Fully Integrated Optical Spectrometer in Visible and Near-IR in CMOS

Lingyu Hong, *Student Member, IEEE*, and Kaushik Sengupta, *Senior Member, IEEE*

***Abstract*—Optical spectrometry in the visible and near-infrared range has a wide range of applications in healthcare, sensing, imaging, and diagnostics. This paper presents the first fully integrated optical spectrometer in standard bulk CMOS process without custom fabrication, postprocessing, or any external optical passive structure such as lenses, gratings, collimators, or mirrors. The architecture exploits metal interconnect layers available in CMOS processes with subwavelength feature sizes to guide, manipulate, control, diffract light, integrated photodetector, and read-out circuitry to detect dispersed light, and then back-end signal processing for robust spectral estimation. The chip, realized in bulk 65-nm low power-CMOS process, measures 0.64 mm $\times$ 0.56 mm in active area, and achieves 1.4 nm in peak detection accuracy for continuous wave excitations between 500 and 830 nm. This paper demonstrates the ability to use these metal-optic nanostructures to miniaturize complex optical instrumentation into a new class of optics-free CMOS-based systems-on-chip in the visible and near-IR for various sensing and imaging applications.**

***Index Terms*—CMOS, dispersion, fluorescence, gratings, metal-optics, nano-optics, photodetecton, spectrometry, waveguides.**

I. Introduction

THE next generation of integrated, low-cost, low-power, miniaturized sensing modalities have the potential to bring transformative changes for a wide range of Internet-of-Things (IoT) applications, which is enabled by the ability to capture a wealth of crucial information and data, and process them over the cloud. Among these sensing modalities, optical spectrometry is one of the rapidly growing areas due to its well-known capability to resolve the optical absorption, reflection, and emission (fluorescence) spectra of a variety of materials that carry their chemical information. The ability to monitor chemical composition of materials has significant impact for a range of applications such as in environmental monitoring [1], [2], biomedical sensing [3], healthcare and fitness [4], [5], drug and medicine [6] and food and water quality monitoring. As an example shown in Fig. 1(a), spectrometers have been shown to be able to monitor chlorophyll levels in water which is an indication of the concentrations of nutrients arising from man-made sources like septic system leakage, poorly functioning wastewater treatment plants, or fertilizer runoff [7]. Other examples include detecting toxic Melamine by measuring the Raman spectra. Melamine is an addition in infant milk products [8], which is a flame-retardant plastic misused to boost the nitrogen content (Fig. 1(a)). In addition, many hydrocarbons of critical importance such as ethanol, benzene, aromatic, Toulene, ETBE and MTBE have spectroscopic signatures in the 750–900 nm range. Beyond sensing,

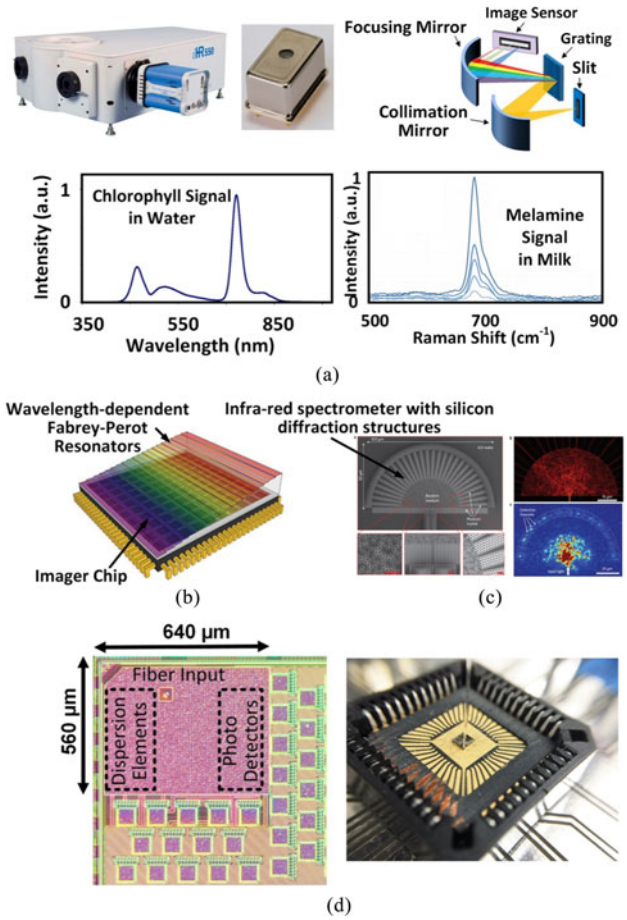


Fig. 1. (a) Classical optical spectrometer with gratings for light dispersion, concave mirrors for focusing and collimation, and other passive optics, and linear image sensors for light detection. Application of optical spectrometry in food and water quality testing [7], [8]. (b) Spectrum sensing with spatially varying wavelength-dependent Fabry-Perot resonators grown precisely on an imager chip [14]. (c) Infrared spectrum sensing with silicon diffraction structures and external detector arrays [15]. (d) Single chip CMOS spectrometer between 500–830 nm and an optical fiber interface (this work).

Manuscript received June 27, 2017; revised September 28, 2017; accepted October 22, 2017. Date of current version December 29, 2017. This work was supported in part by the National Science Foundation and in part by the Qualcomm Innovation Fellowship. This paper was recommended by Associate Editor R. Genov. *(Corresponding author: Kaushik Sengupta.)*

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: lingyu@princeton.edu; kaushiks@princeton.edu).

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

Digital Object Identifier 10.1109/TBCAS.2017.2774603

hyperspatial imaging in the near-IR and visible range also has wide ranging applications in remote sensing, machine vision for autonomous UAVs, medical imaging, and security and surveillance.

However, the classical configuration of optical spectroscopy in the visible and near infrared (NIR) range involves a collection of discrete optical and electronic components assembled together, including gratings for light dispersion, concave mirrors for focusing and collimation, lenses and other passive optics, and linear image sensors for light detection (Fig. 1(a)). This results in a typically non-integrated and expensive system. While there have been prior works on miniaturizing each of these individual components, fundamentally they rely on the same set of components in very similar arrangements. For example, there are efforts to miniaturize the bench-top instrument into a mini-spectrometer by combining the dispersion and imaging component into one single discrete element, a concave grating [9]. While the cost and size of these miniatures have indeed been reduced at the expense of performance such as lower spectral resolution, the cost and size can still be prohibitive for future IoT and sensing applications that require full integration, low power, compactness and low-cost. In recent years, there have been efforts to integrate the optical spectroscopy system into a chip-scale form in the visible and IR range. In the visible range, optical filter arrays have been fabricated on top of image sensors to resolve the spectrum. The filter arrays can be realized either by patterning multiple dielectric layers [10] or by depositing precisely grown Fabry-Perot resonant cavities of variable thickness resulting in different optical transmission bands spatially distributed across the different pixels [11]–[14]. Evidently, these design methodologies require multi-step post fabrication processes and precise alignment which can add considerable cost. On the other hand, in the infrared region, particularly in the telecommunication band around 1.5 μm, narrow-band chip-scale spectrometers based on multiple optical scattering inside a silicon waveguide have been demonstrated [15]. However, its implementation requires off-chip detector arrays and evidently, the silicon-waveguide-based design cannot be employed in the visible range as silicon is highly absorptive.

In this paper, we discuss the co-integration and co-design of the optical nano-structures along with detection and read-out electronics in a single chip to enable realization of a complete spectrometer on-chip in a standard bulk CMOS process in the visible and NIR range. This requires no post-processing, custom fabrication or any other external optical elements enabling a truly low-cost, low-power, and extremely compact (mm-sized) spectrometer for the next-generation sensing and imaging applications, as shown in Fig. 1. The paper presents the end-to-end design aspect that covers the entire stack from interfacing with an external optical fiber to guiding of light inside the chip, controlled dispersion with on-chip scattering nanostructures and finally, detection and processing on-chip. The method to manipulate and guide optical fields inside CMOS in the visible range is achieved by exploiting sub-wavelength lithographic features in the interconnect layers to realize complex metal-optic structures. This allows us to eliminate all external optical elements and miniaturizes the spectrometer in a single CMOS chip operating across 500–830 nm. Integration of such nano-optical systems in CMOS can open the door to a wide range of new applications ranging from compact sensing interfaces to spectroscopy, imaging and health diagnostics [17]–[27]. The preliminary result of this work was presented in [28]. This paper expands on the theory, the co-design methodology of electronics and nano-optics, design details, analysis trade-offs, and measurements.

The paper is organized as follows. Section II presents the overview of the CMOS spectrometer architecture focusing on the co-design of both optics and electronics in CMOS. Section III discusses in details the optical design on-chip including quasi-single-mode wave guiding, controlling the flow of light and design of the integrated dispersion structures . Section IV discusses the detection and read-out circuitry, noise analysis and measurement and electrical characterization of the system. Section V presents optical characterization, spectroscopy measurement results and the fundamental limits on achievable resolution, limit of detection and efficiency.

II. CMOS Optical Spectrometer: System Overview

The ability to manipulate optical fields in the visible range in CMOS have been demonstrated in prior works in creating diffraction structures with angle-sensitive imagers [29]. In this work, the lowest metal layers were used to create diffractive structures that allow incident light to be split and create angle-dependent light-field imaging. The dimensions of these metal-optic structures are comparable to the wavelength of light. However, when these metallic structures reach sub-wavelength feature sizes, their interaction with optical fields enters a different regime with remarkable ability to confine, concentrate, guide and filter light with nanoscale noble metals, opening up many exciting applications including deep sub-diffraction imaging, extraordinary optical transmission through sub-wavelength hole arrays and perfect lensing [30], [31]. In [32], [33], vertically coupled surface-plasmon waveguides were demonstrated to enable angle and scattering insensitive filtering to enable a fully integrated fluorescence biosensor on chip with integrated nano-filter arrays. The system presented here also exploits the copper interconnect layers in modern day CMOS processes to guide and diffract light in a controlled fashion to be detected and processed for spectrum information.

The overall system architecture including the light guiding path with an optical fiber interface is shown in Fig. 2. The perspective and top view of the CMOS chip and coupling of light into the chip, light scattering, detection, and processing therein are shown in Fig. 2(a) and (b) respectively. Firstly, the incident light is coupled through an optical fiber into a small aperture on top of the chip, which is realized on the 7th metal layer (M_7) of the 65 nm CMOS process. Through the aperture, the light is coupled into a horizontal waveguide realized between the 4th (M_4) and M_7 layer. These together form a metal-insulator-metal (MIM) waveguide inside the chip supporting a low-loss mode that propagates horizontally. The waveguide mode is then scattered by the carefully designed metallic dispersive structures, realized also between M_4 and M_7 with metal as well as via layers.

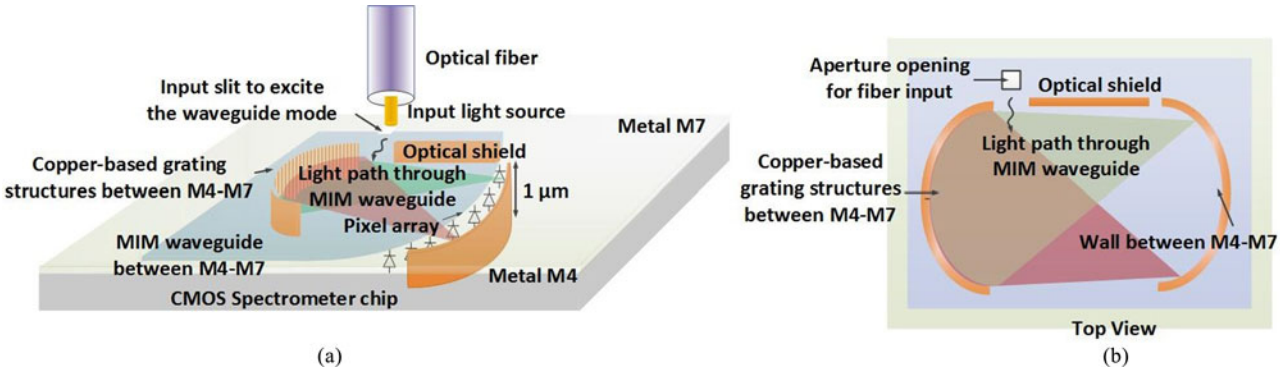


Fig. 2. CMOS optical spectrometer with integrated light guiding path, diffraction structures, detection and read-out circuitry. (a) Perspective view. (b) Top view.

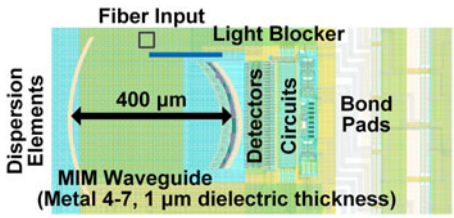


Fig. 3. Co-design and layout of the integrated optical structures and detection circuitry.

The structures are designed to resolve the incident light into different wavelengths and focus them to different horizontal positions at the other (right) end of the chip. As can be seen in Fig. 2, an optical shield is employed between M_4 and M_7 to prevent light from propagating inside the waveguide without being scattered by the dispersion elements. The light is then reflected from an optical wall at the right end of the waveguide into an array of photo-detectors underneath. The signal at the photodetectors is then processed on-chip and finally digitized off-chip for further analysis. The optical spectrum is estimated from the measured signal of the detector array through back-end signal processing as will be discussed in detail in the following sections.

Fig. 3 shows the layout of the chip, including the optical dispersive elements, MIM waveguide, photodetector arrays, circuits and bond pads which are co-designed together for optimal performance. Fig. 1(d) shows the entire chip micrograph with the top aperture and the active region of the chip measures 0.65 mm $\times$ 0.54 mm in size.

III. Integrated Nano-Optical Structures

A. Light to Waveguide Mode Conversion

As illustrated in Fig. 2, the free space light at the output of the optical fiber is firstly coupled into the MIM waveguide modes propagating inside the chip. The coupling from the fiber into the propagating mode can be enhanced by employing a grating structure designed on the metal layers. Fig. 4(a) and (b) show respectively the structure and the electric field profile of free-space-light to waveguide-mode conversion simulated in a Finite Difference Time Domain (FDTD) simulator, the main simulation tool for Section III. Note that in this work, all simulations

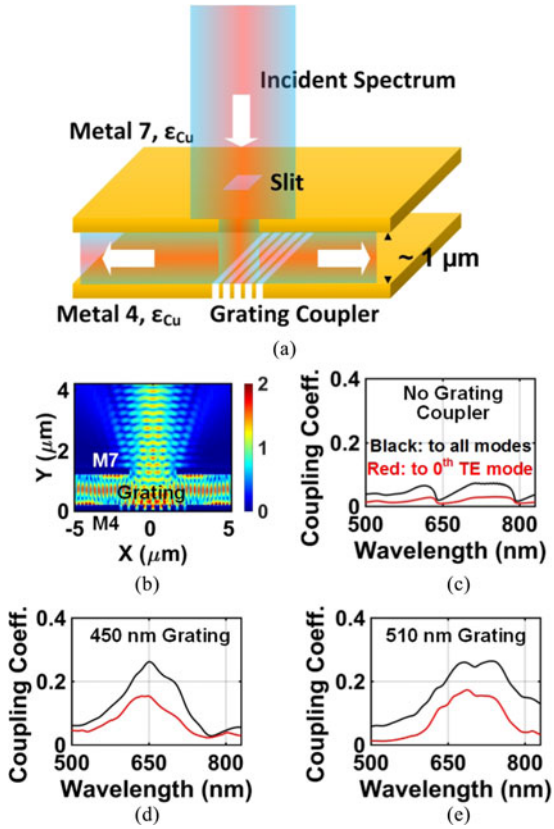


Fig. 4. Improving coupling efficiency from the fiber into the chip with a grating coupler structure. (a) Grating structure designed in M_4 with alternate metal and dielectric layers. (b) Simulated electrical field showing launching of horizontal propagating mode. (c) Coupling efficiency with no grating coupler. (d) and (e) Coupling enhancement with grating coupler designed to cover different optical bands. Efficiency is calculated for unpolarized light with black line showing the total coupling efficiency to all modes and red line showing the coupling efficiency to 0th order TE mode.

as well as experiments are performed under the unpolarized light condition, which is a realistic condition for the majority of applications of a chip-scale spectrometer.

As can be seen in the figure, when the light from the fiber is incident onto the aperture, a horizontally propagating mode is coupled in the waveguide. Expectedly, direct impinging of light onto the aperture can result in poor coupling efficiency (Fig. 4(c)) which can be significantly enhanced with a grating

structure on the bottom metal plane of the waveguide (M_4) with alternative metal and dielectric structure (Fig. 4(d) and (e)). The efficiency is calculated for unpolarized light with black line showing the total coupling efficiency to all modes and red line showing the coupling efficiency to 0th order TE mode, which is the dominant mode in the waveguide, as discussed in the following section. Additionally, as shown in the figure, the conversion efficiency against wavelengths can be tuned and optimized by the grating pitch, which can be employed for spectrometer chips operating at different bands [34]. It can be noted that 0th order mode has an effective index of mode propagation that is very close to the index of refraction of the dielectric material (silicon oxide) as a result of the large waveguide thickness (1 μm). This results in a simple approximate relation between the grating pitch and the wavelength of maximum conversion efficiency which can be used as a first-pass design guideline:

$$\frac{2\pi}{\Lambda} = n_{eff}\frac{2\pi}{\lambda_M} \cong 1.45\frac{2\pi}{\lambda_M} \tag{1}$$

where Λ is the grating pitch, n_{eff} is the effective index of the 0th order mode, and λ_M is the wavelength for peak conversion efficiency. The implemented chip does not carry the grating structure, which affects the overall efficiency, but can still achieve spectral resolution with tens of nWatt level of incident power. Of course, this can be significantly enhanced with the grating structure at the interface.

B. CMOS Integrated Copper-Based MIM Waveguide

Once the light is coupled in the waveguide, it is important to make sure that the waveguide maintains a low-loss single mode propagation, as shown in Fig. 5. Conversion to higher-order modes causes unpredictability in scattering behavior, which reduces robustness and makes the accuracy in spectral estimation sensitive to process variations and external factors. If multiple modes exist with vastly different propagation properties and different effective indices that affect their scattering angles from the dispersion elements, slight variations in the fiber-to-chip coupling can result in drastically different outputs of the detectors, making the system sensitive to the coupling condition.

In this design, the choice of the metal layers for the waveguide was decided to ensure a dominant single mode propagation. The 1 μm spacing between the metal waveguide plates can theoretically result in a multi-mode propagation between M_4 and M_7. However, the loss characteristics of copper at optical frequencies result in vastly different propagation losses for different modes. Propagation loss through modal analysis of the MIM copper waveguides with sandwiched silicon dioxide dielectric layer shows that the first-order TE mode (where the electric field vector is perpendicular to the propagation direction and parallel to the metal slab) is suppressed drastically compared to the 0th order TE mode, as the wave reaches the end of the 500 μm long waveguide towards the photo-detection array. The electric field profiles of the propagating wave of the dominant (0th) and the first-order TE modes are shown in Fig. 5(b) and (c) respectively. The variation of the propagation losses with the thickness of the MIM-waveguide in Fig. 5(d) and (e) show that the losses of both the 0th and 1st order modes decrease with thickness and at the same time, the loss difference of the two modes decreases. This is the design trade-off between modal purity and loss. For the reported on-chip implementation of the MIM waveguide, M_4 and M_7 are chosen as the metal layers for a dielectric thickness of 1 μm. This results in at least 28 dB loss difference for the dominant and higher order modes, making it a predominantly single-mode waveguide across the wavelengths. This ensures that the interaction of the wave front against the scattering elements is robust. Higher order TE modes as well as TM modes have much higher losses and therefore, suppressed. The loss for the dominant waveguide mode remains around 7 dB for wavelength above 640 nm and increases towards lower wavelengths. The higher losses with the metallic waveguides are expected at optical frequencies, but does not preclude spectral decomposition of the incident light with power as low as 40 nW entering the chip.

C. The Spectral Dispersion Element

As shown in Fig. 2, the wave coupled inside the chip travels horizontally and impinges on the specially designed dispersive elements inside the chip that provide the wavelength-dependent scattering which is the working principle of a spectrometer. In a classical spectrometer, this can be achieved with a concave grating structure that allows free-space light of different wavelengths to focus into different spots in space, and then imaged by a linear detector array. In this work, the dispersion elements are designed to disperse waveguide modes rather than free-space light, serving as both dispersion and focusing functionality with minimal spectral aberration.

The dispersion element is designed according to the classical Rowland configuration of concave grating [35], as shown in Fig. 6(a). The light wave propagating through the MIM waveguide impinges on the dispersion element and is spectrally separated into different wavelengths which focus spatially at different spots lying on a circle of radius R. This circle is known as the Rowland circle, where the source and image points lie. On the other hand, the grating structure is designed along a small arc of the larger circle with radius $2R$, as shown in Fig. 6. Suppose that the grating pitch is d_1 and d_2 at point A and point B, respectively. In order for both points to image the source P to image point Q constructively at wavelength λ_i, d_1 and d_2 need to satisfy:

$$\begin{aligned} d_1\left(\sin\alpha_1 - \sin\beta_1\right) &= \lambda_i \\ d_2\left(\sin\alpha_2 - \sin\beta_2\right) &= \lambda_i \end{aligned} \tag{2}$$

Since points P and Q are on the Rowland circle, and in the case of small grating size, points A and B are also approximately on the Rowland circle, we have $\alpha_1 \approx \alpha_2$, and $\beta_1 \approx \beta_2$. Therefore, from (3), d_1 and d_2 are almost identical, which implies that the grating has a constant pitch, unlike a Fresnel lens. As different wavelengths are focused at different positions with different angle (β) on the Rowland circle, the constant pitch d ensures minimum aberration for different wavelengths. This allows the dispersion grating to resolve the optical spectrum over a wide range of incident wavelengths. Fig. 6(a) shows the

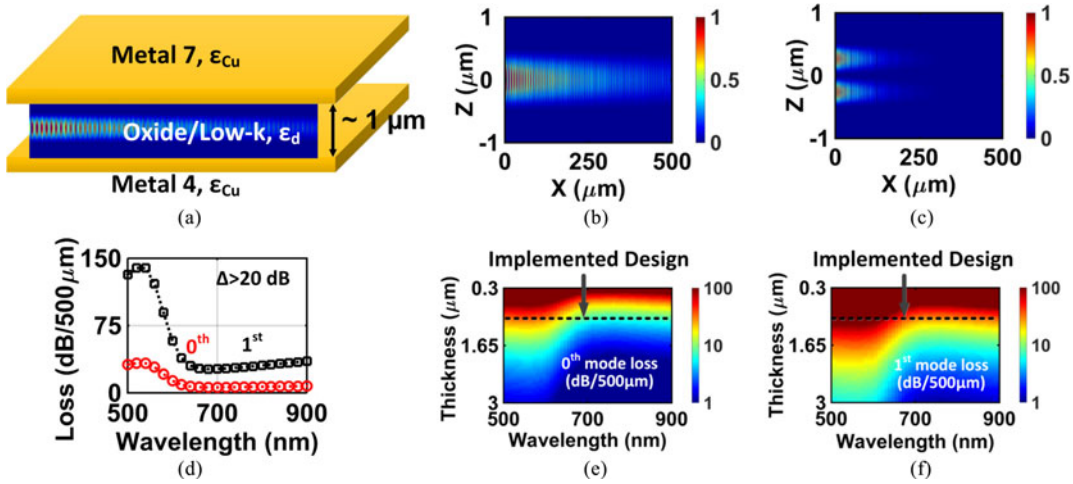


Fig. 5. On-chip quasi-single-mode horizontal waveguide (a) Structure of the metal-insulator-metal waveguide implemented between M_4 and M_7 with 1 μm dielectric thickness. (b) and (c) Propagation of the modes in the implemented waveguide showing strong suppression of the higher order modes making it primarily a single-mode waveguide. (d) and (e) Variation of loss of the 0th and 1st order (TE) modes with dielectric thickness showing lower loss and lower mode purity for thicker waveguides. (f) Variation of propagation loss of the two modes with wavelengths at implemented waveguide thickness (1 μm).

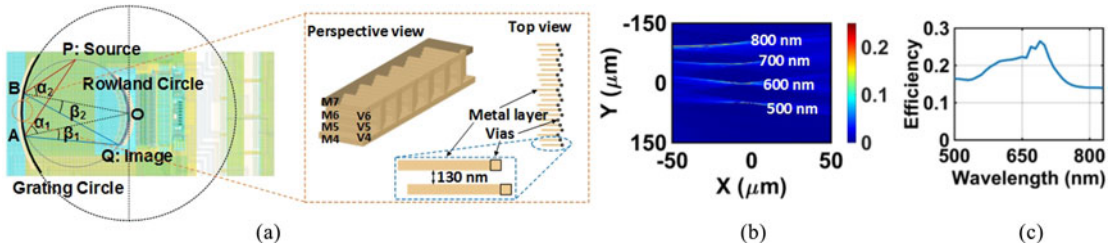


Fig. 6. Integrated concave grating structure. (a) Design of the Rowland gratings inside the chip exploiting the metal and via layers between M_4 and M_7. (b) Simulated focal curvature and the wavelength dependent focal spots within the chip. (c) Simulated absolute grating efficiency considering both grating diffraction and reflection loss.

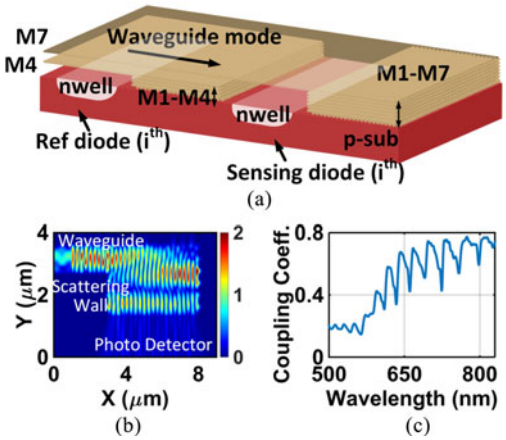


Fig. 7. (a) Detecting the diffracted light traveling through the waveguide by deflecting it onto an array of photodetectors with a vertical wall. (b) FDTD simulation showing the bouncing of the light waves from the wall and traveling downwards to be detected. (c) Simulated coupling efficiency from the waveguide into the array of photodetectors.

zoomed perspective and top view of a small portion of grating structure at the center of grating. As can been seen, the grating is designed using interconnect layers between M_4 and M_7 including both metal layers and via layers. The grating structure is designed to minimize aberration in compliance with the 65 nm CMOS Design Rule Checks (DRC) ensuring high yield. As shown in Fig. 6(a), a single unit cell of the grating comprises of metal strips of 100 nm width (M_5 and M_6) and via layers with 100 nm $\times$ 100 nm dimensions realized with V_4, V_5, and V_6. While the vertical spacings between unit cells is constant at the minimum allowable distance of 130 nm, the other (horizontal) position of the unit cells are chosen to minimize aberration over the wavelengths between 500–900 nm. The details of this design procedure are presented in Appendix, but overall, the radius R is chosen to be 200 μm and the grating dimension is chosen to be around 400 m (vertical). As demonstrated in the simulation in Fig. 6(b), light of various wavelengths are focused at different spots and the focusing points are not significantly broadened at the extreme end wavelengths near 500 nm and 800 nm, demonstrating the effectiveness of the aberration minimization methodology. Fig. 6(c) shows the simulated absolute grating efficiency considering both grating diffraction and reflection loss.

D. Detecting the Dispersed Light

The last piece of the optical path is the reflecting wall that allows the spectrally separated optical fields to be detected and processed by the on-chip photodetectors and low-noise

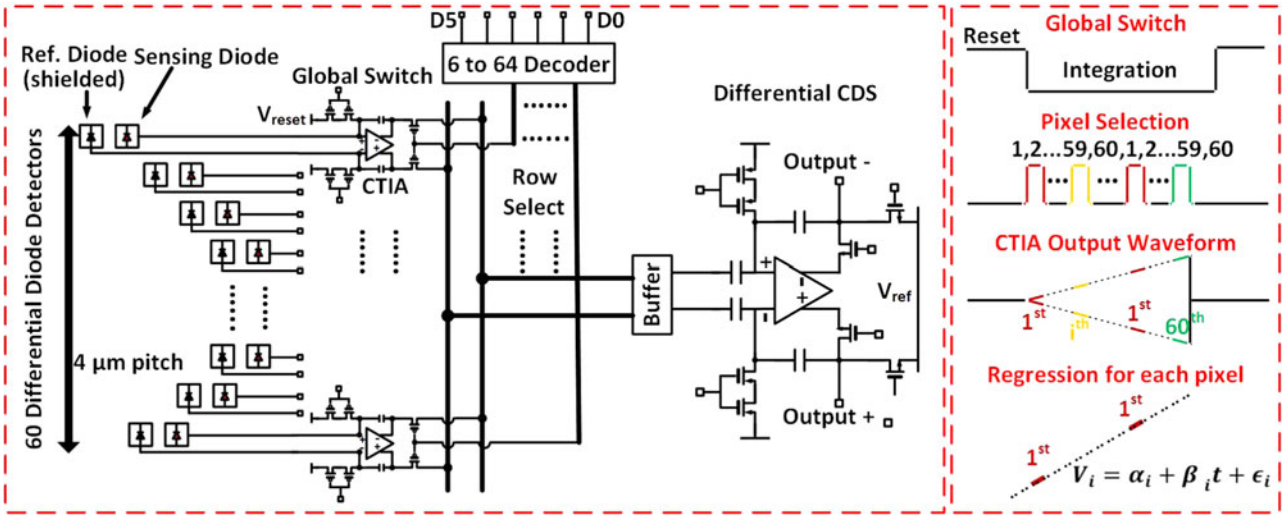


Fig. 8. (a) Photo-detection circuits laid along the focal plane of the concave grating and the corresponding read-out circuitry with CTIA and CDS. (b) Time-multiplexed read-out scheme of the detectors and the CDS outputs. Regression methods using the raw CTIA data can be used off-chip for noise suppression.

electronic circuitry. As shown in Fig. 7, as the waves of different wavelengths propagate and focus at different spots at the end of the waveguide, they are reflected by corrugated optical walls into the array of photodetectors underneath the waveguide. The optical walls are implemented by metal and via layers between M_1 to M_7, as shown in Fig. 7(a), and are naturally corrugated to increase the conversion efficiency. This reduces the number of reflections between the two walls by increasing the scattering into the photodetectors. Fig. 7(b) shows the simulated electric field profile of the propagating-wave to free space light conversion for detection. The efficiency is close to 80% for longer wavelengths and decreases slightly for shorter wavelengths as the loss of copper increases at these frequencies approaching the bulk plasmonic resonance frequency.

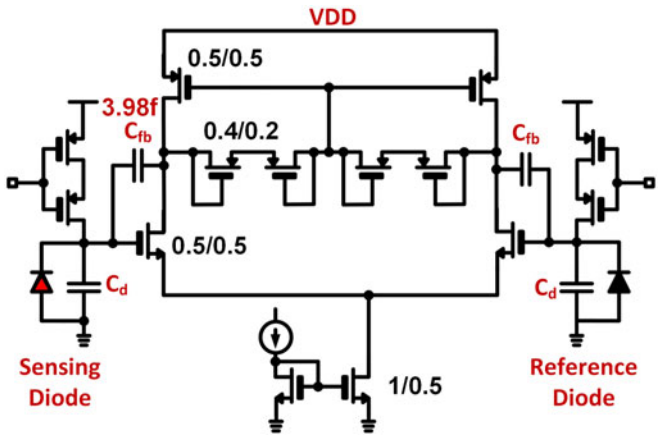


Fig. 9. Schematic of the implemented CTIA.

IV. Sensor Circuitry, Architecture and Electrical Characterization

A. Detection and Read-Out Circuits

Fig. 8 shows the array circuit architecture where 60 differential photo-diode arrays are laid out following the optical focusing curvature. The array of 60 detectors is realized with n-well/ p-sub junctions that has been found to have the largest responsivity in the implemented CMOS process [33]. The signal is integrated by a capacitive transimpedence amplifier (CTIA). Realized in a digital 65-nm bulk CMOS (non-custom imager process), this allows us to eliminate the dependence of the diode capacitance on the integration period. The photo current is integrated in a CTIA, as $V_{op,TIA} = \int \frac{i_{ph}(t))}{(C_{fb}(1+1/A)+C_d/A)} dt \approx \int \frac{i_{ph}(t)}{C_{fb}} dt$ where C_{fb} = 3.98 fF is the feedback capacitance, A is the OTA gain ($\approx$ 35 dB) and C_d is the diode capacitance whose effect is significantly reduced. It is to be noted that the grating structure, the array and the diode layouts need to be co-designed for optimal performance and resolution. As an example, the dimension of the photodetectors are kept at 4 μm, which corresponds to the size of the optical focusing spot. As shown in the figure, each signal is sensed and processed differentially with a reference diode and differential CTIA to minimize common-mode perturbations. A 6-bit decoder is employed to select each pixel and the signals are further processed by the differential correlated double sampling (CDS) circuits to suppress offset and 1/f noise. The signal is finally digitized off-chip.

Fig. 9 shows the schematic of the implemented CTIA. The OTA is designed to have gain around 35 dB, bandwidth around 20 MHz, and pseudo-resistors are employed with transistors in cut-off to minimized layout overhead. Fig. 10 shows the detail of the CDS circuits and the timing diagram for its operation. The pixels can be selected and processed in sequence when the CDS function is turned on, as shown in the timing and simulation result in Fig. 10(b). Alternatively, as shown in the timing diagram in Fig. 8, a global switch can be used to reset and integrate all pixels simultaneously, followed by the pixel read-out in sequence. In order to reduce the total amount of time for acquisition and to reduce the noise via post-processing, the pixel

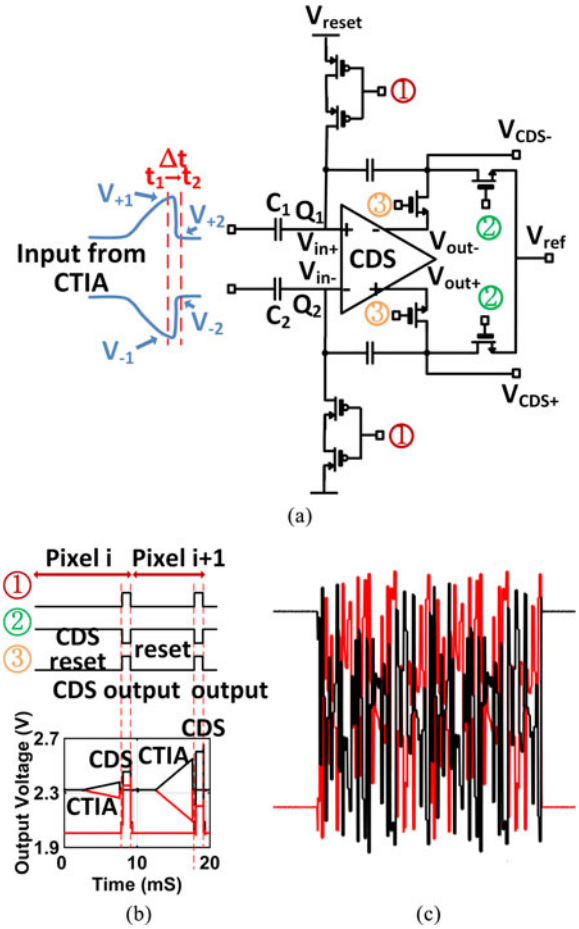


Fig. 10. (a) Implemented correlated double sampling (b) Simulated waveforms for CDS operation when the pixels are selected and processed in sequence. (c) Measured CTIA output waveforms in the alternate reading scheme when all 60 pixels are integrated and readout twice during one single integration period.

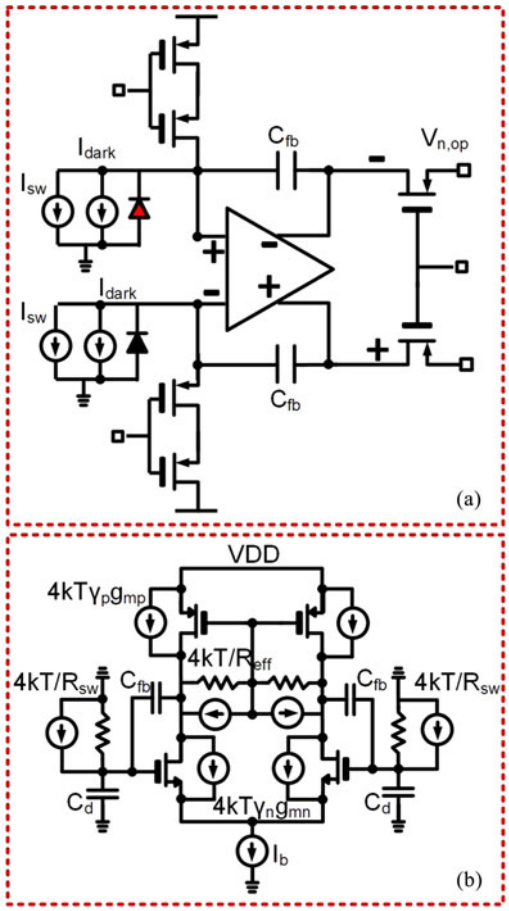


Fig. 11. Noise Sources in the detection and read-out circuitry.

1 to pixel 60 are selected in sequence twice during the common integration period, and the entire analog waveform of the CTIA is directly readout, digitized and analyzed in the time domain. Fig. 10(c) also shows the measured analog CTIA output for all pixels within one data acquisition period lasting around 10 ms. Once the analog waveforms are acquired, regression analysis can be applied to reduce uncorrelated noise as illustrated more in the following section. The chip is operating at 3.3 V supply voltage, each pixel including its individual CTIA consumes around 8.9 μA current and the CDS block consumes around 0.77 mA current.

B. Noise Analysis

Fig. 11(a) shows the details of the circuit schematic for the differential diode and CTIA and Fig. 11(b) shows the contributing noise sources. Such stochastic processes in the system contributed by circuit noise, photon shot noise, and external readout noise can limit achievable light-detection sensitivity and spectral resolution. The total circuit noise primarily consists of the thermal noise and shot noise of the related circuit components and 1/f noise which can be largely suppressed through the CDS processing.

Photon Shot Noise: Photon shot noise is typically the dominant shot noise for an imager. However, in this case, since the incident optical signal is spatially distributed across an array of photodetctors in a non-uniform fashion, the photon shot noise is also non-uniform. If $P_{in,\lambda}$ represents the incident power on the dominant photodiode, then the output noise due to the photon shot noise can be represented as

$$\overline{V_{n,ph,op}^2} = \frac{qP_{in,\lambda}R_fT}{C_{fb}^2} \tag{3}$$

where R_f, T an q are the responsivity (A/W), integration time and electron charge respectively. If the detector is integrated for the maximum integration time allowable by the voltage swing V_{sw} for the CTIA output, then $T \leq T_{\max} = \frac{V_{sw}C_{fb}}{P_{in,\lambda}R_f}$. This makes the maximum photon shot noise for the dominant detector to be constant of the form

$$\overline{V_{n,ph,op}^2}(\max) = \frac{qV_{sw}}{C_{fb}} \tag{4}$$

The resulting photon shot noise can be calculated to be $\sqrt{\overline{V_{n,ph,op}^2}(\max)} = 6.3$ mV for $V_{sw} = 1$ V. While the shot noise power increases linearly with integration time T and V_{sw}, the signal power increases quadratically with these parameters, showing the trade-off between power dissipation and SNR.

Reset Switch Noise: During the reset period, the effective noise current for the switch transistor is $\overline{i^2_{n,sw}}(res) = 4kT/R_{sw}$, where R_{sw} is the on-resistance of the reset switch operating in the linear region in the reset mode. This interfaces with the input reflected capacitance contributed dominantly by the feedback capacitance $C_{in} = C_d + (1 + A(j\omega))C_{fb}$. Since the time-constant contributed by R_{sw} and C_{in} is comparable to that of the dominant time-constant of the OTA ($A(j\omega)$), we need to consider the frequency response of the OTA which is given by

$$A(j\omega) = \frac{g_{mn} - j\omega C_{fb}}{\frac{1}{R_{eff}} + \frac{1}{r_{on}} + \frac{1}{r_{op}} + j\omega C_{fb}} \tag{5}$$

where g_{mn} is the transconductance of the NMOS transistors of the OTA and R_{eff} is the effective resistance of the pseudo-resistors, as shown in Fig. 11(b).

Therefore, the differential output noise voltage of the CTIA during the reset period due to the reset transistor noise is:

$$\overline{V^2_{n,sw,op}}(res) = \int_0^\infty 2 \times \frac{4k_bT}{R_{sw}} \times \left|\frac{A(j\omega)}{\frac{1}{R_{sw}} + j\omega(C_d + C_{fb} + A(j\omega)C_{fb})}\right|^2 df \tag{6}$$

where the factor 2 represents the differential noise power. The above integral is evaluated to be $\sqrt{\overline{V^2_{n,sw,op}}}(res) = 7.2\ mV$.

During the integration period when the reset transistors are switched off, their leakage currents and the diode dark currents are the main noise contributors. Since the voltage noise signal is sampled at intervals of the integration time (T) after resetting every time, the effective noise power at the output of the CTIA is then the mean-square of the time series $V_{n,sw,op}(t)(int) = V_{n,CTIA}(t) - V_{n,CTIA}(t - T)$. This results in the noise power spectral density of the reset transistors during the integration period to be a function on the integration time (T) as

$$\frac{\overline{V^2_{n,sw,op}}}{\Delta f}(int) = \frac{|1 - e^{-j\omega T}|^2}{\omega^2 C^2_{fb}} 4q(\overline{I_{sw}} + \overline{I_{dark}}) \tag{7}$$

where the $\overline{I_{sw}}$ and $\overline{I_{dark}}$ are the average switch leakage current and dark current. This gives the total noise voltage as:

$$\overline{V^2_{n,sw,op}}(int) = \int_0^\infty \frac{\overline{V^2_{n,sw,op}}}{\Delta f}\, df = \frac{2Tq(\overline{I_{sw}} + \overline{I_{dark}})}{C^2_{fb}} \tag{8}$$

This shot noise is often negligible. However, due to the small value of C_{fb} used in this design, it is non-negligible and contributes to the total noise. For our estimated leakage and dark current value ($\sim$10 fA) and integration time of 10 ms, the switch noise during the integration stage is evaluated to be $\sqrt{\overline{V^2_{n,sw,op}}}$(int) = 1.4 mV.

CTIA noise: The thermal noise of the CTIA can be derived and analyzed as follows:

$$V_{n,CTIA,op} = Z_{tot}(j\omega)(i_{nl} + i_{nr} + 2i_{Rl} + 2i_{Rr} + i_{pl} + i_{pr}) \tag{9}$$

TABLE I
SUMMARIZATION OF THE NOISE CONTRIBUTIONS OF THE CHIP

	Notation	Reset Period	integration Period
Reset Switch Noise (Analytical)	$V_{n,sw,op}$	7.2 mV	1.4 mV
CTIA Noise (Analytical)	$V_{n,CTIA,op}$	7.4 mV	2.6 mV
Total Noise in Dark (Analytical)	$V_{n,op}$	10.3 mV	2.9 mV
Total Noise in Dark (Measured)	$V_{n,op,meas}(dark)$	9.36 mV	3.21 mV
Total Noise in Dark (Measured, Regressed)	$V_{n,op,meas}(reg,dark)$	1.2 mV	
Total Noise in light (Measured, Regressed)	$V_{n,op,meas}(reg,light)$	2.1 mV	

where $i_{nl}, i_{nr}, i_{Rl}, i_{Rr}, i_{pl}, i_{pr}$ are the noise currents for the left and right NMOS transistors, pseudo-resistors, and PMOS transistors, as shown in Fig. 11(b) and

$$\frac{1}{Z_{tot}(j\omega t)} = \frac{1}{R_{eff}} + \frac{1}{r_{on}} + \frac{1}{r_{op}} + \frac{j\omega C_{fb}(g_m + \frac{1}{R_{sw}}) - \omega^2 C_d C_{fb}}{\frac{1}{R_{sw}} + j\omega(C_{fb} + C_d)} \tag{10}$$

where r_{on} and r_{op} are the output resistance of the NMOS and PMOS of the OTA, respectively. Therefore, the total noise voltage at the output can be evaluated as follows:

$$\overline{V^2_{n,CTIA,op}} = \int_0^\infty 2|Z_{tot}(j\omega)|^2 \times \left(\frac{8kT}{R_{eff}} + 4kT\gamma_n g_{mn} + 4kT\gamma_p g_{mp}\right) df \tag{11}$$

The expressions in (9)–(11) hold true for both reset and integration periods except that the value of the reset switch resistance R_{sw} for the two periods are different. During the reset period, R_{sw} represents the on-resistance in the linear region, while during the integration period the reset transistor is off and R_{sw} is much larger than the other contributing resistances. Therefore, the noise contribution by the CTIA can be evaluated as $\sqrt{\overline{V^2_{n,CTIA,op}}}(res) = 7.4$ mV and $\sqrt{\overline{V^2_{n,CTIA,op}}}(int) = 2.6$ mV for the reset and integration periods, respectively.

Total noise (dark): Therefore, in dark operation, the total white combining the noise of the photodiode dark current, switch transistor and the CTIA in both reset and integration stage, can be evaluated to be $\sqrt{\overline{V^2_{n,op}}} = 10.7$ mV where $\sqrt{\overline{V^2_{n,op}}}(res) = 10.3$ mV and $\sqrt{\overline{V^2_{n,op}}}(int) = 2.9$ mV. In summary, the noise during the reset period (contributed by both the reset switch as well as the CTIA circuits) is dominant and the analytically derived calculations are summarized in Table I.

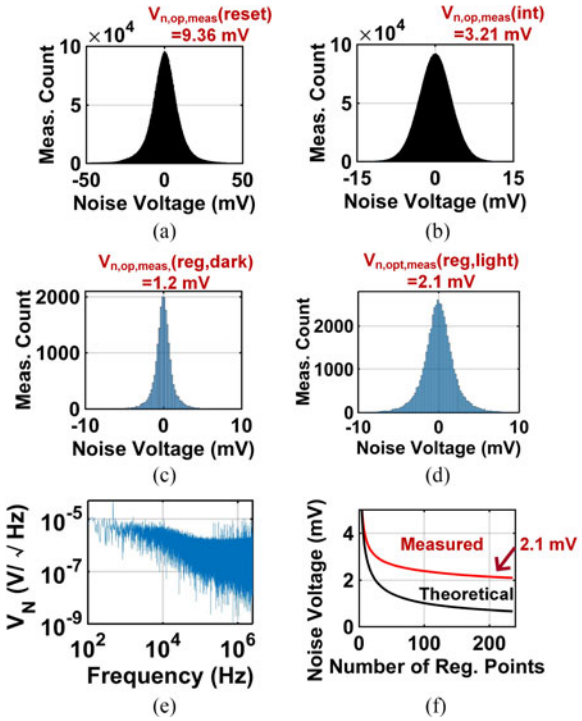


Fig. 12. (a) Measured circuit noise and distribution during the reset period in dark. (b) Measured circuit noise and distribution during the integration period in dark. (c) Reduced total measured noise in dark by applying regression to the CTIA outputs. (d) Reduced total measured noise in light by applying regression to the CTIA outputs. (e) Measured output noise spectrum in light. (f) Measured reduction of noise against number of sample points used in regression showing final tapering off due to correlated noise at the output.

C. Noise Measurement and Suppression Through Regression

Noise of the constituent circuits are measured under both light and dark conditions. This is measured in time-domain by digitizing the signal with a 16-bit off-chip ADC and analyzing the variance of the measured signal. As shown in Fig. 12(a), the reset noise is measured to be $\sqrt{\overline{V^2_{n,op,meas}}}(res) = 9.36$ mV which is close to the theoretically predicted value of $\overline{V^2_{n,op}}(res) = 10.3$ mV. The slightly higher value of the predicted noise is likely due to the over-estimation of the intrinsic gains and output resistances of the stages which increases the noise proportionately as expressed in (11). The noise during the integration period is measured to be $\sqrt{\overline{V^2_{n,op,meas}}}(int) = 3.21$ mV, as shown in Fig. 12(b) which is slightly higher than the analytically derived value of $\sqrt{\overline{V^2_{n,op}}}(int) = 2.9$ mV. In both analytical and measured results, we see that the reset noise dominates over the noise during the integration period (Table I). The total measured noise is, therefore, $\sqrt{\overline{V^2_{n,op,meas}}}(dark) = 9.89$ mV.

In order to reduce the noise levels to increase spectral resolution, regression analysis is performed on the digitized outputs of the CTIA during the integration period. This is achieved by modeling the output of the CTIA as $V_i = \alpha + \beta t_i + \epsilon_i$, where t_i are the sampling times and ϵ_i is the sampled white noise process of variance σ^2_ϵ. When N is the total number of points used for regression, the standard deviation of the regression coefficient β_{reg} can be expressed as $\sigma_\beta \propto \sigma_\epsilon / N^{3/2}$. Therefore, the output noise voltage can be suppressed by regressing over large N as $\sqrt{\overline{V^2_{n,op}}} = (t_n - t_1)\sigma_\beta \propto \sigma_\epsilon / \sqrt{N}$ [33].

By this method, not only the entire array can be read out in around 10 ms of total integration time, the total noise for each pixel under dark can be also reduced from $\sqrt{\overline{V^2_{n,op,meas}}}(dark) = 9.89$ mV to $\sqrt{\overline{V^2_{n,op,meas}}}(reg, dark) = 1.2$ mV, as indicated in the variance reduction in Fig. 12(c). This method can also be applied during normal spectroscopy operation under light when the photon shot noise contributes significantly to the total noise in the system. As shown in Fig. 12(d), the total noise measured under operation with applied regression is given by $\sqrt{\overline{V^2_{n,op,meas}}}(reg, light) = 2.1$ mV. The measured noise spectrum is shown in Fig. 12(e). Fig. 12(f) shows the measured reduction of noise with the number of samples acquired for regression (N). While it initially reduces with increasing number of samples over a period of time, ultimately the measured reduction is limited by the correlated processes in the output noise.

D. Spectral Reconstruction, Robustness and Signal Processing

Mathematically, optical spectroscopy converts an incident optical spectrum vector into a sensor response vector through a responsivity matrix. Consider an incident spectrum discretized into a M dimensional vector $\mathbf{S_{inc}} \in \mathbf{R^{M\times 1}}$ be incident on the spectrometer with N sensors with outputs $\mathbf{V_{op}} \in \mathbf{R^{N\times 1}}$. Then the sensor response can be characterized as a linear transformation as

$$\begin{bmatrix} V_{op,1} \\ V_{op,2} \\ \vdots \\ V_{op,N} \end{bmatrix} = \begin{bmatrix} R_{opt1,1} & \dots & R_{opt1,M} \\ R_{opt2,1} & \dots & R_{opt2,M} \\ \vdots & \ddots & \vdots \\ R_{optN,1} & \dots & R_{optN,M} \end{bmatrix} \begin{bmatrix} S_{inc,1} \\ S_{inc,2} \\ \vdots \\ S_{inc,M} \end{bmatrix} + \begin{bmatrix} V_{n,1} \\ V_{n,2} \\ \vdots \\ V_{n,N} \end{bmatrix} \tag{12}$$

where $\mathbf{R_{opt}} \in \mathbf{R^{N\times M}}$ represents the responsivity matrix which captures the dispersion properties of the grating and $\mathbf{V_n} \in \mathbf{R^{N\times 1}}$ is the output sensor noise. To solve uniquely for the spectrum, we need $N \geq M$. In classical spectrometer, for example, the responsivity matrix is diagonal, which allows different wavelengths to be focused to different locations. However, since this is a linear transfer function, any responsivity matrix which is full rank and well-conditioned can be employed for optical spectroscopy. This is critical since for an on-chip implementation, both stray and scattered light can contribute to the responsivity matrix. In that case, the responsivity matrix needs to be characterized once. Therefore, when an incidence spectrum $\mathbf{S_{inc}}$ creates a sensor response $\mathbf{V_{op}} = \mathbf{R_{opt}}\mathbf{S_{inc}} + \mathbf{V_n}$, the unknown spectrum can be estimated from the measured response $\mathbf{V_{op}}$ and known $\mathbf{R_{opt}}$ through minimum least-square estimation as $\mathbf{S_{inc}} = (\mathbf{R^T_{opt}}\mathbf{R_{opt}})^{-1}\mathbf{R^T_{opt}}\mathbf{V_{op}}$. However, this may not be robust enough in presence of noise [36], [37]. Particularly, the estimated spectrum can be composed of large positive

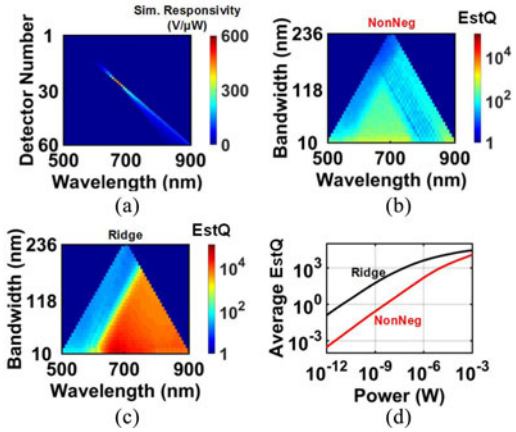


Fig. 13. (a) Simulated responsivity matrix for the grating structure implemented in the design showing near perfect focusing of different wavelengths at different spots. (b) Estimation Quality ($EstQ$) for non-negative least square regularization (c) Estimation Quality ($EstQ$) for ridge regression with Tikhonov regularization. (d) Average $EstQ$ against light power for the two regularization methods.

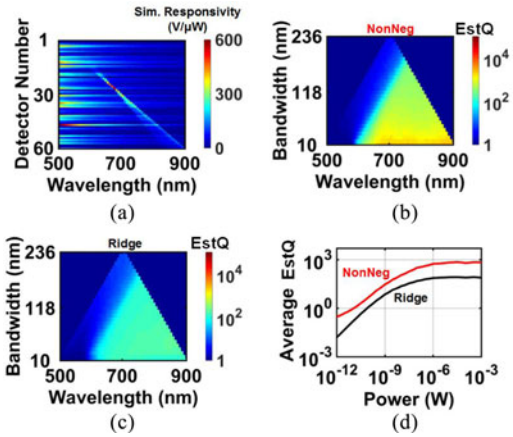


Fig. 14. (a) Simulated responsivity matrix for the grating structure with simulated wavelength-dependent scattering. (b) Estimation Quality ($EstQ$) for non-negative least square regularization. (c) Estimation Quality ($EstQ$) for ridge regression. (d) Average $EstQ$ against light power for the two regularization methods.

and negative spikes which can still mathematically lead to a finite sensor response.

Robust estimation of signals in a linear system in presence of noise is a classical signal processing problem and there are many algorithms that can be applied. In this case, we focus on two. Firstly, we know that all elements in $\mathbf{S_{inc}}$ are positive since this represents the magnitude of the spectrum. This reduces the solution space and eliminates negative spectral estimates in the non-negative least-square estimation as shown below

$$\min_{\mathbf{S_{est}}>\mathbf{0}} ||\mathbf{V_{op}} - \mathbf{R_{opt}}\mathbf{S_{est}}||^2 \tag{13}$$

Secondly, we also investigate the classical LASSO technique with ridge regression and Tikhonov regularization which also simultaneously minimizes the energy of the spectrum, again eliminating the possibility of large positive and negative predictions.

$$\min_{\mathbf{S_{est}}} ||\mathbf{V_{op}} - \mathbf{R_{opt}}\mathbf{S_{est}}||^2 + \lambda||\mathbf{S_{est}}||^2 \tag{14}$$

where λ is regularization parameter that can be optimized for minimum variance depending on the spectrum to be measured [36]. This is a classical technique in regression and machine learning which trades-off variance with bias. The detailed discussions of each in the context of trade-offs is beyond the scope of the paper. However, we will try to highlight these two methods intuitively with the spectroscopic problem a hand.

In absence of scattering, $\mathbf{R_{opt}}$ is close to a diagonal matrix as seen from the focusing simulation in Fig. 6(b). This is represented in Fig. 13(a) which represents the sensor response when the chip is excited at each of the characterization wavelengths. Spectral estimation with this matrix is simulated for varying incident spectra in presence of measured sensor noise voltage of $\sigma(V_n) = 2.1$ mV. The spectral estimation results obtained by employing the two methods *i.e.* the non-negative least-square estimation (13) and ridge regression (14) are shown in Fig. 13(b) and (c) respectively. Here, $EstQ$ which represents the quality of the estimation is defined as

$$EstQ \equiv \frac{1}{Err} = \sqrt{\frac{\Sigma_\lambda |S_{inc}(\lambda)|^2}{\Sigma_\lambda |S_{inc}(\lambda) - S_{est}(\lambda)|^2}} \tag{15}$$

Firstly, the 'triangular' areas in the top left and top right of both Fig. 13(b) and (c) represent areas of non-estimability since for these bandwidths, the spectrum spills beyond the characterization wavelengths between 500–900 nm. In the central triangular region (estimable spectral area), it can be seen that the estimation quality ($EstQ$) increases with wavelength up to around 700 nm (where the focusing spot is minimum), and then slowly decreases with wavelength as the focusing spot spreads out. Ridge regression is shown to perform better than the non-negative least square method for this responsivity matrix. The average $EstQ$ for all wavelengths and bandwidths represented in the two figures is plotted against incident light power in Fig. 13(d). The minimum detection level achieved by ridge regression can theoretically reach down to 0.13 nW for the measured noise voltages for a desired $EstQ = 10$.

However, in practice, as the fiber launches light into the tiny slit of the chip without elaborate optical shielding and packaging, multiple scattering channels exist and the responsivity matrix is expected to be non-diagonal. We modeled the scattering component by random variables across different detectors having wavelength dependency $\propto \lambda^{-4}$. An example of one such new responsivity matrix is shown in Fig. 14(a). In addition to noise, we also consider the case where the responsivity matrix itself can be perturbed which represents the variation between the calibrated matrix and the one during measurement. In this case, we consider $\sigma(R_{opti,j})/\mu(R_{opti,j}) = 0.1\%$ (12) for all i, j. Different from the previous diagonal-like responsivity matrix, the non-negative least square method (Fig. 14(b)) performs better than the ridge regression (Fig. 14(c)) when scattering is included. Fig. 14(d) shows the average $EstQ$ against incident power for all wavelengths and bandwidths represented in Fig. 14(b) and (c). As can be seen, non-negative least square

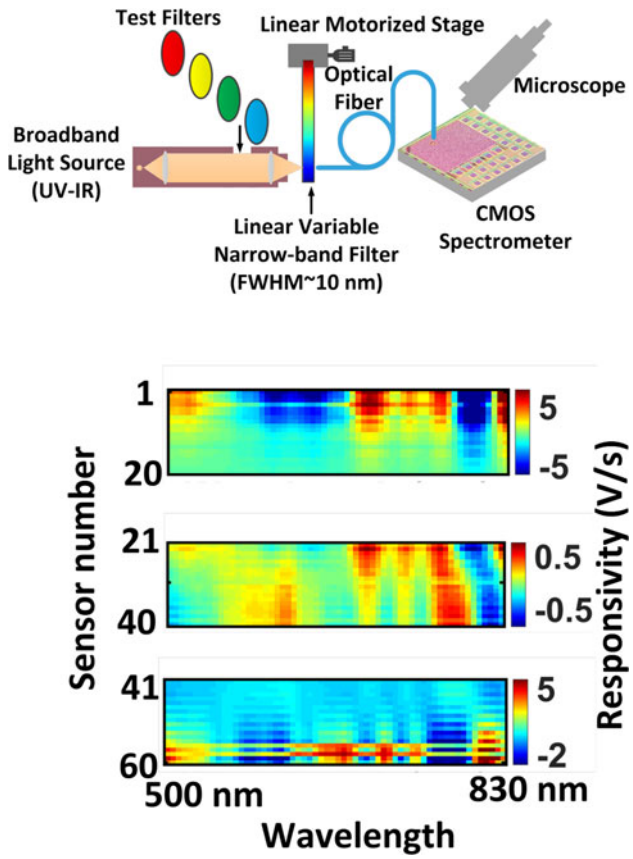


Fig. 15. (a) Spectrometer responsivity characterization with tunable C.W. excitations generated by a broadband source and a linear variable narrow-band filter on a translational motorized stage covering the range of 500–830 nm. The measured spectral responsivity shown here by the pixel outputs against wavelength by removing the average response demonstrates the wavelength-dependent variation of the spatial distribution. The scattering is evident in absence of perfect focusing, but the dispersive grating is robust for spectral estimation from C.W. to wide-band excitations.

method performs better than ridge regression and minimum detection level is 0.24 nW for a desired $EstQ = 10$.

V. Optical Measurements

A. Responsivity Matrix Characterization

The chip is fabricated in 65 nm bulk CMOS process and is interfaced with an optical fiber set up as shown in Fig. 1. Firstly, the optical responsivity matrix $\mathbf{R_{opt}}$ is characterized by exciting the chip with various wavelengths and measuring the sensor response. Fig. 15(a) shows the optical measurement setup for this characterization. Tunable CW excitations are generated from a wide-band source with multiple tunable narrow-band filters with full-width-half-maximum $FWHM \approx 10$ nm. These are realized with a linear variable filter that can be translated across the source input on a motorized translational stage to create very narrowband excitations approximating C.W. sources (Fig. 15). The responsivity matrix is a one-time characterization, and the spectral dependence of the spatial distribution of intensity is shown in Fig. 15(b). In the presence of stray and scattered light, the responsivity matrix is non-diagonal as expected. However, as shown in the analysis before, it is full rank and well-conditioned to allow for spectral reconstruction over the visible and NIR range for varying optical spectra.

B. Narrow-Band and Wide-Band Spectra Measurement

Once the responsivity matrix is captured, the chip is tested for both narrow-band and wide-band excitations. Multiple narrow-band spectral measurements are performed by moving the linear variable filter to various positions corresponding to and between the characterization wavelengths. Wide-band spectral measurements are performed by replacing the linear variable filter by other broadband filter sets in the optical path. The spectral estimation method is shown in Fig. 16(a). The estimated spectrum is represented by a vector and then converted into multiple Gaussian spectra to take into account the 10 nm linewidth of the light source for C.W. excitations. This results in spectral spreading in the final reconstructed spectrum as shown in Fig. 16(a).

When the chip is excited with the same wavelengths of characterization, a near perfect prediction of the incident narrow-band spectra between 500–830 nm is evident from Fig. 16(b). This shows that the implemented nano-optical structures and the single mode waveguides within the chip create a robust and repeatable dispersive profile which can enable accurate spectral estimations in presence of noise. Now, when the chip is excited with light between the characterization wavelengths, the prediction is also very accurate as shown in Fig. 16(b). It can be noted that Fig. 16(b) and Fig. 16(c) are obtained using a responsivity matrix with a step of around 5 nm. The peak prediction accuracy can be improved down to 1.4 nm by averaging over five responsivity matrices shifted by 1 nm, as shown in Fig. 16(d) and (e). It should be noted that the peak prediction error is separate from resolution of a spectrometer which is discussed in details in Section V-D.

The chip is tested with various broadband excitations by employing multiple bandpass filters at the source. This includes a bandpass filter around 600 nm with FWHM of around 40 nm (Fig. 17(a)), bandpass around 800 nm with FWHM of around 40 nm (Fig. 17(b)), and high-pass at around 750 nm. In all cases, the estimation results were fairly accurate with normalized estimation error (Err) defined in (15) kept below 20%. The difference in the incident and estimation is attributed to the effect of external scattering which can be minimized with better optical shielding and packaging. It is known that variable stray light scattering can cause unintended errors in spectral estimation, which are typically eliminated by elaborate optical shielding in classical spectrometers. For a chip-scale spectrometer, such external effects can be minimized externally by black epoxy packaging and internally with on-chip optical walls that prevent propagation of light in undesired directions (Fig. 2). However, in spite of the die being completely unpackaged and unshielded from external variable scattering, the chip shows fairly robust estimation results for both C.W. and wideband optical signals in the visible range.

C. Fluorescence Emission Spectra Measurements

To show the application of the spectrometer in a biosensing set-up, we choose fluorescence spectrometry as an example. In

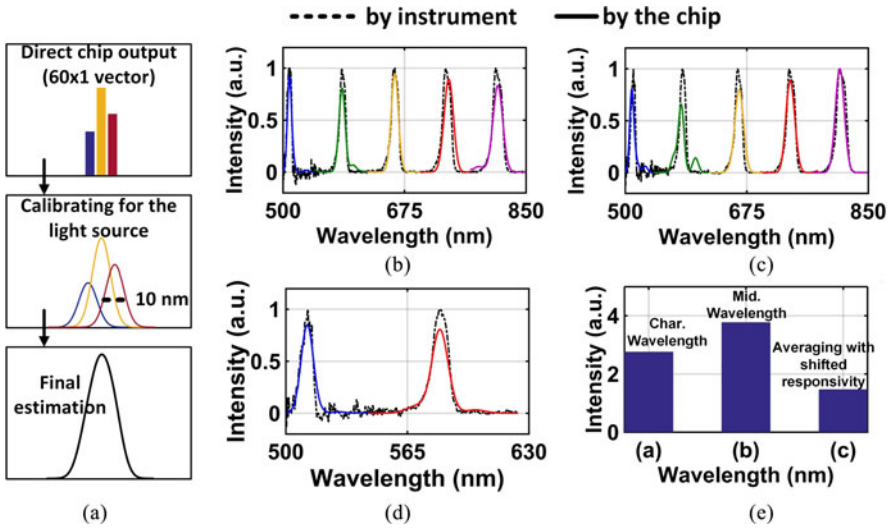


Fig. 16. (a) Spectral estimation method to convert estimated discrete spectrum vector into the final reconstructed spectrum to take into the 10 nm linewidth of the input light source for C.W. excitations (b) and (c) Measured spectral estimation under C.W. excitation at and between characterization wavelengths with estimated 40 nW of optical power entering the chip. (d) With an initial estimation of the spectrum using a responsivity matrix with a step of 5 nm, progressive narrowing of peak estimation by averaging with five responsivity matrices shifted by 1 nm. (e) Averaging increases accuracy of peak wavelength for C.W. excitations down to 1.4 nm.

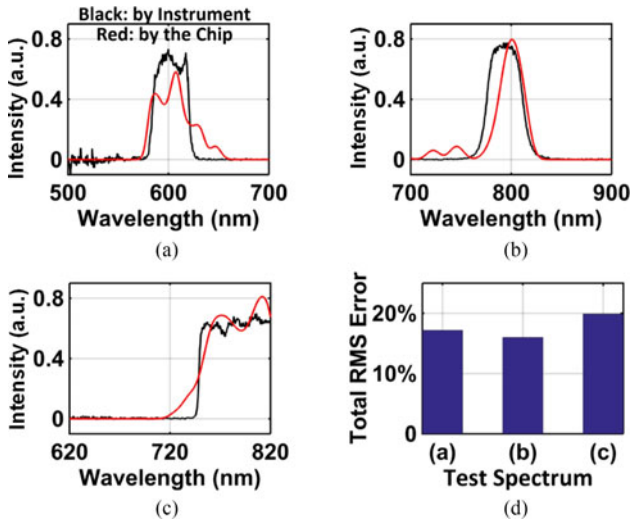


Fig. 17. (a)–(c) Measured spectral estimation with broadband excitations. The light sources of varying spectrum are created with a broadband optical source and multiple bandpass and highpass filters. The results show reasonably good agreement with the incident spectra which can be improved with optical shielding and packaging from external scattering. (d) Normalized estimation errors as defined in (15) in (a)–(c).

particular, we choose quantum dots as the fluorescent tags which have distinct properties such as photo-stability, brightness, and large Stokes Shift that are superior to traditional organic fluorescent dyes [38]. Their sizes are also fairly small to cause any appreciable steric effects in assay experiments and therefore, they are starting to be widely accepted as biological labels for a variety of applications such as nucleic acid and protein detections.

Two types of quantum dots of different wavelengths (centered at around 705 nm and 800 nm, respectively [39]) are measured by the reported CMOS spectrometer chip and compared with their reference spectra measured by bench-top spectrometers. Firstly, the Qdots are measured as fluorescence tags with sandwich immuno and DNA assays (Fig. 18(a)). In particular, DNA hybridization assay and a sandwich immunoassays with target DNAs (42 bases) and target IL-6 cytokines respectively are chosen. The latter plays a critical role in regulation of biological processes in various cell types and in auto-immune processes of many diseases. Biotinylated capture DNA and Human IL-6 antibody probes are used with Biotin-PEG linkers for nucleic acid assays and by employing silane chemistry for immunoassay respectively. As a proof-of-concept, both the assays are tested with target concentrations varying between 1–100 pM and 5–125 pM respectively using streptavidin-conjugated Qdot as the fluorescence label. In both cases, the intensity is shown to vary linearly with concentration which shows the effectiveness of the Qdot to be used as a tag for an affinity-based assay. An example of the captured fluorescence image for 10 pM target DNA concentration after incubation for 10 minutes and washing is shown in Fig. 18(a).

In order to spectrally characterize the fluorescence emission from the two types of tags, a fiber fan-out cable is utilized to collect light from the quantum dots. As shown in the Fig. 18(b)–(c), one branch of the fiber is placed vertically, and quantum dots can be dropped using a pipette on top of it with small volumes (0.2 μL for each drop). Both types of the quantum dots are excited by a diode laser operating at 405 nm wavelengths. The measured fluorescence spectra by the CMOS spectrometer chip for 705 nm and 800 nm quantum dost are compared against their reference spectra measured in a benchtop spectrometer. The results are shown in Fig. 18(d), showing reasonably good agreement, including precisely predicating the shift in the fluorescence emission peak. This resolution can be considerably enhanced with improved optical shielding and packaging which can reduce external variable scattering.

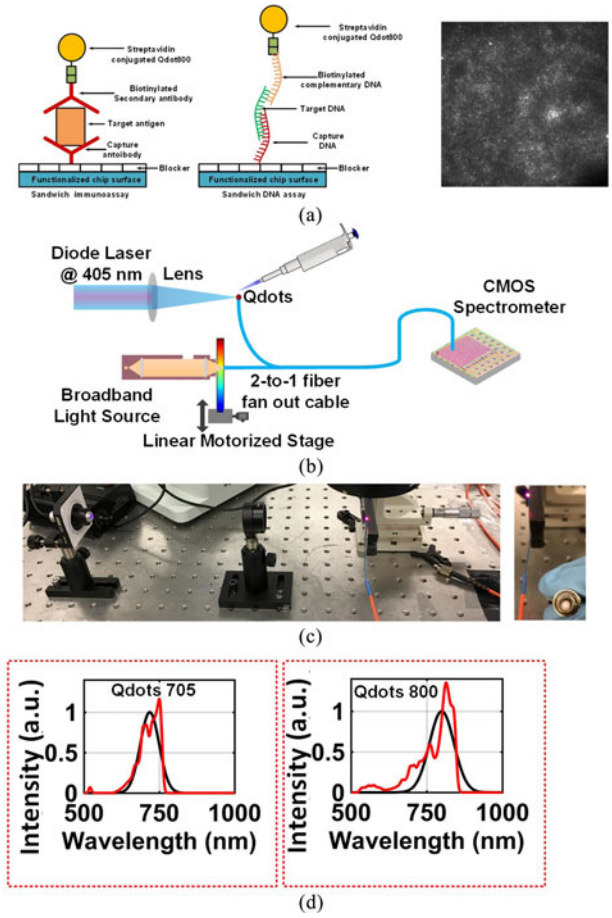


Fig. 18. (a) Using quantum dots for nucleic acid and immunoassay shows linear variation of intensity with target conjurations. The snapshot shows the fluorescence image of a DNA hybridization assay with 10 pM target concentration after incubation for 10 minutes and washing. (b) and (c) Setup for quantum dot based fluorescence spectra measurements with the tags being on top of a fiber and excited by a source at 405 nm. (d) Estimated quantum dot spectra (red) measured by the chip compared against the spectra measured by bench-top spectrometer instrument (black) for Qdot-705 and Qdot-800 from Thermo Fisher Scientific.

D. Resolution and Limit of Detection (LoD)

It can be noted that in a classical spectrometer, the focusing location of any given wavelength remains unchanged regardless of the input light power. In the presented chip-scale design, however, the noise contribution becomes key to the fidelity of the spectral reconstruction. This is because the estimation quality based on the methodology expressed in (13) and (14) is dependent on the noise processes. Low SNR can cause erroneous predicted peaks decreasing achievable spectral resolution. Therefore, collectively two main factors play a key role in determining the resolution of the presented work:

1) Pure optical resolution which is determined by the dimension of the system, the input slit width, the grating aperture, the pixel size and dispersion concave grating designed in compatibility with the CMOS process
2) Signal-to-noise ratio (SNR) which is determined by the input light intensity, loss of the system and the noise processes.

Therefore, to account of both of these effects, the resolution of the spectrum is defined as summation of the FWHM of a C.W. excitation and the spectral prediction error of that given wavelength. This can be expressed as

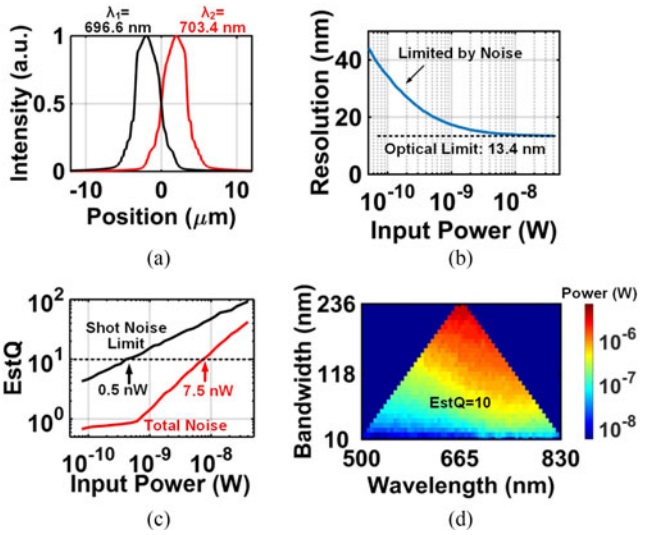


Fig. 19. (a) Simulated H representing FWHM of the image of the slit for two wavelengths $\lambda_1 = 696.6$ nm and $\lambda_2 = 703.4$ nm at focal plane, representing the distance required to separate two images of different wavelengths in order be distinguished. (b) Resolution of the spectrometer as a function of the incident power, calculated from the measured responsivity matrix and varying narrow band incident light with fine steps of around 1nm), and measured circuit noise and photon shot noise. (c) Variation of average $EstQ$ for narrow band incidence across the 500–830 nm wavelength range with incident power calculated from the measured responsivity matrix with Monte Carlo simulations showing a minimum detectable signal of around 7.5 nW for a desired $EstQ = 10$. (d) Minimum detectable power for spectral reconstruction with $EstQ = 10$ as a function of wavelength and bandwidth of the incident spectra.

$$RES_\lambda = FWHM_\lambda + ERR_\lambda \tag{16}$$

where the wavelength prediction error ERR_λ is defined as:

$$ERR_\lambda = \left|\frac{\Sigma[\lambda S_{inc}(\lambda)]}{\Sigma[S_{inc}(\lambda)]} - \frac{\Sigma[\lambda S_{est}(\lambda)]}{\Sigma[S_{est}(\lambda)]}\right| \tag{17}$$

In the case where SNR is very high, the spectral resolution is purely dependent on the optical design, which can be expressed as:

$$\Delta\lambda_{res} = D(\lambda, L) \times \max(2p, H) \tag{18}$$

where D(λ, L) (unit of nm/μm) represents the difference in wavelengths ($\Delta\lambda$) that creates a focal spatial separation of 1μm, and p and H are the pixel size (in μm) and the FWHM (in μm) of the image of the slit at the focal plane respectively. If $H > 2p$, then the spectral resolution is limited by the slit width. Otherwise, the spectral resolution is limited by the finite pixel size.

$D(\lambda, L)$ represents the dispersion capability of the system, and is simulated to be smaller than 1.7 nm/μm. Indeed, this number is inversely proportional to the size of the system. Therefore, scaling the dimension of the whole spectrometer system will proportionally improve the spectral resolution, but at the cost of increasing the loss of the optical modes traveling inside the chip. The value of H is simulated to be around 4 μm with the implemented 2-D concave grating structure and with a 3 μm uniform slit input (Fig. 19(a)). In the case of non-ideal optical structure, we expect the H to be increased by around a factor of

TABLE II
COMPARISON TABLE FOR VARIOUS TYPES OF SPECTROMETER SYSTEMS

	Classical Spectrometers	Mini Spectrometer [9]	Chip-scale Spectrometer with F-P Cavity [14]	Chip-Scale Spectrometer with Random Scatters [15]	This Work
Dispersion Mechanism	Rotating grating in freespace	Concave grating in freespace	Fabry-Pérot cavities of varying thickness	Random scatterers in waveguide	On-chip concave grating in waveguide
Discrete Components	Grating, mirrors, slits, rotator, CCD Imager/PMTs	Concave grating, silts, CCD/CMOS imager	None	External detectors	None
Manufacturing/ Fabrication	Assembly	Custom fab of concave grating, packing of various components	CMOS+post-processing	Custom fabrication	Standard bulk CMOS (no post-processing or custom fabrication)
Dimension	0.1–1 m	20 mm	5–10 mm	0.1 mm (without detectors)	0.5–1 mm
Cost	∼$1000–10000	∼$100–$1000	Determined by custom post-processing	N.A.	<$10 upon mass manufacting
Slits	10–200 μm	50 μm	N.A.	N.A.	3 μm
Spectral Range	Configurable by different gratings	340–780 nm	560–1000 nm	25 nm	500–830 nm
Spectral Resolution	0.1–1 nm	15 nm	10 nm	0.75 nm	13.4 nm
Sensitivity	pW-fW	∼pW	N.A.	N.A.	∼nW

two, which corresponds to twice the pixel size. Therefore, in this design, the effect of finite slit width and pixel width have near equal effects on the spectral resolution. This results in a balanced design and the resolution is calculated to be smaller than 13.6 nm. It should be noted that the spectral resolution is almost inversely proportional to the dimension of the system. Therefore, the resolution can be improved by increasing the overall dimension of the waveguide and grating structure to achieve a sharper focus. The loss, in such a case, can be minimized by increasing the thickness of the waveguide as illustrated before.

Using the measured responsivity matrix and r.m.s. noise voltage, Fig. 19(b) shows the Monte Carlo simulation of the resolution of the spectrometer as a function of the incident power. At large SNR with input power being greater than 10 nW, the resolution is limited by the optical design of the FWHM. The measured value of 13.4 nm matches well to that predicted in (18). When the incident power is reduced, the spectral reconstruction therefore degrades the achievable resolution as shown in Fig. 19(b).

Apart from the spectral resolution, the EstQ is also a metric to characterize the spectral reconstruction quality, as defined in prior section. Expectedly, the noise processes in the circuit limits the sensitivity of the system. Fig. 19(c) shows the Monte Carlo simulation for the average $EstQ$ across the 500–830 nm wavelength range of the narrow-band reconstructed spectra with the measured circuits noise of 2.1 mV and the measured responsivity matrix. As shown in the figure, reducing the input power results in decreased $EstQ$. Input signals with as low as 7.5 nW entering the chip can be resolved maintaining $EstQ > 10$. Evidently, sensitivity can be increased by reducing the circuits noise. As a reference, when the noise sources are purely limited by the photon shot noise, the minimum power for $EstQ = 10$ reaches 0.5 nW.

Expectedly, the minimum detectable power for $EstQ = 10$ will depend on the wavelength as well as the bandwidth of the incident spectrum. This is illustrated in Fig. 19(d) which shows the minimum detectable power as a function of wavelength and bandwidth of the incident spectra for $EstQ = 10$. As the incident spectrum becomes more broadband, the power is less concentrated in the frequency domain, which results in degraded spectral reconstruction quality. Therefore, in order to maintain the same $EstQ$, more incident power is required in the presence of noise for wider band spectra.

Table II shows the comparison of the chip specifications against commercially available instruments and prior works on chip-scale spectrometers with post-fabricated resonators. The spectral resolution of the chip compares well with the miniaturized spectrometers and chip-based works, but a fully integrated spectrometer in a standard CMOS process with no external optics, post-fabrication or custom options can enable new sensing applications where low power, integration, cost and compactness are key.

VI. CONCLUSION

In conclusion, this paper presents a CMOS and nano-optics co-design approach where copper-based nano-optical structures and complex electronics are realized in the same substrate leading to first fully-integrated optical spectrometer in CMOS without any post-processing or custom fabrication. The sub-wavelengths features in copper interconnects are utilized for optical dispersion, wave-guiding, stray-light blocking, as well for traditional circuits routing. Integrated with photodetection circuitry, CTIAs, and CDS, the chip encompasses the entire spectrometer. Extensive measurements with various incident spectra from narrow band, wide band to fluorescent quantum dots of different wavelengths are demonstrated. The CMOS spectrometer demonstrates spectral peak accuracy estimation of 1.4 nm, spectral resolution limit of around 13 nm and minimum resolvable optical power entering the chip of around 7.5 nW. Integration of complex optical systems on chip in the visible and near-IR in CMOS can enable a wide range of sensing

applications in medicine, biomedical, environment, food quality monitoring and industrial applications.

APPENDIX

The appendix details the design of the grating structure. As shown in Fig. 6(a), a single unit cell of the grating comprises of metal strips of 100 nm width (M_5 and M_6) and via layers with 100 nm $\times$ 100 nm dimensions realized with V_4, V_5 and V_6. The spacings between the metal stripes is constant at the minimum allowable distance of 130 nm as shown in Fig. 6(a). Therefore, the y position of each unit cell is determined by

$$Y_i = ip \tag{19}$$

where Y_i is the y position of each unit cell where the origin is set at the center of the Rowland circle, and the pitch p is 230 nm (100 nm metal width and 130 nm spacing). The x position of the unit cells is determined by the following equation

$$\sqrt{(X_i - R)^2 + Y_i^2} + \sqrt{X_i^2 + (Y_i - R)^2} = D_i = D + n_i\lambda_0 \tag{20}$$

where X_i is the x position of each unit cell, n_i is integer, and λ_0 is the center value of the effective wavelength of the 0th order mode propagating inside the waveguide, and D is the optical distance from the source to the center of the grating then to the center of the image $D = (2 + \sqrt{2})R$. This ensures that all unit cells interfere constructively. However, in order to have minimal spectral abbreviation, the value of n_i should be calculated such that (X_i, Y_i) follows the grating circle with radius of 2R as close as possible, which is determined below:

$$n_i = \text{round} \times \left(\frac{\sqrt{(x_{gi} - R)^2 + Y_i^2} + \sqrt{x_{gi}^2 + (Y_i - R)^2} - D}{\lambda_0} \right) \tag{21}$$

$$x_{gi} = R - \sqrt{4R^2 - Y_i^2} \tag{22}$$

where x_{gi} represents the grating's x position with reference to the center of the Rowland circle.

ACKNOWLEDGMENT

The authors would like to thank all members of the Integrated Microsystems Research Lab for technical discussions.

REFERENCES

[1] J. P. Carmo *et al.*, "A review of visible-range Fabry–Perot microspectrometers in silicon for the industry," *Opt. Laser Technol.*, vol. 44, no. 7, pp. 2312–2320, 2012.

[2] J. B. Reeves, G. W. McCarty, and V. B. Reeves, "Mid-infrared diffuse reflectance spectroscopy for the quantitative analysis of agricultural soils," *J. Agriculture Food Chem.*, vol. 49, no. 2, pp. 766–772, 2001.

[3] A. Sakudo, Y. Suganuma, T. Kobayashi, T. Onodera, and K. Ikuta, "Near-infrared spectroscopy: Promising diagnostic tool for viral infections," *Biochem. Biophys. Res. Commun.*, vol. 341, no. 2, pp. 279–284, 2006.

[4] H. Huang, H. Yu, H. Xu, and Y. Ying, "Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: A review," *J. Food Eng.*, vol. 87, no. 3, pp. 303–313, 2008.

[5] M. Safar, D. Bertrand, P. Robert, M. Devaux, and C. Genot, "Characterization of edible oils, butters and margarines by Fourier transform infrared spectroscopy with attenuated total reflectance," *J. Amer. Oil Chem. Soc.*, vol. 71, no. 4, pp. 371–377, 1994.

[6] A. C. Moffat, S. Assi, and R. A. Watt, "Identifying counterfeit medicines using near infrared spectroscopy," *J. Near Infrared Spectrosc.*, vol. 18, no. 1, pp. 1–15, 2010.

[7] Ocean Optics, Largo FL, USA, "Water quality monitoring: chlorophyll a and suspended solids," 2015. [Online.] Available: http://oceanoptics.com/wp-content/uploads/OceanViewWater-Quality-Monitoring.pdf

[8] Ocean Optics, Largo FL, USA, "SERS quantifies toxic melamine in infant formula," [Online.] Available: https://oceanoptics.com/200-pg-melamine-in-infant-formula-on-ram-sers-au/

[9] Hamamatsu, Hamamatsu City, Japan, Micro-spectrometers. [Online.] Available: https://www.hamamatsu.com/eu/en/C12666MA.html

[10] A. Emadi *et al.*, "Fabrication and characterization of ic-compatible linear variable optical filters with application in a microspectrometer," *Sensors Actuators A, Phys.*, vol. 162, no. 2, pp. 400–405, 2010.

[11] J. H. Correia, G. De Graaf, M. Bartek, and R. F. Wolffenbuttel, "A CMOS optical microspectrometer with light-to-frequency converter, bus interface, and stray-light compensation," *IEEE Trans. Instrum. Meas.*, vol. 50, no. 6, pp. 1530–1537, Dec. 2001.

[12] C.-P. Chang and R.-S. Huang, "A 16-channel array-type microspectrometer using integrated Fabry–Perot etalons and lateral pin photodetectors," in *Proc. IEEE Sensors*, vol. 1, 2003, pp. 675–678.

[13] H. Saari *et al.*, "Novel miniaturized hyperspectral sensor for UAV and space applications," *Proc. SPIE*, vol. 7474, 2009, Art. no. 74741M.

[14] N. Tacka, A. Lambrechts1a, P. Soussana, and L. Haspeslagh, "A compact, high-speed and low-cost hyperspectral imager," *Proc. SPIE*, Feb. 2012, Art. no. 82660Q.

[15] B. Redding, S. F. Liew, R. Sarma, and H. Cao, "Compact spectrometer based on a disordered photonic chip," *Nature Photon.*, vol. 7, no. 9, pp. 746–751, 2013.

[16] S.-J. Han, H. Yu, B. Murmann, and N. Pourmand, "Fully integrated optical spectrometer with 500-to-830nm range in 65nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2017, pp. 462–463.

[17] B. Jang, P. Cao, A. Chevalier, A. Ellington, and A. Hassibi, "A CMOS fluorescent-based biosensor microarray," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2009, pp. 436–437.

[18] T. C. D. Huang, S. Sorgenfrei, P. Gong, R. Levicky, and K. L. Shepard, "A 0.18-m CMOS array sensor for integrated time-resolved fluorescence detection," *IEEE J.Solid-State Circuits*, vol. 44, no. 5, pp. 1644–1654, May 2009.

[19] H. J. Yoon, S. Itoh, and S. Kawahito, "A CMOS image sensor with in-pixel two-stage charge transfer for fluorescence lifetime imaging," *IEEE Trans. Electron. Devices*, vol. 56, no. 2, pp. 214–221, Feb. 2009.

[20] R. R. Singh, D. Ho, A. Nilchi, G. Gulak, P. Yau, and R. Genov, "A CMOS/thin-film fluorescence contact imaging microsystem for DNA analysis," *IEEE Trans. Circuits Syst. I, Regul. Papers*, vol. 57, no. 5, pp. 1029–1038, May 2010.

[21] D. Tyndall, B. Rae, D. Li, J. Richardson, J. Arlt, and R. Henderson, "A 100 Mphoton/s time-resolved mini-silicon photomultiplier with on-chip fluorescence lifetime estimation in 0.13 μm CMOS imaging technology," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2012, pp. 122–124.

[22] D. Ho, M. O. Noor, U. J. Krull, G. Gulak, and R. Genov, "CMOS tunable-wavelength multi-color photogate sensor," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, no. 6, pp. 805–819, Dec. 2013.

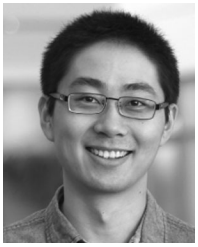
[23] G. Patounakis, K. L. Shepard, and R. Levicky, "Active CMOS array sensor for time-resolved fluorescence detection," *IEEE J. Solid-State Circuits*, vol. 41, no. 11, pp. 2521–2530, Nov. 2006.

[24] D. Stoppa, D. Mosconi, L. Pancheri, and L. Gonzo, "Single-photon avalanche diode CMOS sensor for time-resolved fluorescence measurements," *IEEE Sensors J.*, vol. 9, no. 9, pp. 1084–1090, Sep. 2009.

[25] T.-c. D. Huang *et al.*, "Gene expression analysis with an integrated CMOS microarray by time-resolved fluorescence detection," *Biosensors Bioelectron.*, vol. 26, pp. 2660–2665, 2011.

[26] A. Wang and A. Molnar, "A light-field image sensor in 180 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 257–271, Jan. 2012.

[27] X. Lu, L. Hong, and K. Sengupta, "An integrated optical physically unclonable function using process-sensitive sub-wavelength photonic crystals in 65 nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2017, pp. 272–273.
[28] L. Hong and K. Sengupta, "Fully integrated optical spectrometer with 500-to-830 nm range in 65 nm CMOS" in *Proc. IEEE Solid-State Circuits Conf.*, Feb. 2017, pp. 462–463.
[29] A. Wang, P. R. Gill, and A. Molnar, "An angle-sensitive CMOS imager for single-sensor 3D photography," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2011, pp. 412–414.
[30] T. W. Ebbesen, H. J. Lezec, H. F. Ghaemi, T. Thio, and P. A. Wolff, "Extraordinary optical transmission through sub-wavelength hole arrays," *Nature*, vol. 391, pp. 667–669, 1999.
[31] N. Fang, H. Lee, C. Sun, and X. Zhang, "Sub-diffraction-limited optical imaging with a silver superlens," *Science*, vol. 308, no. 5721, pp. 534–537, 2005.
[32] L. Hong, S. McManus, H. Yang, and K. Sengupta, "A fully integrated CMOS fluorescence biosensor with on-chip nanophotonic filter," in *Proc. Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2015, pp. 206–207.
[33] L. Hong, H. Li, H. Yang, and K. Sengupta, "Fully integrated fluorescence biosensors on-chip employing multi-functional nanoplasmonic optical structures in CMOS," *IEEE J. Solid-State Circuits*, vol. 52, no. 9, pp. 2388–2406, Sep. 2017.
[34] D. Taillaert, P. Bienstman, and R. Baets, "Compact efficient broadband grating coupler for silicon-on-insulator waveguides," *Opt. Lett.*, vol. 29, no. 23, pp. 2749–2751, 2004.
[35] H. Beutler, "The theory of the concave grating," *J. Opt. Soc. Amer.*, vol. 35, no. 5, pp. 311–350, 1945.
[36] X. Wu and K. Sengupta, "A 40-to-330 Ghz synthesizer-free THz spectroscope-on-chip exploiting electromagnetic scattering," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2016, pp. 428–429.
[37] X. Wu and K. Sengupta, "On-chip THz spectroscope exploiting electromagnetic scattering with multi-port antenna," *IEEE J. Solid-State Circuits*, vol. 51, no. 12, pp. 3049–3062, Dec. 2016.
[38] U. Resch-Genger, M. Grabolle, S. Cavaliere-Jaricot, R. Nitschke, and T. Nann, "Quantum dots versus organic dyes as fluorescent labels," *Nature Methods*, vol. 5, no. 9, pp. 763–775, 2008.
[39] Thermo Fisher Scientific, Waltham, MA, USA, Qdot Probes. [Online.] Available: https://www.thermofisher.com/us/en/home/life-science/cell-analysis/fluorophores/qdot-800.html



Lingyu Hong (S'14) received the B.S. degree in physics from Peking University, Beijing, China, in 2012, during which he researched in nanophotonics and plasmonics and received Peking University Academic Excellence Reward and various scholarships. In May 2013, he joined Prof. Sengupta's Laboratory. He is interested in the study, research, and implementation of interdisciplinary knowledge in photonics, electronics, and others for lab-on-chip systems, specifically for biomedical applications. He received the Analog Device Outstanding Student Designer Award in 2015, Qualcomm Innovation Fellowship in 2015, and the IEEE Solid State Circuits Society Pre-doctoral Achievement Award in 2017, and was the Qualcomm Innovation Finalist in 2017.

Kaushik Sengupta (M'12–SM'17) received the B.Tech. and M.Tech. degrees in electronics and electrical communication engineering from IIT Kharagpur, Kharagpur, India, in 2007, and the M.S. and Ph.D. degrees in electrical engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 2008 and 2012, respectively. In 2005, he joined the University of Southern California, Los Angeles, CA, USA, and then joined the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2006, where he was involved in nonlinear integrated systems for high-purity signal generation and low-power RF identification tags. In 2013, he joined the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA, as a Faculty Member. His current research interests include high-frequency ICs, electromagnetics, and optics for various applications in sensing, imaging, and high-speed communication.

Dr. Sengupta received the Young Investigator Program (YIP) Award from the Office of Naval Research in 2017, and the Charles Wilts Prize in 2013 from the Department of Electrical Engineering, Caltech, for outstanding independent research in electrical engineering leading to a Ph.D. He serves on the Technical Program Committee of the IEEE European Solid-state Circuits Conference, IEEE Custom Integrated Circuits Conference and PIERS. He was thrice selected to the Princeton Engineering Commendation List for Outstanding Teaching in 2014, 2016 and 2017. He was a recipient of the IBM Ph.D. Fellowship from 2011 to 2012, the IEEE Solid State Circuits Society Predoctoral Achievement Award in 2012, the IEEE Microwave Theory and Techniques Graduate Fellowship in 2012, the Analog Devices Outstanding Student Designer Award in 2011, the Prime Minister Gold Medal Award of IIT Kharagpur in 2007, the Caltech Institute Fellowship, the Most Innovative Student Project Award of the Indian National Academy of Engineering in 2007, and the IEEE Microwave Theory and Techniques Undergraduate Fellowship in 2006. He was a co-recipient of the IEEE RFIC Symposium Best Student Paper Award (1st prize) in 2012 and the 2015 IEEE MTT-S Microwave Prize.