Miniaturized Diffraction Grating Design and Processing for Deep Neural Network

Lidan Lu, Lianqing Zhu, Qiankun Zhang, Bofei Zhu, Qifeng Yao, Mingxing Yu, Haisha Niu, Mingli Dong, Guoshun Zhong, and Zhoumo Zeng

***Abstract*— Researchers of UCLA reported the fully connected optical neural network, the high computational rate, and low power consumption was realized, while the diffraction grating system based on the Terahertz source is expensive and bulky. In this letter, a long-wave infrared source with a wavelength of 10.6 um is used to establish an optical neural network transfer model using the Sommerfeld diffraction theory. Diffraction grating design, processing, and error analysis applied to deep neural networks are carried out. The MNIST handwritten database is used as the data set to train and optimize the phase parameters by forward propagation and backpropagation. The neuron size is 5um; the number of neurons is 200∗200; the entire grating area is 1mm. Compared with the existing light diffraction neural network, the feature size of the deep learning neural network is reduced by 80 times. The *Ge* (Germanium)-based diffraction grating of 5 layers of neurons with four relative step heights is engraved by semiconductor standard processing technology. The surface-infrared high-efficiency anti-reflection film increased the grating transmission efficiency to over 90%.**

***Index Terms*— Random diffraction grating, processing technology, deep learning neural network, phase parameters.**

I. Introduction

DEEP learning has become a standard component of almost every image recognition [1], speech recognition [2], and machine translation system [3], [4]. The limitations, such as speed and energy consumption, have been explored dramatically in the literature. Lin *et al*. moved the neural network from the chip to the real world, utilizing the propagation of light to achieve near-zero energy consumption and zero-latency deep learning. This solution is called D2NN (Diffractive Deep Neural Network) [5]. D2NNs were then improved by modifying the loss function to reduce the vanishing gradient problem during backpropagation [6]. Although some researchers deem D2NN as a 'mischaracterization' of the system due to linearity and passivity [7], the original authors have overturned the 'mischaracterization' with detailed introduction [8]. They also demonstrate systematic improvements in diffractive optical neural networks based on a differential measurement technique that mitigates the strict non-negativity constraint of light intensity [9]. The systems above are using electromagnetic waves in the terahertz band, which requires large and expensive equipment. There is a strong demand for a different scheme which makes diffractive neural networks more practical. The solution proposed in this letter is to use short-wavelength sources, gratings, and detectors.

To obtain a smaller diffraction grating system and reduce the cost of the diffraction grating of D2NNs. A diffraction grating fully connected neural network system based on carbon dioxide laser (CO2) is studied. Compared with the method of Terahertz source, the CO2 with wavelength of the 10.6 μm and its infrared HgCdTe amplified photodetectors have been widely used in industry. Thus, we study the design and fabrication of diffraction gratings for D2NN based on a 10.6 μm source. As far as the diffraction grating itself is concerned, the pixel size is reduced from 400 μm to 5 μm (reduced to 80-fold), and 3-fold reduces the distance between each layer.

II. Model of Diffract vie Deep Neural Network

A. Optical and Fully Connected Neural Network Theory

The D2NN structure is adopted, possessing the function of a digital classifier. Each neuron has a size of $d = 5$ μm (d$< \lambda/2$, $\lambda = 10.6$ um), the total size of each layer of neurons is 1mm $\times$ 1mm). The relationship between neuron size (d) and the wavelength (λ) of light source is consistent with the Rayleigh Sommerfeld diffraction theory. A single neuron can be used as a point source of a wave, and the model is shown in Fig.1 (b), the optical mode can be expressed as [5], [10]:

$$w_i^l(x, y, z) = \frac{z - z_i}{r^2}\left(\frac{1}{2\pi r} + \frac{1}{j\lambda}\right)\exp\left(\frac{j2\pi r}{\lambda}\right) \tag{1}$$

where l represents the l-th layer of the network; i represents the i-th neuron of layer l; $r = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2}$ located at (x_i,y_i, z_i) indicates the Euclidean distance between the node i of the layer l and the node of the layer $l + 1$; and $j = \sqrt{-1}$.

The phase of the neurons is trained via backpropagation, as in [5]. The gradient ϕ_i^l is calculated for each layer concerning the loss function and updated accordingly. More details can be found in the appendix.

Manuscript received September 26, 2019; accepted October 14, 2019. Date of publication October 21, 2019; date of current version December 19, 2019. This work was supported by the Program for Changjiang Scholars and Innovative Research Team in University under Grant No. IRT_16R07. *(Corresponding authors: Lianqing Zhu; Zhoumo Zeng.)*

L. Lu, G. Zhong, and Z. Zeng are with the State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin 300072, China (e-mail: zhmzeng@tju.edu.cn).

L. Zhu and M. Dong are with the Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science and Technology University, Beijing 100192, China.

Q. Zhang, Q. Yao, M. Yu, and H. Niu are with the Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science and Technology University, Beijing 100192, China (e-mail: lqzhu_bistu@sina.com).

B. Zhu is with Beijing ZX Intelligent Chip Technology Company, Ltd., Beijing 100876, China.

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LPT.2019.2948626

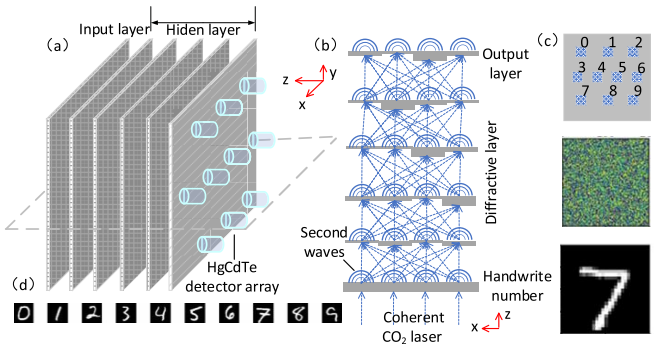
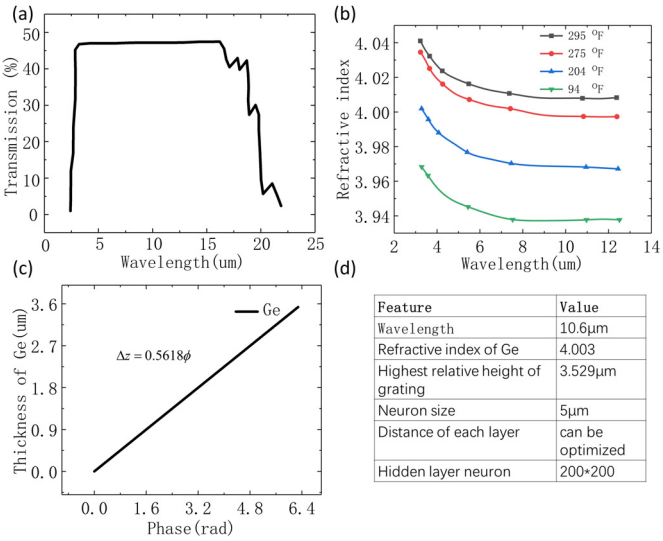


Fig. 1. (a) Structure for digital classifier, (b) Diffraction full connection principle diagram for diffractive deep neuron network, (c) the position labels corresponding to 10 numbers, (d) digital display in handwritten digital library.



Feature	Value
Wavelength	10.6μm
Refractive index of Ge	4.003
Highest relative height of grating	3.529μm
Neuron size	5μm
Distance of each layer	can be optimized
Hidden layer neuron	200×200

Fig. 2. (a) Transmission spectrum of Ge in the case of a thickness of 10mm (b) refractive index n versus wavelength at four different temperatures. (c) The relative height changes when the phase change from 0 to 2π. (d) feature of deep diffraction neuron network.

B. Structure of the System

To train and test the D2NN as a digit classifier, we use the MNIST handwritten digit dataset [5]. The dataset consists of 70000 handwritten digits from teenagers and adults, with the corresponding labels. The images are 28 × 28 pixels, with 55000 images in the training set, 5000 in the validation set, and 10000 in the test set [11], part of the image is shown in Fig.1(d). The verification set is used to further determine the hyperparameters in the model, such as the number of hidden layers, the number of neurons. The test set is used to evaluate the accuracy of the model, that is, the generalization ability.

The size of the hidden layer is 200∗200, to achieve a fully connected neural network. First, the input data set is up-sampled, that is, the pixel size is changed from 28∗28 to 200∗200. Then, the input image is binarized, that is, the pixel value is represented by 0 and 1, or the image is subjected to gray normalization, that is, the pixel value 0 to 255 is normalized to [0, 1].

C. Parameter Design for Diffraction Grating

To obtain a smaller size of the diffractive grating, a diffraction grating fully connected neural network system based on CO_2 source is adopted. There are two main reasons why a wavelength of 10.6um is adopted in this system. Firstly, the CO_2 sources and HgCdTe photovoltaic detectors have been widely used in the market. Secondly, in terms of the choice of diffraction grating processing technology, it needs flexibility, low fabrication and integration complexity, robustness, all while maintaining compatibility with CMOS technology. The current processing conditions in the laboratory, the lateral processing accuracy of the diffraction grating is higher than 0.5 μm, and the longitudinal machining accuracy is higher than 30nm.

Ge is a versatile infrared material commonly used in imaging systems and instruments in the 2 to 12 microns spectral region shown as Fig.2(A) in the case of a thickness of 10mm. It can be used as a substrate for lenses, windows, and output couplers for low-power pulsed CO_2 lasers [12]. As depicted in Fig.2(B), the higher the temperature, the greater the refractive index of *Ge*. At room temperature, the refractive index of *Ge* is 4.003 [13]. Compared with the material having a refractive index of 1.7227 used in Reference [5], the difference in refractive index of *Ge* from air is large, possessing a stronger ability to limit the light field. Therefore, the designed diffraction grating has a small size and has a higher demand on the processing precision. The relationship between neuron's phase values and the relative height of *Ge* can be expressed as:

$$\Delta z = \frac{\lambda \phi}{2\pi\, \Delta n} = 0.5618\phi \tag{2}$$

where λ is the incident light, n is the difference between the refractive index of *Ge* and air, thus $\lambda = 10.6um$, $\Delta n = n_{Ge} - n_{air} = 4.003 - 1 = 3.003$. when the phase changes by 2π, the relative height is up to 3.529um shown in Fig.2 (c).

To achieve full connectivity between neurons of each layer, optical diffraction connects the neurons at different layers of the network. If the feature size of neuron (d_s) is larger than half the wavelength, the maximum half-cone diffraction angle can be formulated as: $\varphi_{\max} = \sin^{-1}(\lambda/(2d_s))$, the size of hidden layer neurons is 200∗200, if d_s is assumed to be 6um, then $\varphi_{\max} = 62.1°$, the distance between each layer should be more extensive than 530 μm, otherwise the distance is too small to achieve full connectivity. According to the Rayleigh Sommerfeld diffraction theory, if d_s is set to be less than half of the wavelength, that is 5 μm, full connection can be achieved regardless of the distance between layers. Then, the digital classifier can be optimized to find a suitable distance, to the extent that rightness rate is the highest.

D. Calculation and Optimization

The digit classifier is trained with MNIST datasets and achieved the desired mapping functions between the input and output planes after 50 epochs. The training batch size is set to be 10, for the digit classifier network. To achieve optimal performance, there are many hyperparameters that must be tuned, including the number of layers, the number of neurons, and the diffraction grating thickness after stepping. Handwritten digit recognition in the scheme is a multi-classification, its performance evaluation adopts error analysis, accuracy (Np/N)

can be expressed as the percentage of the number of positive samples (Np) in the total number of samples (N) [14]. The accuracy rate of all digital classifications in the following text is calculated based on the confusion matrix. The test flow more extensive in the S.1. The input data is 10,000 test handwritten digital images after preprocessing, and the output is the final optimized parameter value (phase Ø).

Neuron Numbers. According to Rayleigh Sommerfeld diffraction theory, a single neuron of each layer of a diffractive neural network can be used as a point source of a wave. The system has $200*200*5 = 0.2$ million neurons, each having a trainable phase term. The spacing of each diffraction grating is set to 10mm. The number of diffractive layers is five, the confusion matrix with the number of neurons $100*100$ and $200*200$ are shown in Fig.3(a) and Fig.3(b), the test accuracy rates are 85.7% and 88.2%, respectively. Input digital handwriting layer and five hidden layers corresponding light intensity distribution are shown in Fig.3(c) ($100*100$ and $200*200$). The last hidden layer possessing $200*200$ neurons corresponds to a more pronounced square focus than that of $100*100$ neurons in the position of handwritten number 7.

The number of the network layers. A single diffractive layer cannot achieve the same level of inference that a multi-layer D2NN structure can perform. Multi-layer architecture of deep diffraction neuron networks provides a large degree-of-freedom within a physical volume to train the transfer function between its input and the output planes, which, in general, cannot be replaced by a single phase-only layer. The confusion matrix of a 2, 3, and 5-layer network are shown in Fig.S.2(a), Fig.S.2 (b) and Fig.S.2 (c), the test accuracy rates are 74.1%, 79.1% and 88.2% respectively. Comparing Figure 3(b) with Figure.S.2 (c), both of them use five layers of hidden layers; each layer is with $200*200$ neurons. As shown in Fig.S.2 (d), the last hidden layer with a total of five layers corresponds to a square focus on the position of handwritten number 4.

Layer-to-layer distance. It knows from the manuscript that the feature size is set to be less than half of the wavelength ($\lambda = 10.6\mu$m), that is 5um, achieving Rayleigh Sommer diffraction. To facilitate the align of the grating mask in the experiment, a distance of more than 5mm between each hidden layer is needed. The test accuracy rates of a 5mm, 10 mm, 20mm between each layer s are 74.1%, 88.2% and 79.1% respectively. Since the research on experimental testing is not covered in the text, according to the experimental requirements, the spacing of each diffraction grating is set to 10 mcm, which facilitates the alignment of the grating neurons.

In order to facilitate the actual processing, the 40000 phases of the trained $0\sim2\pi$ interval are classified into four heights, and the stepped phase is calculated to be $0\sim \pi/2$, $\pi/2\sim \pi$, $\pi \sim 3\pi/2$, $3\pi/2\sim 2\pi$, the phase in the interval is classified as $\pi/2$, π, $3\pi/2$, 2π, respectively. The distance between adjacent layers is 1000μm; the number of diffractive layers is 5; the number of neurons is 200×200 in every single layer. The confusion matrix with the stepped network layers is shown in Fig.S.3 (a), and the accuracy rate is 74.4%, which is 13.4% lower than the accurate rate before uncategorized. As shown in Fig.S.3 (b), the last hidden layer with a total of 5 layers corresponds to a clear square focus on the position of handwritten number 3, which shows a little blurry than the uncategorized d steps. To facilitate the processing and

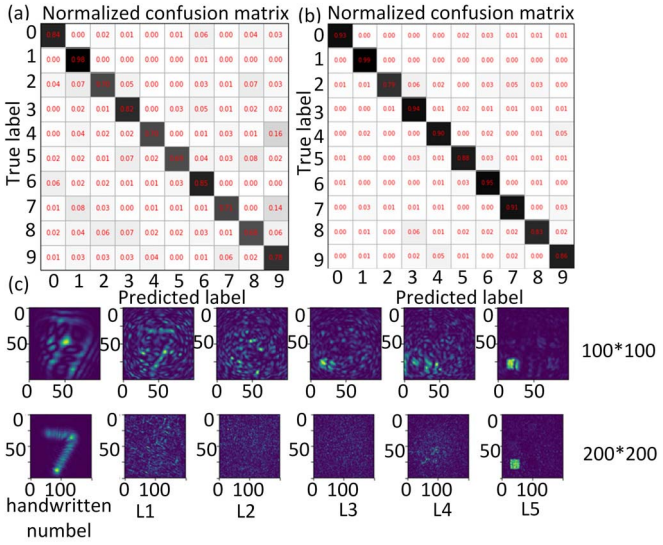


Fig. 3. (a) and (b) show the confusion matrix with the number of neurons $100*100$ and $200*200$, after the training phase, the light intensity distribution of five different layers (L1, L2,…, L5) of the handwritten digit classifier are shown in (c).

verify the feasibility of the processing, only four (2^2) kinds of step heights are classified. We can engrave $8(2^3)$, $16(2^4)$, $32(2^5)$ even 64 (2^6) kinds of step heights through multiple photoetching. This way improves the test accuracy of the classifier.

III. Diffraction Grating Processing and Error Analysis

According to the calculation and optimization results, the grating adopts five hidden layers, each layer of neurons is $200*200$, and the neuron size is 5μm$*5\mu$m. Grating mask design with engraving schemes, it's not wise to adopt metal chromium as a masking layer for alignment marks and scratch marks. To increase the light transmittance, the antireflection coating should be coated on both sides of the grating, and transmittance reaches to 1 ideally, the surface of the antimony-plated anti-reflective film requires no metal on it. If the surface of the crucible is cleaned with a de-chromizing solution, the chromium solution will react with the sulfhydryl group. Thus, the first mask is in a 4-inch wafer, with five hidden layers, simultaneously with alignment marks and scratch marks. The second mask only needs to be aligned with the first mask, as shown in Fig. 4(a)-4(c).

The *Ge* with a height of 170 μm will have a natural warping phenomenon, and although its absorption is small, the alignment error is big. The thickness of the *Ge* is 700 μm, which is rigid, but will increase the loss, so the *Ge* needs to be coated with double-sided anti-reflection film.

The relative step heights of the *Ge* corresponding to the phases $\pi/2$, π, $3\pi/2$, and 2π are 0.88 μm, 1.76 μm, 2.64 μm, and 3.5um, respectively. In order to form four heights, a set of engraving is used, and the processing steps are shown in Fig. 4(d), firstly, a lithographic negative adhesive NR9-3000PY with a relatively large etching is choose, a thickness of 4 um is formed at 4000 r/m; secondly, a photolithography machine is used for pattern transfer; lastly,

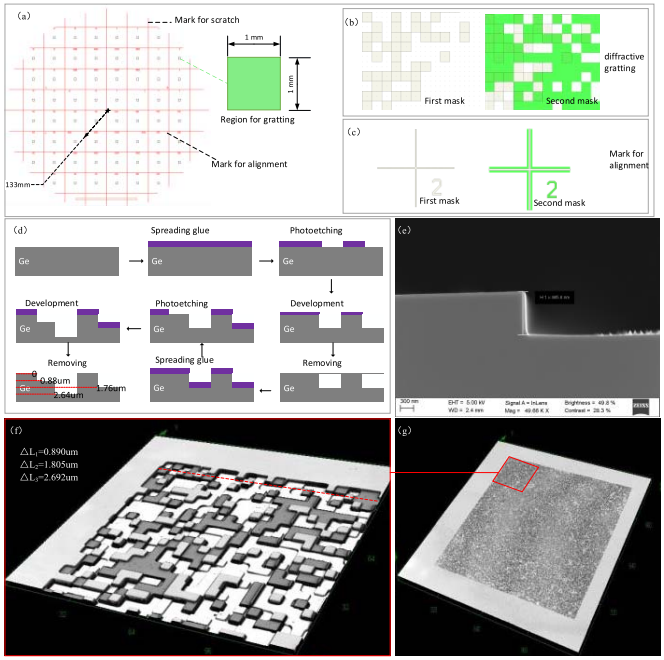


Fig. 4. (a) Mask layout for diffractive grating, (b) the structure of part of the mask layout and (c) alignment marks, (d) processing steps, (e) etching verticality, (f) magnified view about cross section of Ge characterized by SEM after etching, (g) cross section of Ge characterized by SEM after etching.

an inductively coupled plasma (ICP) etching machine is used for etching, the main gas component is hydrogen bromide. The etching rate is about 10 nm/s, and the step sidewalls formed are steep characterized by SEM, as shown in Fig. 4(e). A 3d microscope characterizes the step thickness of the diffraction grating, as shown in Fig. 4(f) and 4(g). The maximum measurement error after multiple measurements is about 50nm, and it is fed back into the model. To fully understand the influence of the step height processing error, the tolerance model with the step height error of 0nm∼50nm was tested, and the accuracy of the test set was only 77.4%. Therefore, processing errors must be reduced. It requires process groping to maximize machining accuracy. The results show that the diffraction grating needs to adopt this semiconductor processing technology. In order to make a_i^l in S. Eq. (1) be a constant, ideally 1, *Ge* needs to be coated with double-sided anti-reflection film, which is processed at *Beijing Qifeng Landa Optical Technology Development Co., Ltd*. In the scheme of terahertz as the light source, the 3D printing of the diffraction grating is difficult to meet the miniaturization requirements of the grating.

IV. Conclusion

The parallel computing capacity and power consumption of optical systems are gradually demonstrated in the advantages of artificial neural networks. Compared with previous electronics-based learning methods, deep learning fully connected neural networks provide a unique all-optical learning method. Diffraction grating possesses tens of thousands of neurons and inartificial fully connected properties. On the one hand, generation and detection of Terahertz source require highly precise and custom-made optical components. Design based on the CO_2 source is easy to obtain from the market. On the other, comparing with the existing D2NN, the feature size of the diffractive grating is reduced by 80 times. It laid the foundation for further large-scale integration of photonic computing chips. Mature photoelectric detection makes the application of optical diffraction method revolutionary. The *Ge*-based wafer is processed in batches compatible with the semiconductor processing technology. A single engraving process can realize the stepping process. Future applications of this method include image analysis and feature detection. May be due to different light source bands, or training optimization parameters are not optimal, the classification accuracy of 88.2% is lower compared to 91.75% shown in reference [5]. The forward plan is to be more compact and can be extended to the visible light band or 1.5um communication band. In this case, high-precision diffraction grating processing schemes such as 3D lithography direct writing technology [15] and femtosecond laser micromachining [16] are required. These initial findings demonstrate a proof of concept that a wavelength of 10.6um diffractive deep neuron network can realize data classifier, more detailed investigations are currently underway.

References

[1] T.-H. Kim, D. Kang, K. Pulli, and J. Choi, "Training with the invisibles: Obfuscating images to share safely for learning visual recognition models," 2019, *arXiv:1901.00098*. [Online]. Available: https://arxiv.org/abs/1901.00098

[2] J. Gao, X. He, and L. Deng, "Deep learning for Web search and natural language processing," in *Proc. 8th ACM Int. WSDM Conf.*, Shanghai, China, 2015.

[3] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: https://arxiv.org/abs/1406.1078

[4] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, pp. 441–446, Jun. 2017.

[5] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–1008, Sep. 2018.

[6] D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Analysis of diffractive optical neural networks and their integration with electronic neural networks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 1, Jan./Feb. 2018, Art. no. 3700114.

[7] H. Wei, G. Huang, X. Wei, Y. Sun, and H. Wang, "Comment on 'All-optical machine learning using diffractive deep neural networks'" 2018, *arXiv:1809.08360*. [Online]. Available: https://arxiv.org/abs/1809.08360

[8] D. Mengu, Y. Luo, Y. Rivenson, X. Lin, M. Veli, and A. Ozcan, "Response to Comment on 'All-optical machine learning using diffractive deep neural networks'" 2018, *arXiv:1810.04384*. [Online]. Available: https://arxiv.org/abs/1810.04384

[9] J. Li, D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Class-specific differential detection in diffractive optical neural networks improves inference accuracy," Jun. 2019, *arXiv:1906.03417*. [Online]. Available: https://arxiv.org/abs/1906.03417

[10] J. W. Goodman, *Introduction to Fourier Optics*. Mexborough, U.K.: Roberts, 2005.

[11] *The MNIST Database of Handwritten Digits*. Accessed: 2012. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[12] *Optical Materials-Germanium (Ge)*. Accessed: 2016. [Online]. Available: http://www.iiviinfrared.com/Optical-Materials/ge.html

[13] H. W. Icenogle, B. C. Platt, and W. L. Wolfe, "Refractive indexes and temperature coefficients of germanium and silicon," *Appl. Opt.*, vol. 15, no. 10, pp. 2348–2351, Oct. 1976.

[14] P. H. Evangelista *et al.*, "Modelling invasion for a habitat generalist and a specialist plant species," *Diversity Distrib.*, vol. 14, no. 5, pp. 808–817, Sep. 2008.

[15] T. Bückmann *et al.*, "Tailored 3D mechanical metamaterials made by dip-in direct-laser-writing optical lithography," *Adv. Mater.*, vol. 24, no. 20, pp. 2710–2714, May 2012.

[16] R. Schaeffer, *Fundamentals of Laser Micromachining*. Boca Raton, FL, USA: CRC Press, 2016.