based limited-fanout space-variant interconnection architecture was studied in Chapter 5. The expressions for propagation length, system lateral dimension, system aspect ratio, *f*-number of the reconstruction optics, system volume, interconnection density, and space-bandwidth product requirement were derived for a cascadable limited-fanout space-variant interconnection system by choosing an appropriate set of independent parameters. By comparing the system characteristics of the two architectures, we concluded that the limited-fanout architecture can provide a more compact system size for large scale interconnection systems than the fully connected architecture. However, from the simulation for a layer of limited-fanout space-variant

interconnections with 128×128 input nodes, 128×128 output nodes, and 5×5 nearest neighbor weighted interconnections from each input node to the array of output nodes, we conclude that the interconnection crosstalk for the DOE-based limited-fanout architecture is likely too high for use in many applications because of noise from the reconstructions of the surrounding diffractive optical elements. Therefore, we have developed a crosstalk reduction technique based on a modified design for diffractive optical elements, which rearranges the reconstruction pattern such that less noise lands on each detector region. Simulations showed that the technique can substantially reduce the crosstalk for the limited-fanout system. It was found that the crosstalk reduction technique can also be used to reduce crosstalk level and system propagation length at the same time.

The reconstruction of a diffractive optical element will change when the illumination wavelength changes. In Chapter 6, we have studied the effect of an incorrect illumination wavelength on the reconstruction of multiple-phase-level DOEs. The expressions for DOE reconstruction under an incorrect illumination wavelength were derived and the effect was simulated on DOEs (with two, four, and eight phase levels) designed for a 3×3 spot array generator and a 3×3 triangular connection pattern. From the simulation, we concluded that the effect of an incorrect illumination wavelength on the reconstruction of DOEs with more than two phase levels is generally dependent on the phase distributions of the DOEs. This effect also varies for different diffraction orders from a given DOE. On the other hand, for any off-axis diffraction order, the effect of an incorrect illumination wavelength on the reconstruction of $\{0, \pi\}$ -binary-phase-level DOEs is not dependent on the phase distributions of the DOEs and this effect is the same for all the off-axis

diffraction orders. The effect of an incorrect illumination wavelength on the reconstruction for any off-axis diffraction order depends on the ratio of the designed wavelength and actual illumination wavelength. The effect for the zeroth diffraction order is DOE-dependent and the diffraction intensity will always increase if the wavelength is incorrect. The above special properties for $\{0, \pi\}$ -binary-phase-level DOEs were simulated and theoretically proved. We have also analyzed the effect of an incorrect illumination wavelength on the reconstruction for binary-phase-level DOEs with phase delays other than 0 and π. The effect for these unconventional $\{0, \varphi\}$ -binary-phase-level DOEs is somewhat different from the effect for the conventional binary-phase-level DOEs.

The reconstruction of a diffractive optical element will change when the etch depth in fabrication is incorrect. In Chapter 7, we have studied the effect of etch depth error on the reconstruction of multiple-phase-level DOEs, under the assumption that the (possible incorrect) depth of each etch step is uniform across the DOE. The expressions for DOE reconstruction under etch depth error was derived. For the case of $\{0, \pi\}$ -binary-phase-level DOEs, the effect of etch depth error on DOE reconstruction is essentially identical to the effect of an incorrect illumination wavelength on DOE reconstruction. We have also derived the expression for the combined effects of an incorrect illumination wavelength and etch depth error on the reconstruction of binary-phase-level DOEs, then demonstrated a way to adjust the illumination wavelength to cancel the effect of etch depth error for DOE reconstruction. A similar analysis for unconventional $\{0, \varphi\}$ -binary-phase-level DOEs has also been conducted. For the case of DOEs with more than two phase levels,

we have shown that the effect of a constant percentage depth error in all etch steps is essentially the same as the effect of a wavelength change. The expression for the combined effect of an incorrect illumination wavelength and a constant percentage depth error in all etch steps was also derived and the way to cancel the etch-depth error by adjusting the illumination wavelength was similarly described. Finally, we have studied and simulated the effect on the reconstruction of four-phase-level DOEs for the case of etch depth error that varies in percentage from etch step to etch step.

In Chapter 8, we first investigated the Gerchberg-Saxton algorithm for DOE design, then developed a modified Gerchberg-Saxton algorithm based on noise-adjustment scheduling and phase-quantization scheduling for better DOE performance. The performance of the modified algorithm was verified by designing DOEs with two, four, eight, and sixteen phase levels for the 3×3 spot array generator and 3×3 triangular connection pattern. Because both the original and modified Gerchberg-Saxton algorithms performed poorly for designing binary-phase-level DOEs, we have also studied another algorithm, simulated annealing. The effect of cost functions in simulated annealing algorithm was studied and a dual-cost-function simulated annealing algorithm was developed for improving the design performance. Finally, we have developed a new design approach that relieves unnecessary constraints and increases design freedom by allowing an unequally spaced quantization of phase delays in DOEs. This new approaches does increase the complexity of DOE design but generally does not increase the complexity of DOE fabrication. The principles of this new design approach were derived and a modified dual-cost-function simulated annealing algorithm was developed

to implement this new design approach. From the performance of the DOEs designed using the phase-optimizing dual-cost-function simulated annealing algorithm, it is clear that the new design approach yields DOE designs with substantially better performance than the conventionally designed DOEs.

9.2 Future Research Directions

In order to provide the weighted interconnections needed for implementation of artificial neural networks based on arrays of these sub-DOEs in the limited-fanout interconnection system, an array of Fourier transform microlenses are used for the DOE reconstruction optics. Because a Fourier transform microlens is also a phase-only element, it is possible to integrate the reconstruction optics with the sub-DOE array to form a hybrid diffractive element array. Each of these hybrid diffractive elements will have both the weighted fanout function and the lens function in it. This integration can provide the advantages of reducing the number of elements and easing the alignment requirements in a multilayer system. Figure 9-1 shows a conceptual optoelectronic implementation of a multilayer feed-forward neural network with the sub-DOE array and reconstruction optics being integrated together into the hybrid optical element array. On the other hand, integrating the reconstruction optics with the diffractive optical elements to form hybrid optical elements will put additional design constraints on the sub-DOE array. Therefore, the design algorithm needs to be modified to incorporate these additional constraints for these hybrid optical elements. This is an interesting direction for future research.

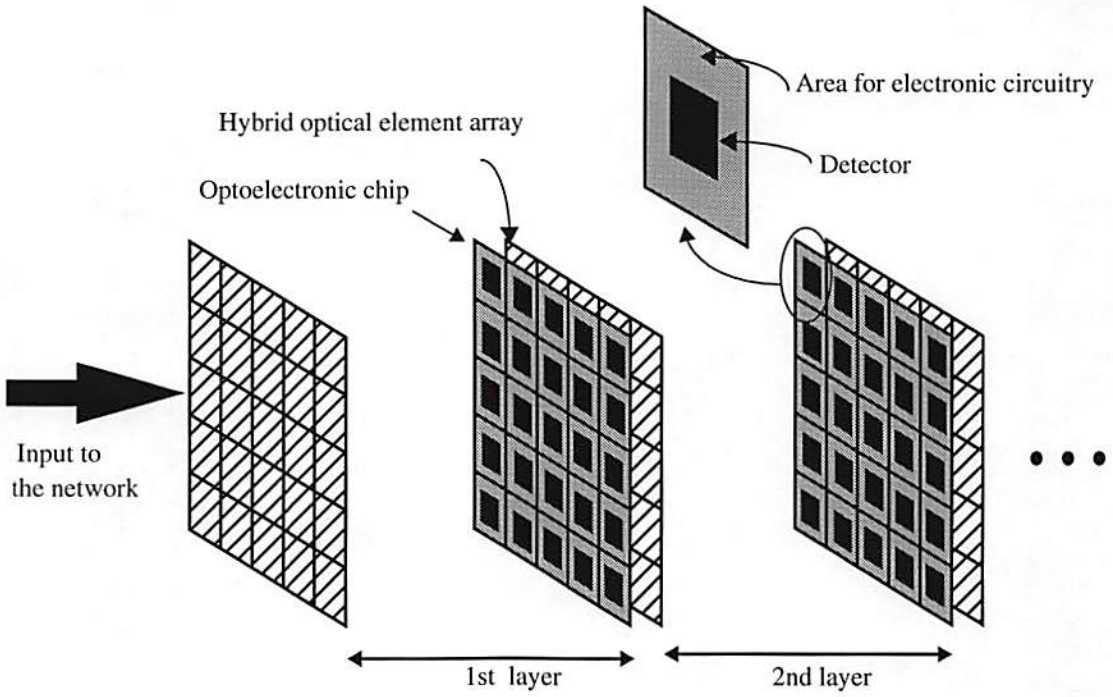

Figure 9-1: Schematic diagram for optoelectronic implementation of multilayer feed-forward neural networks. The sub-DOE array and reconstruction optics have been integrated together to the hybrid optical element array. Note that the light emitter array (or modulator array and readout-illumination optics) on the back side of each optoelectronic chip is not shown. Note that the figure is not to scale.

We have investigated space-variant fully connected and limited-fanout systems. The fully connected interconnection system has the advantage of global interconnection ability but requires a large space-bandwidth product for the sub-DOE array. The limited-fanout interconnection system has the advantage of shorter propagation length and higher connection density than the fully connected system, but only has local interconnection ability. There are other possible interconnection architectures such as space-invariant, hybrid space-variant/space-invariant, or hybrid fully-connected/limited-fanout architectures. A hybrid fully-connected/limited-fanout system organizes the neuron units

into sub-group, is fully connected within each sub-group, and is limited-fanout connected between adjacent sub-groups. This hybrid system is a compromise that provides higher connectivity than the limited-fanout system and shorter propagation lengths than the fully connected system. Figure 9-2 shows one possible realization of a hybrid fully-connected/ limited-fanout interconnection system.

When all the connection patterns of input nodes are essentially the same except that the connection patterns are shifted relative to each other, the interconnection system is space-invariant (as opposed to space-variant, as described above and in previous chapters). Because all the connection patterns are the same, we can use only one diffractive optical element for all the input nodes and greatly reduce the space-bandwidth product requirement in the system. Some appropriate reconstruction optics can then be used to shift the connection patterns. Figure 9-3 shows one possible realization of a space-invariant interconnection system. Usually, a space-invariant interconnection system can implement larger scale neural networks than a space-variant interconnection system. It is also possible to form a hybrid space-variant/space-invariant interconnection system. Such a system can combine the advantages of both variant and invariant systems.

Although the phase-optimizing dual-cost-function simulated annealing algorithm does perform substantially better than other algorithms, the algorithm requires many more computations and therefore is slow. By further analyzing the algorithm, it might be possible to speed up the algorithm by either removing unnecessary computations or other methods.

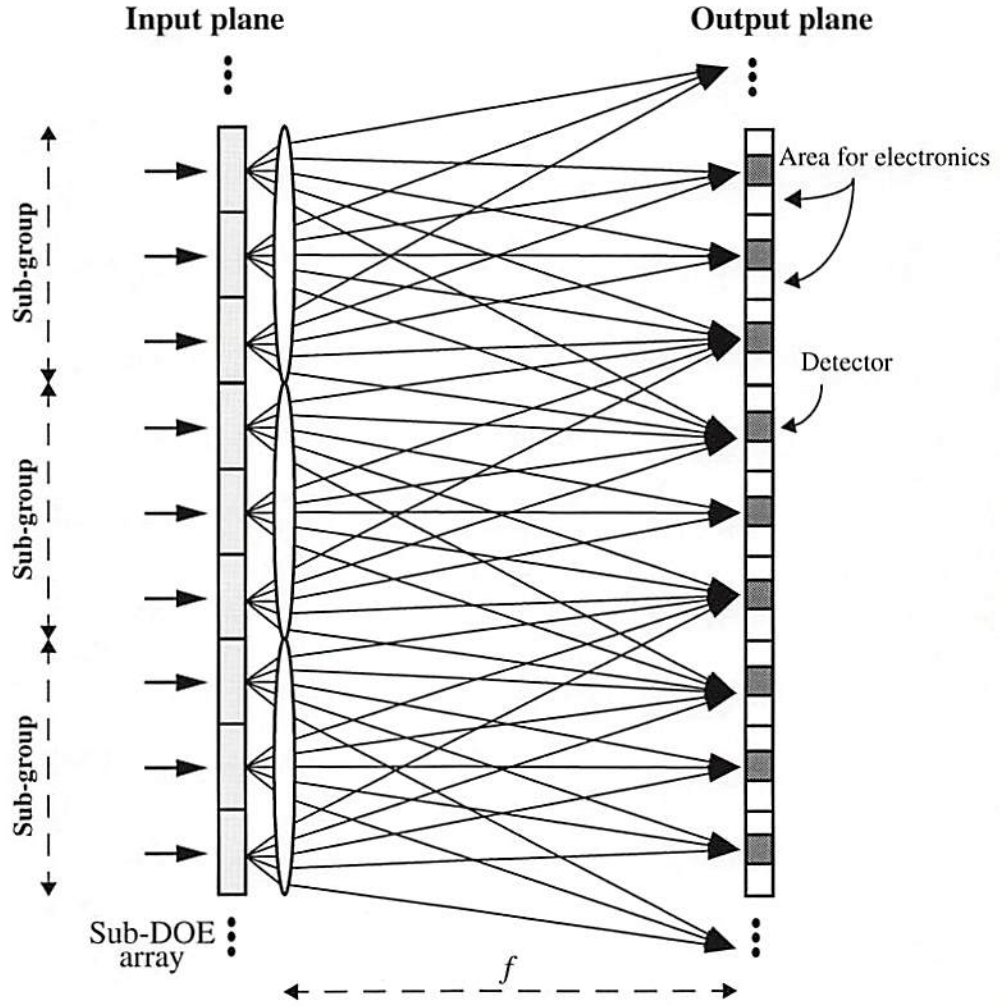

Figure 9-2: One possible realization of a hybrid fully-connected/limited-fanout interconnection system. It is fully connected within each neuron-unit sub-group and has limited fanout between adjacent sub-groups.

Usually the Gerchberg-Saxton algorithm can converge fast to a good DOE phase distribution when there is no phase quantization in the iteration (i. e., infinite number of phase levels). The phase-optimizing dual-cost-function simulated annealing algorithm is trying to find the best phase distribution and the optimum independent phase delays at the

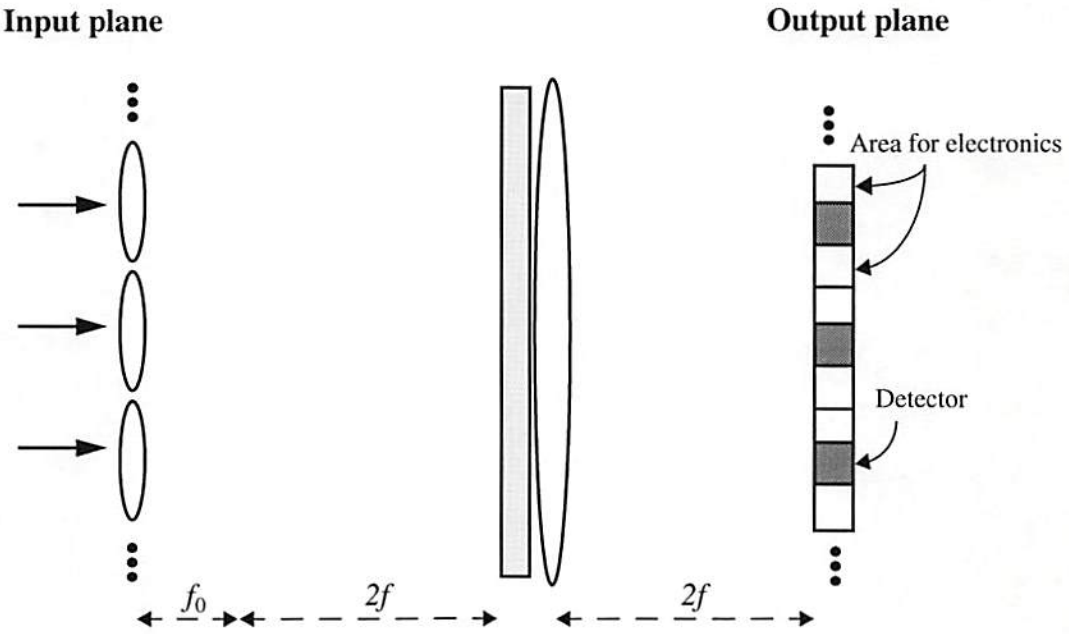

Figure 9-3: One possible realization of a space-invariant interconnection system. The lens array (with focal length f_0) is used for focusing the light coming from neuron units. The reconstruction optics is a bulk lens (with focal length f) for shifting the reconstruction patterns.

same time. It might be possible to use the result of the no-phase-quantization Gerchberg-Saxton algorithm, then apply the phase-delay-optimization concept on the resulting continuous-phase DOE, and obtain a good discrete-phase DOE. This proposed design approach could be much faster than the current phase-optimizing dual-cost-function simulated annealing algorithm.

References

1. M. A. Arbib, Ed., *The Handbook of Brain Theory and Neural Networks* (MIT Press, Cambridge, 1995).

2. Carver Mead, *Analog VLSI and Neural Systems* (Addison-Wesley, New York 1989).

3. Special issue on neural networks, *Applied Optics* 26, December, 1987.

4. H. J. Caulfield, J. Kinser, and S. K. Rogers, "Optical neural networks," *Proceedings of IEEE* 77, No. 10, 1573-1582 (1989).

5. Special issue on neural networks, *Applied Optics* 32, March, 1993.

6. H. J. Caulfield, "Parallel N^4 weighted optical interconnections," *Applied Optics* 26, 4039-4040, 1987.

7. S. Kakizaki and P. Horan, "Limitations of optical lateral intraconnection of smart pixel arrays," in *Optical Computing, 1995 OSA Technical Digest Series* 10, 201-203 (Optical Society of America, Washington, DC, 1995).

8. D. S. Wills, N. M. Jokerst, M. Brooke, and A. Brown, "A two layer image processing architecture incorporating integrated focal plane detectors and through-wafer optical interconnect," in *Optical Computing, 1995 OSA Technical Digest Series* 10, 19-22 (Optical Society of America, Washington, DC, 1995.)

9. G. Yayla, A. V. Krishnamoorthy, G. C. Marsden, and S. C. Esener, "A prototype 3D optically interconnected neural network," *Proceedings of IEEE* 82, No. 11, 1749-1762 (1994).

10. W. B. Veldkamp, "Wireless focal planes: on the road to amacronic sensors," *IEEE Journal of Quantum Electronics* 29, No 2, 801-813 (1993).

11. A. R. Tanguay, Jr., B. K. Jenkins, and A. A. Sawchuk, "Dense 3-D Integrated Electronic/Photonic Computing Structures Enabled by Diffractive Optical Elements", *Advanced Research Projects Agency's 1994 Optoelectronics Review*, Monterey, California, June 13-17, (1994).

12. C. Kyriakakis, Z. Karim, A. R. Tanguay, Jr., R. F. Cartland, A. Madhukar, S. Piazzolla, B. K. Jenkins, C. B. Kuznia, A. A. Sawchuk, and C. von der Malsburg, "Photonic implementations of neural networks," in *Optical Computing, 1995 OSA Technical Digest Series* 10, 128-130 (Optical Society of America, Washington, DC, 1995).

13. C. C. Huang, B. K. Jenkins, and C. B. Kuznia, "Weighted space-variant local interconnections based on micro-optic components: crosstalk analysis and reduction," in *Optical Computing, 1995 OSA Technical Digest Series* 10, 280-282 (Optical Society of America, Washington, DC, 1995).

14. A. V. Krishnamoorthy, G. Yayla, G. C. Marsden, and S. C. Esener, "A scalable Optoelectronic Neural System Using Free-Space Optical Interconnects," *IEEE Transactions On Neural Networks*, Vol. 3, NO. 3, 404-413 (1992).

15. C. Huang, K. Jenkins, and C. Kuznia, "Space-Variant Interconnections Based on Diffractive Optical Elements for Neural Networks: Architecture and Crosstalk Reduction," Submitted to *Applied Optics*.

16. A. Goldstein, "Scalable Photonic Neural Networks for Real-Time Pattern Classification," USC-SIPI report #307, 1997.

17. A. Goldstein, B. K. Jenkins, "Neural-Network Object Recognition Algorithm for Real-Time Implementation on 3-D Photonic Multichip Modules," OSA Annual Meeting/ILS-XII, Rochester, N. Y. (1996).

18. C. Kuznia, C. Huang, K. Ananthanarayanan, C. Chen, and A. Sawchuk "Micro Diffractive Optical Elements for Smart Pixel Fanout Interconnections," 1995 Annual Meeting, Optical Society of America, Portland, Oregon.

19. K. Ananthanarayanan, C. Chen, S. DeMars, C. Huang, D. Su, C. Kuznia, C. Kyriakakis, Z. Karim, B. Jenkins, A. Sawchuk, and A. Tanguay "Multilayer Electronic/Photonic Multichip Modules with Vertical Optical Interconnections," 1995 Annual Meeting, Optical Society of America, Portland, Oregon.

20. H. S. Hinton, T. J. Cloonan, F. B. McCormic, A. L. Lentine, F. A. Tooley, "Free-Space Digital Optical Systems," IEEE Preceding, Vol. 82, N0. 11, 1632-1648 (1994).

21. D. Zaleta, M. Larsson, W. Dachner, and S. H. Lee, "Coupled Kinoforms for Space Variant Optical Interconnection Systems," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 80-83.

22. S. Aoyama and T. Yamashita "Superimposed Grating for use with Magneto-Optical Disk Heads," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 167-170.

23. M. Warren, S. Kravitz, G. Hadley, J. Wendt, G. Vawter, and M. Armendariz "Application of Diffractive Optics to Photonic Integrated Circuit Packaging," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 171-174.

24. P. Mckee, J. Tower, A. Thurlow, and D. Wood, "A Fresnel Lens Array for Optical Fiber Semiconductor Laser Coupling," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 235-238.

25. M. Feldman, "Diffractive optics move into the commercial arena," *Laser Focus World,* 143-151 (Oct. 1994).

26. A. Kathman and E. Johnson, "Binary Optics: New Diffractive Elements for the Designer's Tool Kit," *Photonics Spectra* 125 (Sep. 1992).

27. Feature Issue on "Diffractive Optics: Design, Fabrication, and Applications," *Applied Optics*, Vol. 32, No. 14, May, 1993.

28. *Diffractive Optics: Design, Fabrication, and Applications*, 1992 OSA Technical Digest Series (Optical Society of America, Washington, DC, 1992).

29. *Diffractive Optics: Design, Fabrication, and Applications*, Vol. 11, 1994 OSA Technical Digest Series (Optical Society of America, Washington, DC, 1994).

30. H. Dammann, and K. Gortler, "High Efficiency In-line Multiple Imaging by Means of Phase Holograms," *Optical Communication*, Vol. 3, No. 5, 313-315 (1971).

31. J. Jahns, M. Down, M. Prise, N. Streibl, and S. Walker, "Dammann Grating for Laser Beam Shaping," *Optical Engineering*, Vol. 28, No. 12, 1267-1275 (1989).

32. S. Walker, and J. Jahns, "Array Generation with Multilevel Phase Gratings," *Journal Optical Society of America A*, Vol. 7, No. 8, 1509-1513 (1990).

33. A. Lohmann, W. Lukosz, J. Schwider, N. Streibl, and J. Thomas, "Array Illuminator for Optical Computer," *SPIE Vol. 963 Optical Computing*, 232-239 (1988).

34. F. McCormick, "Generation of Large Spot Arrays from A Single Laser Beam by Multiple Imaging with Binary Phase Gratings," *Optical Engineering*, Vol. 28, No. 4, 299-304 (1989).

35. G. Swanson, and W. Veldkamp, "Diffractive Optical Elements for Use in Infrared Systems," *Optical Engineering*, Vol. 28, No. 6, 605-608 (1989).

36. C. H. Wang, B. K. Jenkins, and J. M. Wang, "Visual cortex operations and their implementation using the incoherent optical neuron model," *Applied Optics* 32, 1876-1887 (1993).

37. D. C. Van Essen and C. H. Anderson, "Information Processing Strategies and Pathways in *the Primate Retina and Visual Cortex*," Chapter 3 in S. F. Zornetzer, J. L. Davis, and C. Lau, Ed., *Introduction to Neural and Electronic Networks* (Academic Press, San Diego, 1990).

38. C. H. Wang, and B. K. Jenkins, "Subtracting Incoherent Optical Neuron Model: Analysis, Experiment, and Applications," *Applied Optics*, 29, 2171-2186 (1990).

39. R. L. De Valois and K. K. De Valois, *Spatial Vision* (Oxford University Press, New York, 1990).

40. M. R. Taghizadeh and J. Turunen, "Synthetic diffractive elements for optical interconnection," *Optical Computing and Processing* 2, 221-242 (1992).

41. J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, San Francisco, 1968).

42. P. Keller and A. Gmitro, "Design and analysis of fixed planar holographic interconnects for optical neural networks," *Applied Optics* 32, 5517-5526 (1992).

43. E. G. Johnson, A. D. Kathman, D. H. Hochmuth, and A. Cook, "Advantages of genetic algorithm optimization methods in diffractive optic design," *Diffractive and Miniaturized Optics* (54-74), SPIE Press, 1993.

44. A. Kathman, and D. Brown, "New Techniques for Genetic Algorithm Optimization of Diffractive Optical Elements," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 137-138.

45. R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination phase from image and diffraction plane pictures," *Optik* 35, 237-246 (1972).

46. F. Wyrowski, "Diffractive Optical Elements: Iterative calculation of quantized, blazed phase structures," *Journal of Optical Society of America A* 7, No. 6, 961-969 (1990).

47. F. Wyrowski, and O. Bryndahl, "Iterative Fourier-Transform Algorithm Applied to Computer Holography," *Journal of Optical Society of America A 5*, 1058-1065 (1988).

48. H. Akahori, "Spectrum Leveling by an Iterative Algorithm with a Dummy Area for Synthesizing the Kinoform," *Applied Optics*, 25, 802-811 (1986).

49. F. Wyrowski, "Iterative Quantization of Digital Amplitude Hologram," *Applied Optics*, 28, 3864-3870 (1989).

50. M. Feldman and C. Guest, "Iterative encoding of high-efficiency holograms for generation of spot arrays," *Optical Letters* 14, 479-481 (1989).

51. A. Vasara, M. R. Taghizadeh, J. Turunen, J Westerholm, E. Noponen, H. Ichikawa, J. M. Miller, T. Jaakkola, and S. Kuisma, "Binary surface-relief gratings for array illumination in digital optics," *Applied Optics* 31, 3320-3336 (1992).

52. M. P. Dames, R. J. Dowling, P. McKee, and D. Wood, "Efficient optical elements to generate intensity weighted spot arrays: design and fabrication," *Applied Optics* 30, 2685-2691 (1991).

53. M. Kim, and C. Guest, "Simulated Annealing Algorithm for Binary Phase Only Filters in Pattern Classification," *Applied Optics*, Vol. 29, N0.8, 1203-1208 (1990).

54. M. Kim, and C. Guest, "Block-quantized Binary-phase Holograms for Optical Interconnection," *Applied Optics*, Vol. 32, N0.5, 678-683 (1993).

55. D. Stack, and M. Feldman, "Recursive Mean-squared-error Algorithm for Iterative Discrete On-axis Encoded Holograms," *Applied Optics*, Vol. 31, N0.23, 4839-4846 (1992).

56. R. Morrison, "Symmetries That Simplify the Design of Spot Array Phase Gratings," *Journal Optical Society of America A*, Vol. 9, No. 3, 464-471 (1992).

57. J. Mait, "Design of Binary-phase and Multiphase Fourier Gratings for Array Generation," *Journal Optical Society of America A*, Vol. 7, No. 8, 1514-1528 (1990).

58. J. Fienup, "Iterative Method Applied to Image Reconstruction and to Computer-generated Holograms," *Optical Engineering*, Vol. 19, 297-305 (1980).

59. J. Fienup, "Phase Retrieval Algorithms: a Comparison," *Applied Optics*, Vol. 21, 2758-2769 (1982).

60. C. Chen, and A. Sawchuk, "Nonlinear Least Square and Phase Shifting Quantization Methods for Diffractive Optical Element Design," to be published in *Applied Optics*.

61. B. K. Jenkins, P. Chavel, R. Forchheimer, A. A. Sawchuk, and T. C. Strand, "Architectural implications of a digital optical processor," *Applied Optics* 23, 3465-3474 (1984).

62. P. Keller and A. Gmitro, "Computer-generated holograms for optical neural networks: on-axis versus off-axis geometry," *Applied Optics* 32, 1304-1310 (1993).

63. A. Gmitro, P. Keller, C. Coleman, and P. Maker, "Design and Fabrication of Multi-Level Phase Holograms for On-Axis Optical Interconnects," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 239-242.

64. M. W. Farn, "Modeling of Diffractive Optics," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 48-51.

65. R. Morrison and M. Wojcik, "A Selective Cell-based Algorithm for Designing High Efficiency Beam Array Generators," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 76-79.

66. U. Krackhardt, "Optimum Quantization Rules for Computer Generated Holograms," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 139-142.

67. Y Han, C. Delisle, and L. Hazra, "Exact Surface Relief Profile of Kinoform Lenses From a Given Phase Function," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 147-150.

68. S. Kirkpatrick, C. Gelatt, Jr., and M. Vecchi, "Optimization by Simulated Annealing," *Science,* 220, 671-680 (1983).

69. N. Yoshikawa and T. Yatagai, "Interpolation Approaches to Computer-Generated Holograms," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 155-158.

70. M. Clark, "A Direct Search Method for the Computer Design of Holograms for the Production of Arbitrary Intensity Distribution," in *Diffractive Optics*, Vol. 11, 1994 OSA Technical Digest Series, 159-162.