Image Sensor Architecture

Yuhao Zhu

Department of Computer Science
Department of Brain and Cognitive Sciences
University of Rochester

yzhu@rochester.edu
https://yuhaozhu.com/
https://horizon-lab.org/

Contents

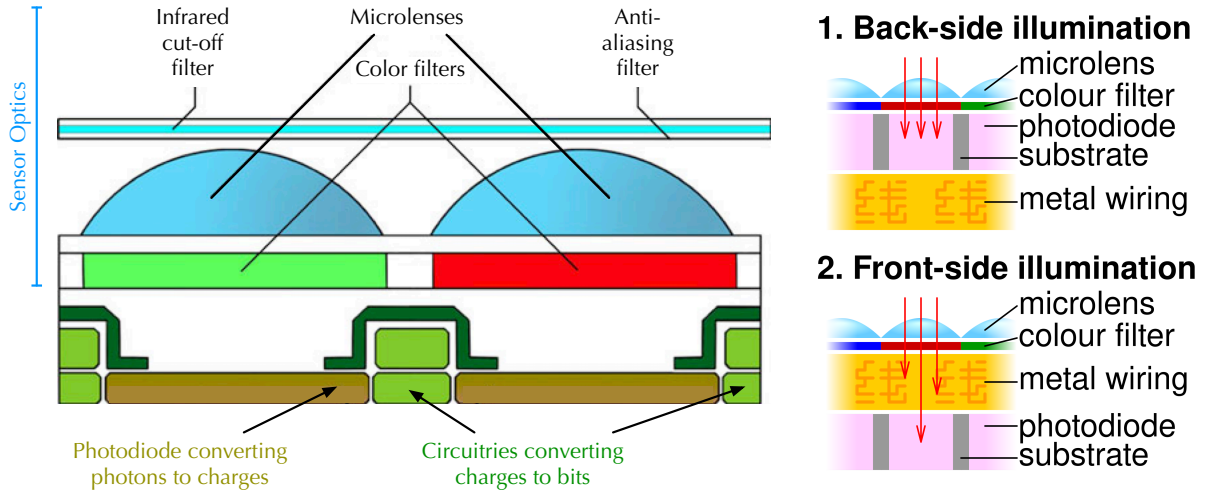


Figure 1: Left: a conceptual, cross-sectional view of the sensor with the optical elements, photodiodes, and the peripheral circuitries; adapted from Adair and Nikitas [2017]. Right: comparison between 1) front-illuminated sensor, where lights have to first traverse through the peripheral circuitries before reaching the light-sensitive photodiodes and 2) back-illuminated sensor, where lights can directly reach the photodiodes; from Cmglee [2019].

1 Overview

The main job of the sensor is to turn optical signals, i.e., the optical image impinging on the sensor plane, to electrical signals, i.e., digital images. This conversion is broken down into two steps, first by converting photons to charges followed by turning charges to digital numbers.

Figure 1 (left) shows a cross-sectional view of the sensor hardware, which has three main components.

- First, there are a set of optical elements sitting on the sensor. These optical elements are not the imaging optics we discussed in the previous chapter, because their main goal is not to form an image.
- Second, under these optical elements are the photodiodes, which turn optical signals carried in photons to electrical signals in the form of electric charges.
- Third, interleaved with the photodiodes is the circuitry that process the output of the photodiodes, turning charges to digital values.

From a computational perspective, we can model an image sensor as a signal processing chain, a transfer function f, that transfers the optical signal to the electrical signal. The optical signal itself has noise, and every step in the signal processing chain not only manipulates the signal itself but also introduces/affects the noise:

$$f : (\mu_p, \sigma_p) \mapsto (\mu_y, \sigma_y), \tag{1}$$

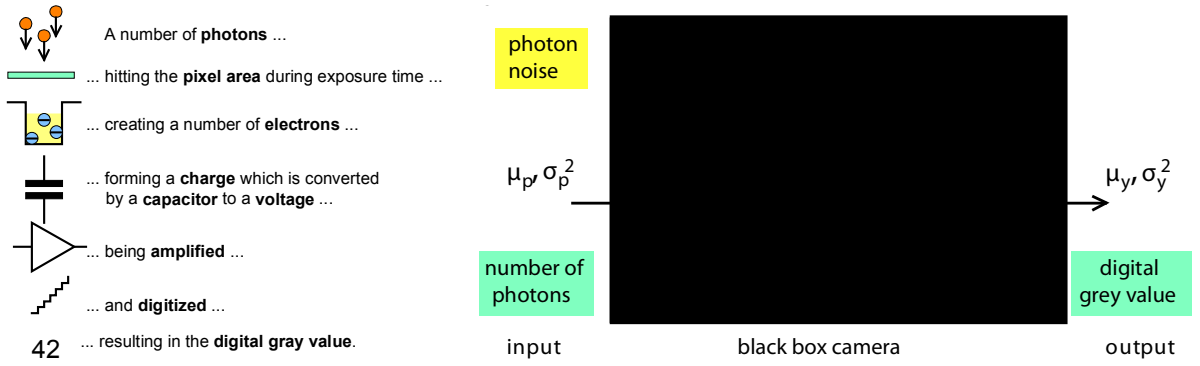


Figure 2: Image sensor can be seen as a chain of processing, or a transfer function, that transfers the optical signal with a mean μ_p and standard deviation σ_p to the electrical signal with a mean μ_y and standard deviation σ_y. The goal of this chapter is demystify this processing chain and build an analytical model for this transfer function. From European Machine Vision Association Standard 1288 [EMVA, 2021, Fig. 1].

where μ_p and σ_p are the mean (signal) and standard deviation (noise) of the input optical signal, respectively, and μ_y and σ_y are the mean and standard deviation of the output electrical signal. The key goal of this Chapter is to build a quantitative model of f.

There are two ways the pixels and the wires that read out the pixel outputs are physically arranged, shown in Figure 1 (right). In the **back-side illumination** (BSI) arrangement, the wiring of the circuitries are behind the photodiodes, which directly interface with the lights. In the **front-side illumination** (FSI) arrangement, the metal wiring sits between the light and the photodiodes. This means lights could be absorbed and scattered through the metal layer before reaching the photodiodes, reducing the chance of a photon being properly captured. While earlier image sensors use FSI, because it is easier to manufacture, almost all commercial image sensors use BSI now [Swain and Cheskis, 2008].

FSI is actually quite similar to the structure of human eyes, where, if you call, the photoreceptors are "hiding" behind other retinal neurons such as the retinal ganglion cells, which are functionally the last layer of retinal processing but anatomically sitting at the first layer on the retina. Different from the FSI sensor, however, the non-photoreceptor neurons on the retinal do very little to lights: they do not absorb or scatter lights much and can be generally thought of as transparent. Metal wires, of course, disrupt incident photons significantly.

2 From Photons to Charges and Digital Numbers

We will talk about how optical signals are first converted to electrical signals in the form of charges, and then talk about how the charges are detected, at which point the electrical signals are manifested as voltage potentials. The voltage potentials are then quantized as Digital Numbers, which are the raw pixel values. We will focus on the basic building blocks that enable these

conversions and leave to the next chapter to discuss how these building blocks are connected in a global sensor architecture. The discussion here assumes monochromatic sensing without noise. We will talk about color sensing and the noise issue later.

2.1 Photons to Charges

What turn optical signals to electrical signals is the light-sensitive photodiode in a pixel. A photodiode is a p-n junction made of silicon, a semiconductor material. When a photon hits silicon and is absorbed, an electron from the silicon *might* be freed/emitted, transforming optical signals to electrical signals. This is called the **photoelectric effect** [Einstein, 1905a,b], the discovery of which won Albert Einstein his Nobel Prize.

In particular, when a photon is absorbed, if its energy is greater than or equal to the **work function** ϕ of the material, which the minimum energy needed to free an electron from the surface of the material, the photon can transfer its energy to an electron and free the electron. A photon's energy is given by the **Planck's relation**:

$$\mathcal{E} = hf = \frac{hc}{\lambda}, \tag{2}$$

where h is the Planck constant, f is the photon frequency, and c is the speed of light. So if $hf > \phi$, an absorbed photon can free an electron. Interestingly, the residual energy $hf - \phi$ becomes the kinetic energy of the electron, so a photon with a shorter wavelength (i.e., higher frequency) would allow the emitted electron to move faster.

It is clear that there is a frequency threshold ϕ/h, lower than which a photon would never be able to free an electron. Higher than the threshold, there is generally a one-to-one mapping between an absorbed photon and an emitted electron: an absorbed photon always frees an electron. Since the work function of silicon is about 1.1 eV (electron volt), absorption of photons with wavelength longer than 1,100 nm would not emit any electron.

A key figure of merit in image sensing is the notion of **quantum efficiency** (QE), which is the ratio between the number of electrons collected and the number of incident photons:

$$QE = \frac{\text{\# of electrons collected}}{\text{\# of incident photons}}. \tag{3}$$

Figure 3 (left) shows the QE spectrum of an image sensor in the Hubble Space Telescope. It might come as surprise that QE is lower than 1 (even for wavelengths well within the 1,000 nm threshold) and is actually wavelength dependent: shouldn't every absorbed photon (within the wavelength threshold) always free an electron? There are two reasons.

First, the denominator in the QE definition is the number of *incident* photons, not the number of absorbed photons. Not all photons that hit the photodetector will be absorbed. Figure 3 (right) shows the spectral absorption coefficient σ (unit 1/cm) of silicon on the left y-axis, and the right y-axis shows the corresponding mean free path l (i.e., the expected length a photon can travel within silicon before being absorbed) at different wavelengths; recall from

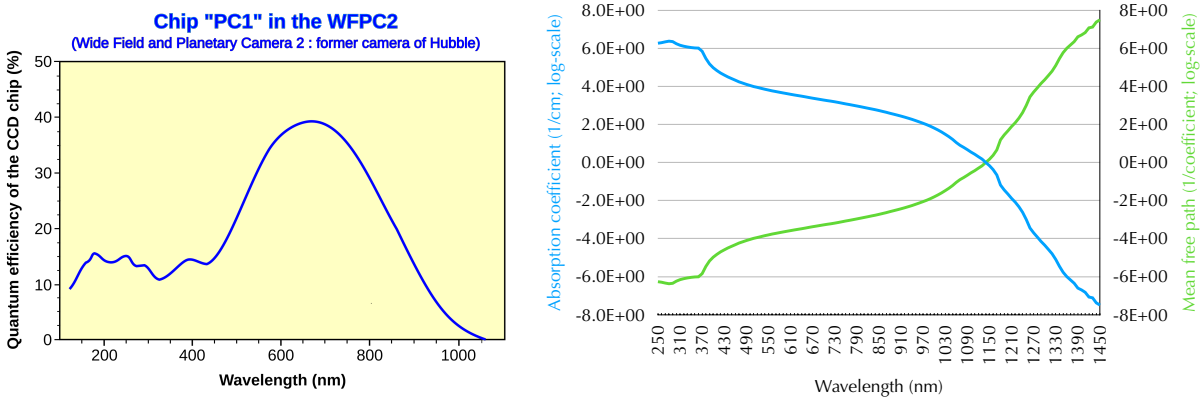


Figure 3: Left: quantum efficiency of a sensor on the Hubble Space Telescope; from Eric Bajart [2010] with data from Biretta and McMaster [2008, Fig. 4.2]. Right: silicon absorption coefficient (left axis) and mean free path (right axis) as a function of wavelength; data from Green and Keevers [1995].

the material chapter that $l = 1/\sigma$. We can see that absorption is strongest for the blue-ish lights but decays very rapidly toward the longer wavelengths. This definition of QE is different from how QE is defined in human vision. Recall from the photoreceptor lecture; there QE is the probability of pigment excitation once the pigment actually absorbs a photon; there, the QE of photopigment is roughly two-thirds and is not wavelength-sensitive.

Second, the nominator in the QE definition is the number of *collected*, not emitted, electrons: even if an electron is freed by an absorbed photon, that electron might not actually be collected and contribute to the electrical signal. Depending on where the electrons are freed, some of them need go through a random walk (think of it as a Brownian motion) before being collected, and you can imagine some electrons can be recombined with the holes during the walk.

Given QE, the total number of emitted electrons after an exposure time T is given by:

$$N = \int_{\lambda} QE(\lambda) Y(\lambda) \mathrm{d}\lambda \tag{4a}$$

$$= \int_{\lambda} QE(\lambda) \frac{\Phi(\lambda) T \lambda}{hc} \mathrm{d}\lambda, \tag{4b}$$

where $Y(\lambda)$ is the number of photons incident on a photodiode at a particular wavelength λ during the exposure time T (assuming Y is invariant during T here). $Y(\lambda)$ is related to the spectral power distribution (SPD) of the incident light $\Phi(\lambda)$ by: $Y(\lambda) = \frac{\Phi(\lambda)T\lambda}{hc}$, where $\Phi(\lambda)T$ is spectral energy distribution. Using the Planck's relation (Equation 2), we can turn the spectral energy distribution to the spectral quantity distribution: $\frac{\Phi(\lambda)T\lambda}{hc}$.

Note that we define QE for the photodiode itself: the denominator in Equation 3 refers to the number of photons incident on the photodiode, not that enters the camera system. This is an important distinction, between many photons that enter the camera would not even make their

ways to the photodiode; some of them are reflected at the lens surfaces and others are absorbed by the various filters (Chapter 4). In many contexts, the QE is reported with respect to the entire camera system, where the denominator *is* the number of photons entering the camera. Always ask what the precise definition of a QE is when reading the literature.

2.2 Measuring Charges

Basic Principle

Now that we have turned photons to charges — the freed electrons move to the n region and the holes move to the p region of the p-n junction, the next step is to measure the charges. The basic principle of doing so is using a capacitor: we use the electrons to discharge a capacitor with a known capacitance; by measuring the voltage difference before and after the discharge, we can then estimate the number of emitted electrons:

$$\Delta V = \frac{Q_{SIG}}{C_{FD}} \times g = \frac{Nq}{C_{FD}} \times g, \tag{5}$$

where Q_{SIG} is the charge in the signal stored in the capacitor, C_{FD} is the capacitance, and g is the voltage gain of whatever device is used to read out the voltage (usually a source follower; see later). Q_{SIG} itself is the product of N, the number of charges in the signal, and q, the elementary charge.

We can see that once we can measure ΔV, we can get an estimate of N. Why do we care about N? Intuitively, the incident light luminance is positively related to N: more incident photons means higher luminance. Luminance L, if we are interested in only grayscale, monochromatic imaging, is ultimately what we want to estimate.

It is important to realize that the actual relationship between L and N is not linear. We know that luminance is defined as:

$$L = \int_{\lambda} V(\lambda)\Phi(\lambda)\mathrm{d}\lambda, \tag{6}$$

where $V(\lambda)$ is the luminance efficiency function (LEF) and $\Phi(\lambda)$ is the SPD of the incident light. Taking Equation 6 and Equation 4b together, we can see that given N we cannot quite estimate L, because L depends on $\Phi(\lambda)$, but estimating $\Phi(\lambda)$ from N is an under-determined problem, as Equation 4b shows. To be exact, L does not necessarily scale linearly with N — it does not even necessarily scale positively with N, but it is perhaps not terribly wrong to informally say a higher charge count means a higher luminance in the scene. We will return to this problem when we discuss color sensing, too.

4T Design

The photodiode (PD) technically acts as a capacitor itself (the n-side neutral region holds electrons and the p-side neutral region holds holes), so we could simply use the PD for that

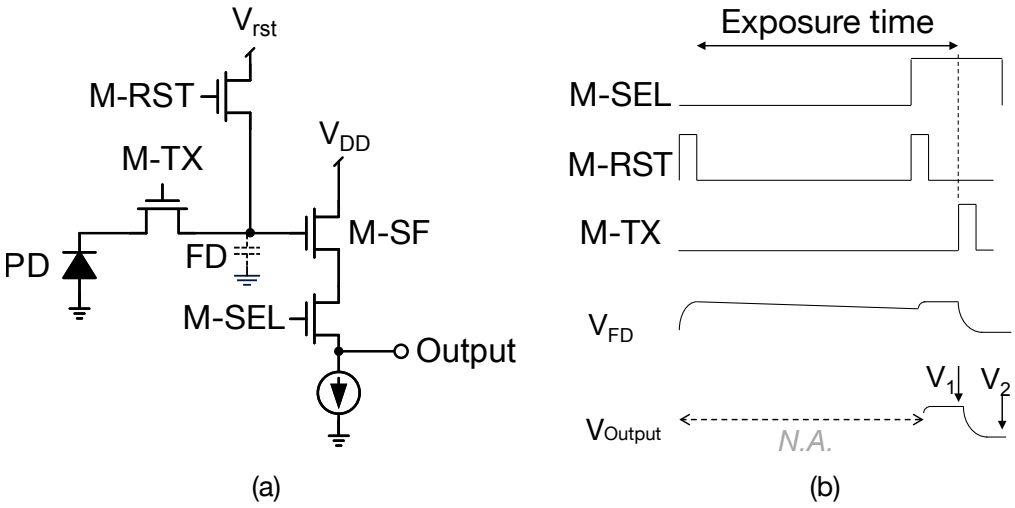


Figure 4: (a): circuit diagram of a typical 4T pixel design; adapted from Ma [2024, Fig. 2.5(a)]. (b): timing diagram of operating a 4T pixel.

purpose. This is indeed how an earlier pixel design works, which we will return to shortly. Modern pixels actually transfer the charges from the PD to a separate measurement node, which we focus on here.

Figure 4(a) shows the circuit diagram of a typical pixel design that detects and measures the charges. The design has a PD and four transistors so it is usually called the 4T design. The `M-TX` switch controls the transfer of the charges accumulated in the PD to the Floating Diffusion (FD)[1], another capacitive area and is sometimes called the *measurement node* or the *sense node*, because that is where the charges are actually being measured. The FD is connected to the NMOS Source Follower (SF) transistor `M-SF`, where the gate terminal is its input and is connected to the FD voltage, the drain is connected to the supply voltage, and the source is the output that faithfully follows/transfers the input with a gain of about 0.9 (g in Equation 5).

The sequence of operation goes roughly like the following, and Figure 4(b) shows the corresponding timing diagram:

1. Before the exposure, we turn on the `M-RST` switch *and* the `M-TX` to drain the charges (electrons) at the PD, which will also, as a byproduct, drain the charges in the FD, reseting their voltage potentials both to V_{RST}. Resetting the FD voltage at this step is of no functional use, as we will shortly see.

2. We then turn off `M-RST` and `M-TX`, and the exposure begins, during which the charges are collected inside the PD. We can see from Equation 5 that in order to measure the charges

[1]For the charges collected in PD to be transferable to the FD, the photodiode needs to be "pinned", which means there is another layer of p+ implant above the p-n junction pinned to the ground (0 V). Such a PD is also called the Pinned Photodiode, or PPD [Fossum and Hondongwa, 2014].

we need to measure the voltage difference *at the FD node* before and after the charges are transferred. So toward the end of the exposure, we turn on the `M-RST` switch again while, importantly, keeping the `M-TX` switch off. This would allow us to reset the FD voltage to V_{rst}, which will be measured through `M-SF` as V_1 in Figure 4(b)[2].

3. We then turn on the `M-TX` switch, which transfers the charges from the PD to the FD. After that, we turn off `M-TX` and read the voltage from `M-SF` for the second time, this time for the voltage at FD after the charge transfer. This is the V_2 in Figure 4(b). The difference between V_1 and V_2 is the ΔV in Equation 5.

As we can see, we read the voltage of the FD twice to obtain the voltage difference caused by the charges collected during the exposure. This is called **Correlated Double Sampling** (CDS), which turns out to also be very important to mitigate many noise sources, which we will discuss later.

To read out the voltage from the SF, the `M-SEL` switch needs to be turned on, which is omitted from Figure 4(b) for simplicity. As we will shortly see in Chapter 3, in most cases (although not all), pixels are read out row by row, so the `M-SEL` switches of all pixels in the same row are connected to the same signal, usually called the row select signal.

The timing diagram in Figure 4(b) is illustrative of the major operations (omitting `M-SF`) but not drawn to scale. The exposure time is usually at tens of milliseconds scale (e.g., 30 FPS means roughly a 33.3 ms exposure time) but the timescale to operate the transistors/switches is at the microsecond level. Also observe, in Figure 4(b), that during the exposure the voltage at the FD (V_{FD}) slowly reduces from V_{rst} after the first reset — because of charge leakage in the FD, just like how DRAM cells leak. This is why we need the second reset to bring the voltage at FD back to V_{rst} before charge transfer. This is also why we say the first reset is of no functional use to the FD (but of course very important to the PD because we want the PD to collect only electrons emitted from the current exposure).

4T APS vs. 3T APS vs. PPS

The (4T) pixel design describe above is called an **Active Pixel Sensor** (APS) design, first conceived by Noble [1968] (see Fossum [1993] for a more modern perspective). An APS has a per-pixel SF (a common-drain amplifier) that "actively" reads out the signal for each pixel by turning its charges to voltage. A simpler and earlier version of the APS design uses only three transistors (3T) without the `M-TX` gate. Figure 5 (left) compares the 4T APS with the 3T APS. Without the transfer gate, the PD is used as the measurement/sensor node itself, so the C_{FD} in Equation 5 is effectively the capacitance of the PD itself.

The 3T APS simplifies the pixel design and, thus, increases the fill factor (without the microlenses). It, however, generally suffers from lower signal to noise ratio (SNR), for a variety of reasons. For instance, the PD has a large inherent photodiode capacitance, so the signal (ΔV in Equation 5) read from the PD is low, making is more vulnerable to noise. In contrast, we get

[2] V_1 and V_{rst} technically are ever so slightly different because the charges might be leaking between resetting and read out.

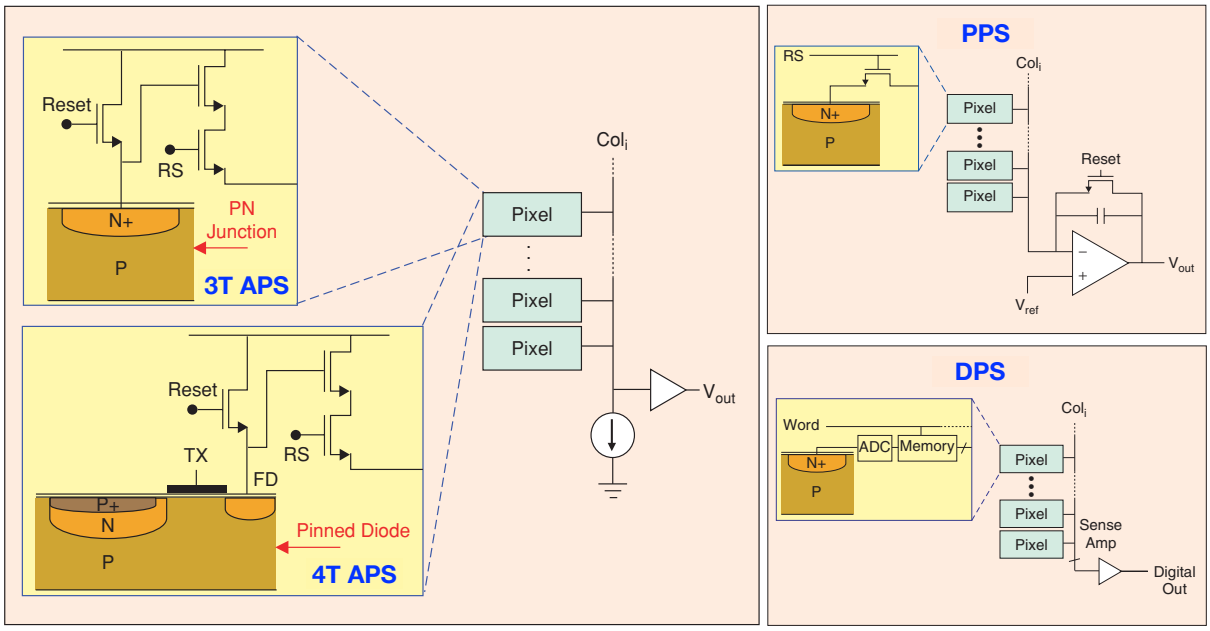


Figure 5: Left: 3T APS vs. 4T APS. Top right: Passive Pixel Sensor (PPS). Bottom right: Digital Pixel Sensor (DPS). Adapted from El Gamal and Eltoukhy [2005, Fig. 5, 10, 11].

to control the FD in the 4T APS, which can be made to have a much lower capacitance, leading to a higher SNR. The CDS for 3T APS is also much less effective in suppressing noise, as we will discuss later.

A precursor to APS was the **Passive Pixel Sensor** (PPS), first suggested in Weckler [1967] and Dyck and Weckler [1968]. A PPS has only one transistor, as shown in the top-right panel in Figure 5. The PPS has no SF that reads out voltage from the PD charges. Instead, the charges (not voltage) in the PD "passively" flow through a column bus and are turned to voltage there through a charge amplifier [Aoki et al., 1982]. This leads to much worse noise profile, because of the large (parasitic) capacitance of the column bus. The SF in APS acts as an active amplifier, which isolates the sense node (whether it is the PD or the FD) from the large column bus capacitance, providing a much higher output current and lower output impedance than a PD does and, thus, improving the SNR [Kozlowski et al., 1998; Ohta, 2020, Chpt. 2.5].

Electronic Shutter

Ideally when we are not capturing lights the photodiodes should not be exposed to lights. This is achieved by a shutter. **Mechanical shutters** do so by physically blocking lights. The sensor is *not* exposed to light normally, blocked by the shutter. The shutter then mechanically opens to expose sensor to light. There are many types of mechanical shutters, of which the most popular one is the focal plane shutter shown in Figure 6 (a). The shutter has two curtains that move in sync with a gap that allows lights in. The size of the shutter opening and the speed of the

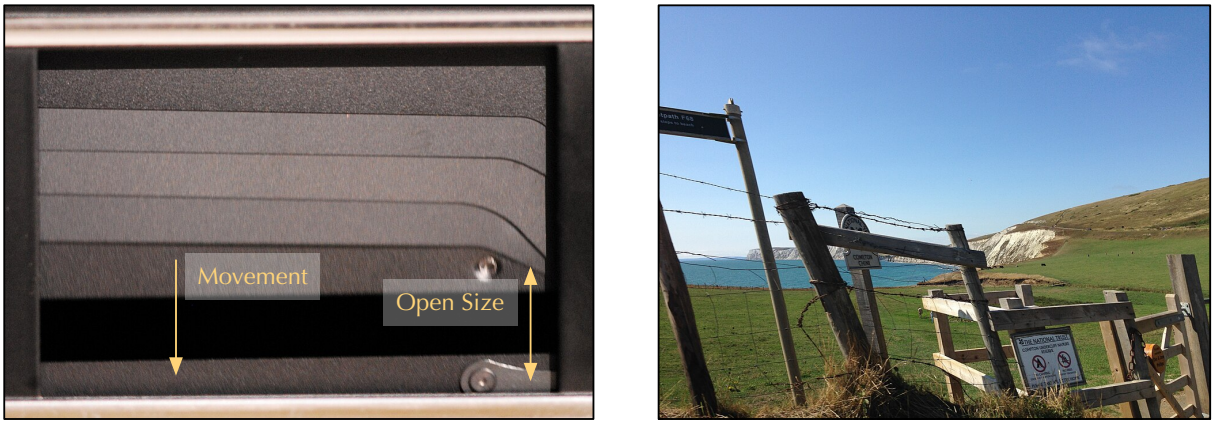


Figure 6: (a): a mechanical focal-plane shutter, which is inherently a rolling shutter; adapted from Ommnomnomgulp [2008]. (b): rolling shutter artifact; from BrayLockBoy [2018].

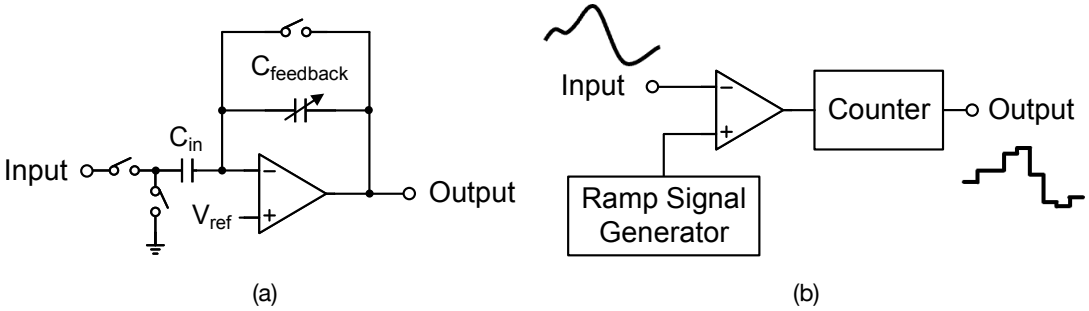


Figure 7: (a): analog CDS and programmable amplifier; from Ma [2024, Fig. 2.5(b)]. (b): a single-slope ADC typically used in image sensors; adapted from Ma [2024, Fig. 2.5(c)]

movement dictates the exposure time: larger opening and slower speed mean longer exposure time. This is called a focal plane shutter because the shutter is located in front of the focal plane (sensor). There is also the leaf shutter, which is usually located at the aperture plane with the lenses.

The 4T pixel design above essentially implements an **electronic shutter** (ES). With an ES, we expose photodiodes to lights *all the time*. The way we mark the start of the exposure is through the `M-RST` switch, which resets the PDs, and the way we mark the end of the exposure is through the `M-TX` switch, which transfers the PD charges for measurement. The time difference between these two steps dictates the exposure time. As you can imaging, the shutter speed (inverse of the exposure time) of an electronic shutter can be much faster than that of a mechanical shutter, since there are no mechanical moving parts.

2.3 Read-out Circuitry

Following the pixel circuitry is the read-out circuitry, which usually has two main components: the programmable-gain amplifier and the Analog-to-Digital Converter (ADC). Figure 7 illustrates the common, simplified designs of the two components.

The amplifier is there to amplify the voltage read from the pixel, and the **gain** of the amplifier is programmable. A programmable gain is useful in imaging and photography to artificially shorten or extend the exposure time (e.g., through the ISO setting in a digital camera). The particular design shown in Figure 7(a) combines CDS with a classical amplifier design with two capacitors. Specifically, the two voltages read-out from the FD (one right after the reset and the other right after the charge transfer) are sampled by the C_{in} capacitor sequentially, which essentially performs an analog-domain subtraction that is required by CDS. The voltage difference is then amplified with a gain $\frac{C_{in}}{C_{feedback}}$. $C_{feedback}$ is usually programmable, allowing us to control the gain.

The amplified voltage difference then goes through an ADC to obtain the digital value. There is a huge amount of ADC designs [Murmann, 2014]. The design that is commonly used in image sensors is one of the Single-Slop (SS) design, whose simplified diagram is shown in Figure 7(b). An SS ADC consists of a comparator, a ramp signal generator, and a counter. The ramp generator provides a monotonically increasing or decreasing ramp signal, which is compared with the to-be-quantized analog signal (output of the amplifier). At every clock cycle, the comparator compares the two inputs while the counter increments. When the two input signals cross, the counter value is recorded and represents the quantized digital value of the analog signal.

The designs in Figure 7 perform CDS in the analog domain (through C_{in}). In many image sensors today, the CDS is performed in the digital domain after the ADC [Nitta et al., 2006]. You would think that such a design might require twice the ADC overhead plus the additional digital subtraction overhead. In reality, the design is quite clever. The ADC would first quantize the first sample (before reset) and the resulting counter value represents the digital value of the first sample. For the second sample, instead of counting from scratch, we would simply turn the counter around so that it counts backward. At the end, the counter value is naturally the digital difference of the two samples.

3 Global Architecture

We have discussed the individual building blocks that are need for a pixel to turn lights to digital values, but how are they put together in an actual image sensor supporting tens of millions pixels? This chapter talks about the global architecture of an image sensor. We will start with a common architecture followed by other variants.

3.1 Column-Parallel Readout

Figure 8 (a) shows a typical arrangement, where pixels are organized as a 2D array, just like a (DRAM/SRAM) memory array, and each column has an amplifier and ADC shared by all

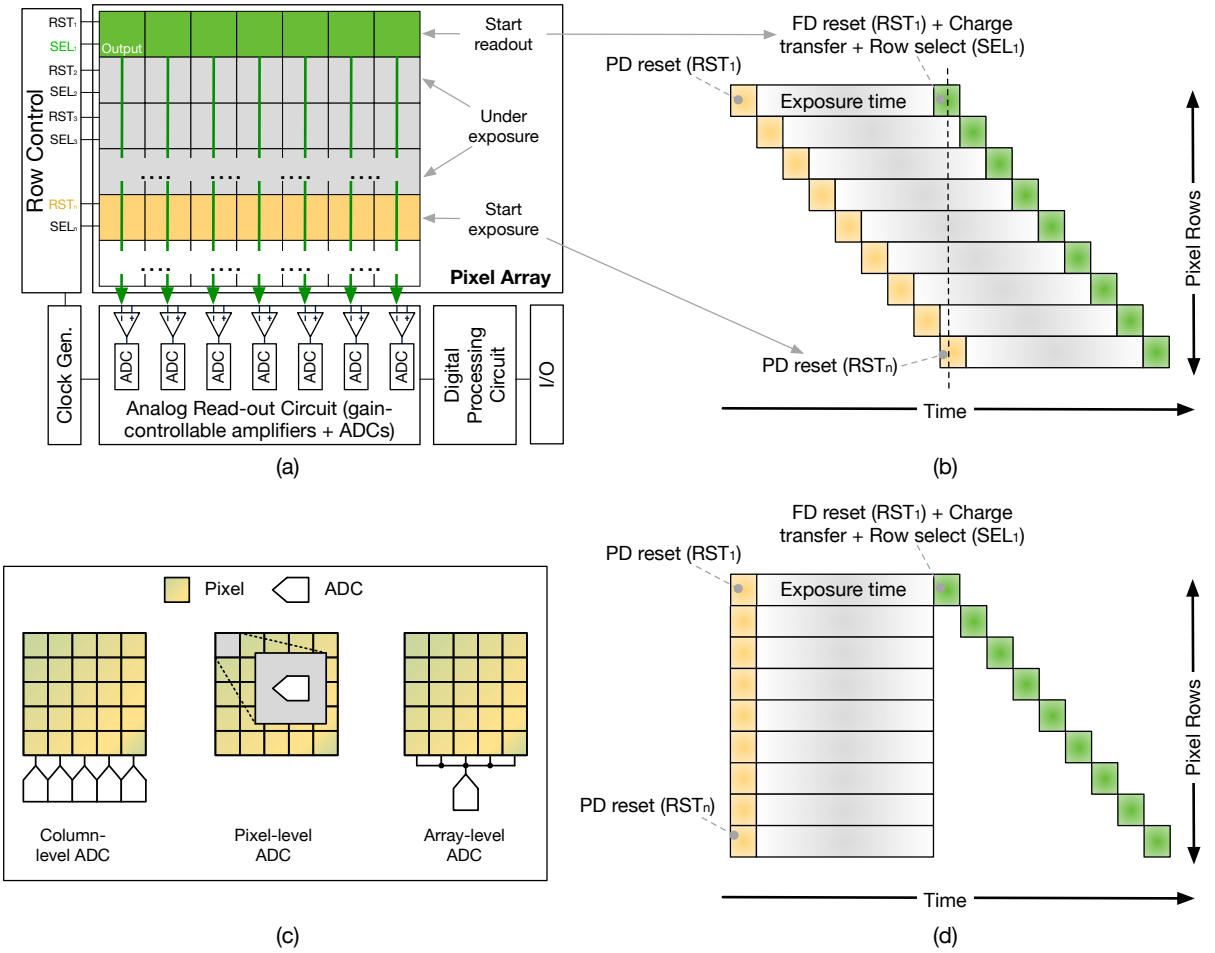


Figure 8: (a): the block diagram of a typical rolling-shutter image sensor with column-level amplifiers and ADCs, where pixels in the same column share the same amplifier and ADC; pixels are exposed and read-out row by row under the control of the `RST` signal (connecting to the `M-RST` switches) and the `SEL` signal (connecting to the `M-SEL` switches) (for simplicity, we omit the per-row `TX` signal, which connects to all the `M-TX` switches in the same row); (b): timing diagram operating the image sensor in (a) with a rolling shutter; technically the FD reset should be overlapped with the exposure time, but is lumped into the readout box for simplicity. (c): comparison of column-level ADC used in (a) with pixel-level ADC and array/chip-level ADC. (d): timing diagram operating the image sensor in (a) with a global shutter.

the pixels in that column. That is, the `Output` pin in Figure 4 of all the pixels in the same column are connected to the same amplifier and ADC. The read-out circuit is then connected to digital processing circuitry, which could potentially perform simple image-space operations such as downsampling, scaling, rotation, etc. There is also an I/O unit that transfers the pixels to the host processor, usually through the MIPI-CSI interface, and transfers commands/configuration data from the host processor, usually through the I2C interface, which has a much lower bandwidth that MIPI (Kb/s vs. Gb/s).

The pixels in the pixel array are addressed row by row through a row scanner logic, shown on the left of Figure 8 (a). Pixels in the same row share three external signals: a reset signal `RST`, which is connected all the `M-RST` transistors in the row, a row-select signal `SEL`, which is connected to all the `M-SEL` transistors of the same row, and a transfer signal `TX` (omitted in the figure) connected to all the `M-TX` switches in the same row.

The operating sequence of the pixel rows is shown in Figure 8 (b); the times are not drawn to scale. Each row of pixels goes through the PD reset, exposure, and readout phases under the control of the three external signals (`RST`, `SEL`, and `TX`). Importantly, the three phases are pipelined across rows. That is, while the first row is being exposed, we can start resetting the PDs for the subsequent rows and preparing them for exposure. For instance in the concrete example of Figure 8 (a), the first row is starting the read-out sequence, the n^{th} row is starting the exposure, while all other rows in-between are currently under exposure. While the exposure times of different rows can overlap, their readout sequences cannot — because pixels in the same column but different rows share the same read-out circuitry.

We can see that the way the pixel array is addressed and operated is similar to how a memory array (e.g., SRAM/DRAM) is, where the data in an entire row is accessed at once. However, since the pixel rows are operated strictly sequentially (unless random sampling is needed [Feng et al., 2024]), the row scanner logic does not need a decoder, which supports random accesses that a typical memory array would need. Instead, one can usually use parallel shift registers to generate the three external signals row by row.

3.2 Rolling vs. Global Shutter

The timing diagram suggests that pixels in different rows technically have slightly shifted exposure times, inherently using a **rolling shutter**. The mechanical focal-plane shutter shown in Figure 6 (a) is inherently a rolling shutter. Rolling shutters introduce noticeable artifacts; one such example is shown in Figure 6 (b), where photo was taken by a camera traveling in a car driving at about 50 mph. As a result, the fence and gate appear slanted, because vertical parts of these objects are taken at different times. Such an artifact is much less visible for more distant objects, such as the cliff (can you reason about why?).

Global shutters address the rolling shutter artifacts by exposing all pixels at the same time. Figure 6 (d) shows the timing diagram of a global shutter sensor; compare that with that of the rolling shutter sensor in Figure 6 (a). All the PDs are reset at the same time and have the same exposure duration.

The pixels are still read-out row by row due to the column-level design of the read-out circuitry. This means the pixel values have to be temporarily held in some form of analog buffer

after exposure and before they are read-out. One could certainly use the FD for this analog buffer — with the caveat that the this prevents the PD from starting a new exposure cycle. This is because starting a new exposure requires resetting the PD, which would also reset the corresponding FD, as shown in Figure 4 (a). For that reason, it is common to implement an additional analog buffer inside each pixel. The buffer can be implemented either in the charge domain before the FD [Yasutomi et al., 2011; Sakakibara et al., 2012; Tournier et al., 2018; Kumagai et al., 2018b; Yokoyama et al., 2018; Kobayashi et al., 2017] or implemented in the voltage domain after the FD [Kondo et al., 2015; Stark et al., 2018; Miyauchi et al., 2020].

3.3 Pixel-Parallel and Chip-Level Readout

We can also arrange the read-out circuitry differently, as illustrated in Figure 8 (c). For instance, we could have a *per-pixel* (gain-controllable) amplifier and ADC and, consequently, a per-pixel digital memory. This essentially allows each pixel to directly output digital values, giving rise to the so called **Digital Pixel Sensor** (DPS) design, which was first reported in Fowler et al. [1994] and is recently gaining tractions [Liu et al., 2019]. The bottom-right panel in Figure 5 shows the pixel design diagram of a DPS, where the in-pixel memory can be, for instance, a 6T SRAM cell. In this case, the entire pixel array is indeed like an SRAM array.

DPS increases the pixel design complexity and pixel sizes, which, without microlenses, reduces the fill factor. This can, however, be alleviated with a stacked design, which we will get to in Chapter 3.5. The main advantage of the DPS is that it massively increases the readout bandwidth due to pixel-parallel ADCs, which could shorten the frame latency when using a global shutter (see Figure 8 (d)), especially when short exposure time is desirable (e.g., high frame rate or "snap-shot" photography).

Yet another read-out arrangement is to have a single gain-controllable amplifier and ADC for the entire pixel array. This is shown in Figure 9 (b). In this case, we not only need logic to scan rows one by one, but also, for each row, scan its columns one by one (e.g., through shift registers). This arrangement is not commonly used due to its slow read-out speed, but is the only option for sensors based on the Charge-coupled Devices (CCD), a design that is different from all the designs we have discussed so far and is our focus next.

3.4 CMOS vs. CCD Sensor

All the sensor designs we have covered so far are called Complementary Metal-Oxide-Semiconductor (CMOS) sensors, because they heavily rely on circuitries implemented using the CMOS techonlogies. CCD sensor is the other major category of sensor design, first reported in Boyle and Smith [1970]. Both CCD and CMOS sensors use silicon to implement the PDs (although the specific implementations can differ [Nakamura, 2006, Chpt. 3.1.2]). The main difference lies in how the charges generated by the PDs are read out. See Fossum [1993, 1997]; El Gamal and Eltoukhy [2005] for the historical background and comparisons.

A CCD sensor directly reads out charges from pixels by *shifting* the collected charges, first row by row and, for each row, column by column to a single, array-level SF amplifier (and potentially a gain-controllable amplifier and ADC afterwards). This architecture is shown in

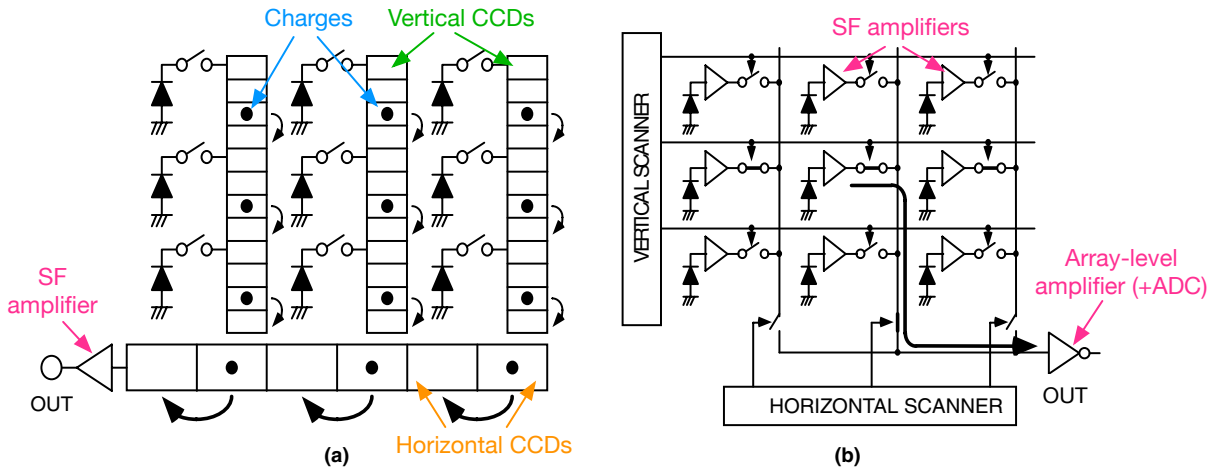


Figure 9: (a) charge-shifting read-out architecture for CCDs; (b) read-out architecture for CMOS image sensors with a global, array-level amplifier. Adapted from Nakamura [2006, Fig. 3.5].

Figure 9 (a); compare that with the CMOS architecture in Figure 9 (b), where the charges are converted to voltages *within* the pixels, and it is the voltage potentials that are being read-out from the pixel array by *addressing*, rather than shifting across, individual rows.

The key to a CCD sensor is the charge-couple devices themselves. A CCD is a set of connected MOS capacitors that store and transfer, between them, charges [Hu, 2009, Chpt. 5], invented by Willard Boyle and George E. Smith [Boyle and Smith, 1970], who won the Noble Prize in Physics in 2009. In a CCD image sensor, the CCDs are connected to the PDs. After the exposure, all the PDs simultaneously transfer their charges to the corresponding vertical CCDs. The vertical CCDs in the same column then act as a shift register, transferring the charges downward to the horizontal CCD at the bottom of the chip. When a row of charges reach the horizontal CCDs, the charges are then transferred horizontally (again, in a shift-register fashion) to the SF amplifier, which turns charges to voltage.

Given this signal read-out architecture, it is perhaps unsurprising to see that CCD sensors inherently support global shutters: the CCDs used for shifting charges naturally store the charges temporarily during the read-out.

CCDs are fabricated using process technologies that are optimized for charge transfer and that are incompatible with the CMOS technologies. In contrast, the read-out architecture of the CMOS sensors can be fabricated using CMOS technologies. This is a huge advantage because non-imaging logics such as control (e.g., clock generation) and analog/digital processing (e.g., ADC, image processing, computer vision tasks) are also based on CMOS technologies. Such logics, in CCD sensors, need to be implemented on a separate chip that interfaces with the CCD chip, rather than integrated with the pixel array on the same chip in a CMOS image sensor.

As modern CMOS technologies mature and gradually take over the semiconductor industry, CMOS image sensors have become more appealing. The main advantage of the CCD sensors is their high SNRs. CCD sensors do not have active devices during read-out and, thus,

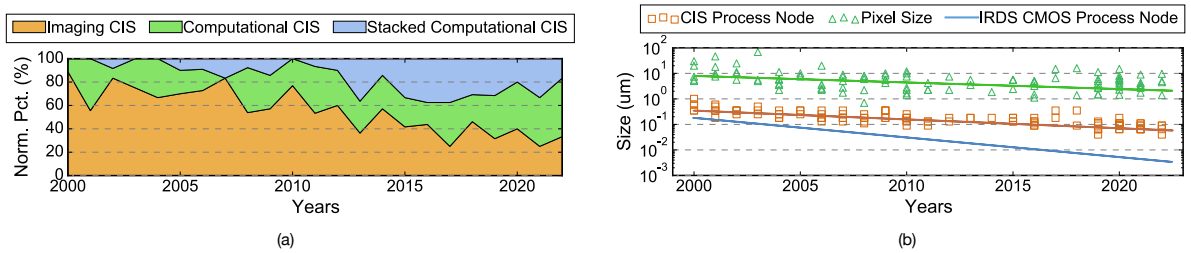


Figure 10: (a) Percentage of conventional CIS, computational CIS, and stacked computational CIS designs from surveying all ISSCC and IEDM papers published between Year 2000 and 2022. Increasingly more CIS designs are computational. (b) CIS process node always lags behind conventional CMOS process node. This is because CIS node scaling tracks the pixel size scaling, which does not shrink aggressively due to the fundamental need of maintaining photon sensitivity. From Ma et al. [2023, Fig. 1, 3].

avoid/minimize many sources of noise that CMOS sensors are vulnerable to, a point we will return to when discussing noise modeling[3]. Because of that, while consumer cameras today mostly use CMOS sensors, CCD sensors are still use widely used in many scenarios where imaging quality is critical, e.g., scientific imaging. For instance, many telescopes for astrophysics (e.g., Sloan Digital Sky Survey) still use CCD sensors.

3.5 Computational and Stacked CMOS Image Sensors

Because the imaging circuitries and the logic processing circuitries both use the CMOS process technologies, a clear trend in CMOS Image Sensor (CIS) design is to move into the sensor computations that are traditionally carried out outside the sensor, which gives rise to the notion of **Computational CIS**.

CIS Scaling Trends

Figure 10 (a) shows the percentage of computational CIS papers in International Solid-State Circuits Conference (ISSCC) and International Electron Devices Meeting (IEDM), two premier venues for semiconductor circuits and devices, from Year 2000 and Year 2022 with respect to all the CIS papers during the same time range. The trend is clear: increasingly more CIS designs integrate compute capabilities.

A key reason why we could integrate processing/computational capabilities into the CIS chip is because the advancements in the CMOS technologies that, for instance, have significantly shrunk the feature size, which is the smallest physical dimension that can be reliably fabricated on a semiconductor chip and is proportional to the transistor size. At the same time, however,

[3]It is worth noting, however, that it is difficult for the CCD sensor to perform CDS because of its read-out architecture (shifting charges to a single SF amplifier).

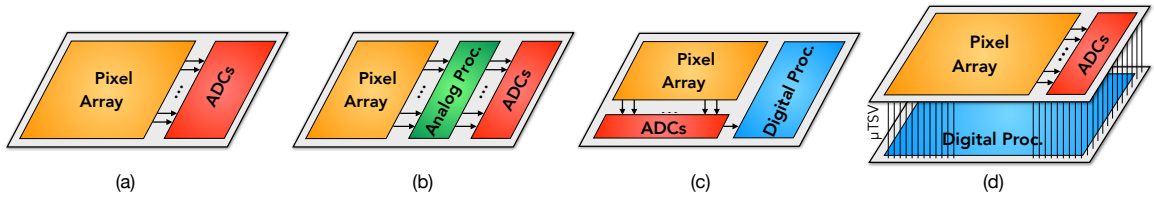


Figure 11: (a) Traditional 2D imaging CIS with the PD array and the ADCs. (b) Computational CIS with analog processing capabilities (before the ADCs). (c) Computational CIS with digital processing. (d) Stacked computational CIS with digital processing in a separate layer. Adapted from Ma et al. [2023, Fig. 2].

the PD size itself has not shrunk proportionally, meaning adding CMOS logics to the sensor increases the total chip area minimally in the grand scheme of things.

This is shown in Figure 10 (b), where triangle markers show the pixel sizes in CIS designs from all ISSCC papers appeared during Year 2000 and Year 2022, which include leading industry CIS designs at different times. We overlay a trend line regressed from these CIS designs to better illustrate the pixel size scaling trend. As a comparison, the blue line at the bottom represents the standard CMOS technology node scaling laid out by International Roadmap for Devices and Systems (IRDS) [IRDS, 2024]. We can see that the gap between the pixel size and the standard CMOS feature size steadily increases. In fact, the pixel size scaling stagnates at around 5 μm, which is long seen as the practical pixel size limit [Fossum, 1997]. As semiconductor manufacturers keep pulling rabbits out of a hat, the CMOS feature size is still, miraculously, shrinking (TSMC/Samsung are shipping products with a 2 nm process node in 2025), so the gap would still exist at least for quite a while.

Computational CIS Architectures

The computations inside a CIS could take place in both the analog and the digital domain. Figure 11 (b) illustrates one example where analog computing is integrated into a CIS chip before the ADC. Analog operations usually implement primitives for feature extraction [Bong et al., 2017b,a], object detection [Young et al., 2019], and DNN inference [Hsu et al., 2020; Xu et al., 2021]. Figure 11 (c) illustrates another example that integrates digital processing, such as ISP [Murakami et al., 2022], image filtering [Kim et al., 2005], and DNN [Bong et al., 2017a].

As the processing capabilities become more complex, CIS design has embraced 3D stacking technologies, as is evident by the increasing number of stacked CIS in Figure 10. Figure 11 (d) illustrates a typical stacked design, where the processing logic is separated from, and stacked with, the pixel array layer. The different layers communicate through hybrid bond or micro Through-Silicon Via (μTSV) [Liu et al., 2022; Tsugawa et al., 2017]. The processing layer typically integrates digital processors; such as ISP [Kwon et al., 2020], image processing [Hirata et al., 2021; Kumagai et al., 2018a], and DNN accelerators [Eki et al., 2021; Liu et al., 2022].

Three-layer stacked designs have also been proposed. Sony IMX 400 [Haruta et al., 2017] is

a 3-layer design that integrates a pixel layer, a DRAM layer (1 Gbit), and a logic layer with an Image Signal Processor (ISP). The DRAM layer buffers high-rate frames before steaming them out to the host. This enables super slow motion (960 FPS): otherwise, the bandwidth of the MIPI CSI-2 interface limits the capturing rate of the sensor. Meta conceptualizes a three-layer design [Liu et al., 2022] with a pixel array layer, a per-pixel ADC layer, and a digital processing layer that integrates a DNN accelerator — using DPS. Stacking makes it easier to implement DPS: the main disadvantage of DPS is the complexity of the pixel design, but with stacking the additional pixel processing circuitry (gain amplifier, ADC, etc.) can be "hidden" on a separate layer than the pixel array layer [Liu et al., 2022, 2020].

Challenges of CIS

Moving computation inside a CIS, however, is not without challenges. Most importantly, processing inside the sensor is far less efficient than that outside the sensor. This is because, while the CIS is implemented using the CMOS technologies, it uses significantly *older* process nodes than that of the conventional CMOS.

This is shown in Figure 10 (b), where the square markers show the process node used in each CIS paper surveyed. As a reference, the IRDS standard CMOS process node scaling line is also shown. At around Year 2000, the CIS process node started lagging behind that of the conventional CMOS node, and the gap is increasing. CIS design today commonly use $65nm$ and older process nodes. This gap is not an artifact of the CIS designs we pick; it is fundamental: there is simply no need to aggressively scale down the process node because the pixel size does not, and can not, shrink much. In fact, from Figure 10 (b) we can see that the slope of CIS process node scaling almost follows exactly that of the pixel size scaling. The reason that pixel size does not shrink much is to ensure light sensitivity: a small pixel reduces the number of photons it can collect, which directly reduces the dynamic range and the SNR, which we will see in the noise lecture[4].

Inefficient in-sensor processing can be mitigated through 3D stacking technologies, which allows for heterogeneous integration: the pixel layer and the computing layer(s) can use their respective, optimal process node. Stacking, however, could increase power density especially when future CIS integrate more processing capabilities. Therefore, harnessing the power of (stacked) computational CIS requires exploring a large design space, and is still an active area of research [Ma, 2024; Feng et al., 2024; Ma et al., 2023].

4 In-Sensor Optics

The on-chip optics server a few purposes: blocking lights in the IR/UV ranges, boosting photon collection efficiency, anti-aliasing, and filtering for color reproduction.

[4]It is interesting to note the fact that there is a fundamental pixel size limit negates one advantage of the CCD sensors, where the pixel design is simpler so one can theoretically make the pixel size smaller, but that is countered by the limit to which the PDs can shrink [Fossum, 1997].

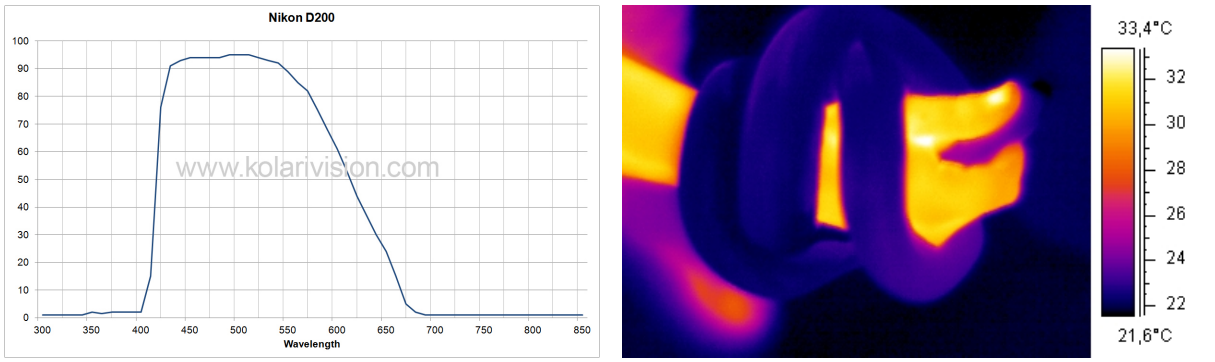


Figure 12: Left: transmittance spectrum of the on-chip cut-off optics on Nikon D200; from Kolarivision [Melentijevic, 2015]. Right: IR thermal imaging uses light power in the IR range to estimate temperature; from Arno / Coen [2006].

4.1 IR/UV Cut-Off Filters

Many cameras have cut-off filters for infrared (IR) and ultraviolet (UV) lights. Their goals are to remove/block IR or UV lights, as much as possible, from the incident light. These filters are transparent in that they predominately absorb lights while scattering very little light. So their optical behaviors can be adequately captured by their transmittance spectra. Figure 12 (left) shows the transmittance spectrum of the cut-off filter on Nikon D200, where light below 400 nm and above 700 nm are essentially blocked from hitting the sensor.

The reason most photographic cameras want to remove IR and UV lights is because the human visual system is not sensitive to IR and UV lights (recall our earlier discussions about the spectra of the cone fundamentals, which drop to 0 beyond roughly the 380 nm and 780 nm range). So for a camera to accurate reproduce the color of an object as if the object is directly viewed by the human eyes, the sensor's sensitivity ideally needs to mimic that of the human eyes. Cutting IR and UV lights for which our photoreceptors are not sensitive to is just the first step. We will discuss in detail in Chapter 5 what other mechanisms are in place for accurate color reproduction in image sensors.

Interestingly, thermographic cameras detect optical power in the IR range to estimate object temperature. Any object above absolute zero radiates, and this is call the **blackbody radiation**. Planck's law governs the electromagnetic power emitted at a particular wavelength at a particular temperature. It turns out that at room temperature (about 300 K), most of the radiation power is in the IR range; very little radiation comes from the visible range. That is why thermal cameras use IR radiation for temperature estimation. Figure 12 (right) shows an example of an IR image visualized as a heatmap, a real heatmap.

4.2 Microlenses

An important figure of merit of image sensors is the **fill factor** (FF), which is defined as the ratio of the photosensitive area of a pixel (A_{pd}), i.e., the photodiode, to the actual pixel area

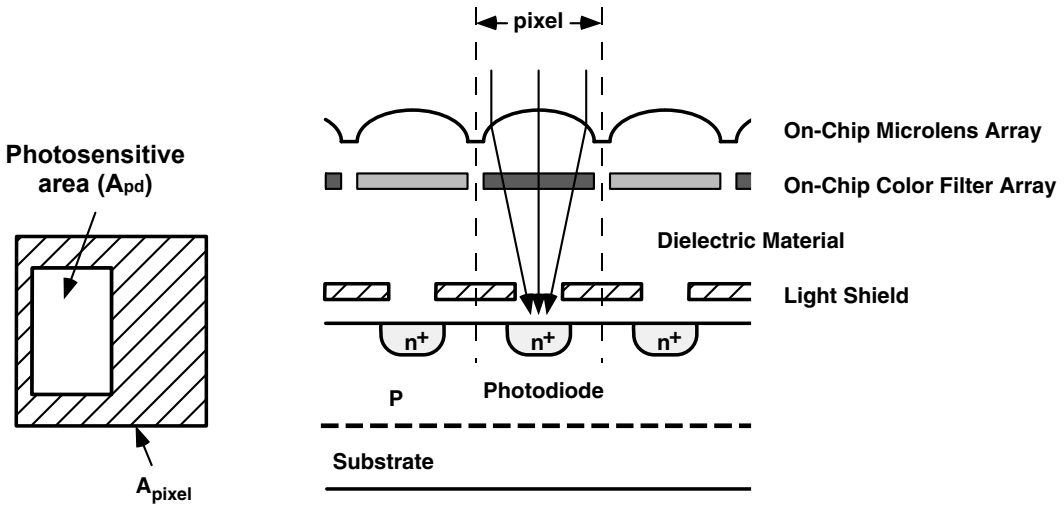


Figure 13: (a): the fill factor of an image sensor is the ratio of the photosensitive area (A_{pd} in the figure) to the total pixel area (A_{pixel}). (b): microlenses increase the effective fill factor of an image sensor. Adapted from Nakamura [2006, Fig. 3.9].

(A_{pixel}). This is illustrated in Figure 13 (a) and defined as:

$$FF = A_{ps}/A_{pixel} \times 100\%. \tag{7}$$

Why would A_{ps} be smaller than A_{pixel}? Recall from Figure 1 that, in addition to the actual photodiode, a pixel contains many other electrical components (capacitors, transistors, and other complex logic gates) that take up the area. Given a fixed pixel area, a larger FF means the pixel collects more photons during exposure, which, as we will discuss later, translates to a higher signal to noise ratio. So it is almost always desirable to have a higher FF.

One common way to increase the FF that is prevalent in almost all image sensors is through microlenses. This is illustrated in Figure 13 (b). Every pixel has a convex lens, which we call a **microlens**, sitting on top of it. The job of the microlens is to, ideally, direct all the photons hitting the pixel to the photodiode, in which case the FF would effectively be 100%, which contemporary image sensors are very close to.

4.3 Anti-Aliasing Filters

Many image sensors also have anti-aliasing (AA) filters, especially photographic sensors. Recall that pixels perform spatial sampling of the optical image, which is continuous, thus introducing aliasing. The classic anti-aliasing method is to pre-filter the continuous signal using a low-pass filter, essentially blurring the signal and reducing its peak frequency. Pharr et al. [2023,

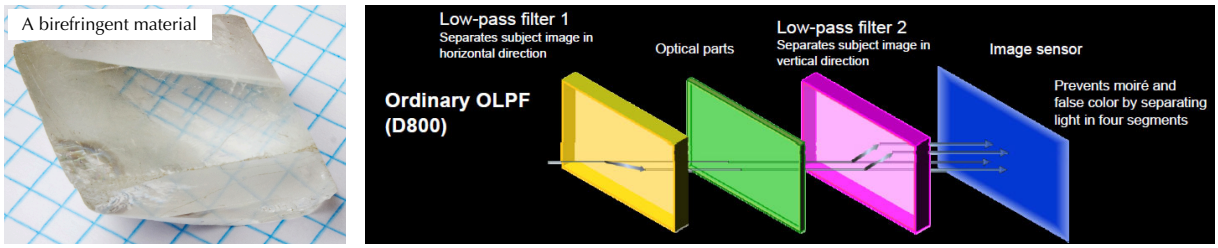


Figure 14: Left: a birefringent material that, through double refraction, split a ray into two; adapted from APN MJM [2011]. Right: Nikon D800 cascades two birefringent materials that splits a ray to four; from Diallo and Britton [2012].

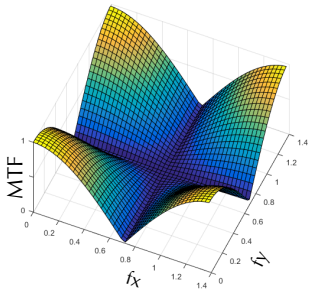
Chpt. 8] and Glassner [1995, Unit II] provide great technical discussions of signal sampling and reconstruction, which we will omit here.

In some sense, the photodiodes themselves and the microlenses act as pre-filters already: they inherently perform spatial 2D box convolutions over the continuous signal impinging upon them. Take the photodiode as an example: each photodiode integrates all the incident photons as we have seen in Chapter 2, and integration is equivalent to convolving/filtering the signal with a 2D box filter.

However, the support of the filter carried by the microlens and the photodiode is small: the microlens filter has a size of the pixel area and the photodiode filter support is even more compact. To more aggressively pre-filter the signal, we need a filter with a wide support. To that end, AA filters use birefringent material, as shown Figure 14 (left), which essentially splits a ray into two rays, each has a different polarization and, thus, takes a slightly different path (recall that the refractive index depends on the polarization of light). If we cascade two such materials, a ray gets split into four rays; this is called a 4-dot beam splitting. This is done by, e.g., Nikon D800e, as shown in Figure 14 (right).

The birefringent material acts as a low-pass filter. The intuition is that if an incident ray is spread over, say, 4 sensor-plane points, then each sensor-plan point, equivalently, integrates information from 4 incident rays, each coming a distinct scene points (assuming a pinhole aperture). We know integration is essentially low-pass filtering.

The way to understand the effect of the AA filter is to analyze its Point Spread Function (PSF) and Modulation Transfer Function (MTF). Assuming a pinhole aperture, a 4-dot beam splitting AA filter essentially imposes a PSF where a scene point is spread over 4 sensor-plane points. The PSF is the sum of 4 Dirac Delta functions placed on a regular grid (where the offset between adjacent grid points depends on the difference in refractive indices and the relative positions between the two splitting plans). An example of the MTF of such a PSF is shown in Figure 15 (left), where the x- and y-axis are the two spatial frequencies and the a-axis is the magnitude of the MTF. We can see that this particular MTF passes low frequencies, cuts off at around 0.7 cycles per pixel, and interestingly also pass high frequencies. Passing high frequencies is generally not a huge concern because power at high frequencies are usually already attenuated by the PSFs of other optical elements (e.g., the main imaging lens). Of course in reality the



MTF of a 4-dot AA

Without AA Filter (Nikon D800e)

With AA filter (Nikon D800)

Figure 15: Left: MTF of a 4-dot AA filter, where low and high frequency bands are allowed to pass; from AlmaPhoto [2016]. Middle: an image taken by Nikon D800e, which doesn't have an AA filter. Right: the same scene taken by Nikon D800, which has a 4-dot AA filter. The two Nikon images are from Cardinal [2012].

aperture is not a pinhole, so the PSF is not simply a sum of four Delta functions, but can similarly analyzed.

Figure 15 (middle) and Figure 15 (right) compare the images taken of the same scene by Nikon D800e, which lacks an AA filter, and Nikon D800, which has a 4-dot AA filter. Look at the AC's condenser coil; the AA image is more blurred but has much less objectionable aliasing effect.

4.4 Monochromatic (Noise-Free) Sensor Model

Each in-sensor optical element adds its own spectral transmittance, so the overall transmittance of the in-sensor optics is the product of them. We will simply use $T(\lambda)$ to represent the overall transmittance. Given what we have discussed so far, we can build an analytical model for a monochromatic, noise-free image sensor. The raw pixel value n of a pixel of size $u \times v$ whose top-left corner is (x, y) and is exposed for a duration of t_{exp} is given by:

$$Q = \int_{\lambda} \int_{t}^{t+t_{exp}} \int_{y}^{y+v} \int_{x}^{x+u} Y(x', y', \lambda, t') T(\lambda) QE(\lambda) \mathrm{d}x' \mathrm{d}y' \mathrm{d}t' \mathrm{d}\lambda, \tag{8a}$$

$$\Delta V = \frac{Qq}{C_{FD}} \times g, \tag{8b}$$

$$n = \lfloor \frac{\Delta V}{V_{max}} (2^N - 1) \rfloor, \tag{8c}$$

where $Y(x', y', \lambda, t')$ is the number of photons incident on position (x', y') at a particular wavelength λ at a particular time t', so it is a quantal counterpart of the spectral irradiance; $T(\lambda)$ is the overall spectral transmittance of the in-sensor optics, $QE(\lambda)$ is the quantum efficiency, q

is the elementary charge; Equation 8a models the total amount of charges collected at the particular pixel, where we integrate spatially, temporally, and spectrally. Equation 8b is essentially Equation 5, and models the voltage difference sensed before and after the exposure. Equation 8c is a crude ADC model, assuming that the voltage range $[0, v_{max}]$ is quantized into N bits, and the output of the ADC model is the digital number, a.k.a., the raw pixel value.

How do we express $Y(x', y', \lambda, t')$, the quantal counterpart of irradiance? The spectral irradiance at position (x', y') and time t' is:

$$E(x', y', \lambda, t') = \int^{\Omega(p,V)} L(p, \omega, \lambda, t') \cos\theta \, \mathrm{d}\omega, \tag{9}$$

where $p = (x', y')$, V is the aperture, $\Omega(p, V)$ is the solid angle subtended by p and V; $L(p, \omega, \lambda, t')$ is the radiance with a wavelength λ incident on p from the direction ω at time t', and θ is the polar angle subtended by ω and the pixel normal vector.

Given Planck's equation (Equation 2), we can turn irradiance E (energy per unit area per unit time) to the quantity Y (photon quantity per unit area per unit time):

$$Y(x', y', \lambda, t') = \frac{E(x', y', \lambda, t')\lambda}{hc}. \tag{10}$$

Plugging Equation 9 and Equation 10 into Equation 8, we have:

$$Q = \int_\lambda \int_t^{t+t_{exp}} \int_y^{y+v} \int_x^{x+u} \int^{\Omega(p,V)} \frac{L(p, \omega, \lambda, t') \cos\theta \mathrm{d}\omega T(\lambda) QE(\lambda) \lambda}{hc} \mathrm{d}x' \mathrm{d}y' \mathrm{d}t' \mathrm{d}\lambda \tag{11a}$$

$$= \int_\lambda \Big(\int_t^{t+t_{exp}} \int_y^{y+v} \int_x^{x+u} \int^{\Omega(p,V)} L(p, \omega, \lambda, t') \cos\theta \mathrm{d}\omega \mathrm{d}x' \mathrm{d}y' \mathrm{d}t' \Big) T(\lambda) QE(\lambda) \frac{\lambda}{hc} \mathrm{d}\lambda \tag{11b}$$

$$= \int_\lambda \mathcal{E}(\lambda) T(\lambda) QE(\lambda) \frac{\lambda}{hc} \mathrm{d}\lambda. \tag{11c}$$

Recall from the material lecture, the inner four integrals in Equation 11b collectively form the so-called camera measurement equation, which calculates $\mathcal{E}(\lambda)$ in Equation 11c, representing the energy at wavelength λ collected by the pixel during the exposure. We have implicitly assumed here that the effects of the in-sensor optics can simply be modeled by the spectral transmittance $T(\lambda)$. This is largely reasonable because 1) in-sensor optics are mostly transparent and 2) they are very close to the pixels so we can ignore rays that incident on the edge of the optics and, after refractions, miss the pixels.

Spectral Sensitivity Function

We can make a few assumptions to simplify our discussion. First, we assume the ADC quantization error is negligible. Second, we assume that the irradiance within a pixel is spatially uniform and temporally uniform during a short exposure time. Equation 8 is then simplified to:

$$n \approx k \int_\lambda Y(x, y, \lambda, t) T(\lambda) QE(\lambda) \mathrm{d}\lambda, \tag{12a}$$

$$= k \int_\lambda Y(x, y, \lambda, t) SSF_{quantal}(\lambda) \mathrm{d}\lambda, \tag{12b}$$

where $Y(x, y, \lambda, t)$ is the (average) number of incident photons at wavelength λ hitting position (x, y) at time t; $k = uvt_{exp} \frac{qg}{C_{FD}} \frac{2^N - 1}{V_{max}}$ is a constant.

Let's define a convenient term: **Spectral Sensitivity Function** (SSF), which is the product of $T(\lambda)$ and $QE(\lambda)$. SSF is the only spectral (wavelength-dependent) term in Equation 12b other than the incident light itself; it represents the phenomenological light sensitivity of the sensor over wavelength. SSF is sometimes also called the camera response function.

The SSF defined in Equation 12b is an "equal-quantal" function because it tells us the relative responses between different wavelengths under the same amount of incident photons. We can turn it to an "equal-energy" or "equal-power" function that operates on energy or power:

$$n \approx k \int_\lambda Y(x, y, \lambda, t) SSF_{quantal}(\lambda) \mathrm{d}\lambda, \tag{13a}$$

$$= k \int_\lambda \frac{\Phi(x, y, \lambda, t)}{t_{exp} \frac{hc}{\lambda}} SSF_{quantal}(\lambda) \mathrm{d}\lambda, \tag{13b}$$

$$= k' \int_\lambda \Phi(x, y, \lambda, t) SSF_{power}(\lambda) \mathrm{d}\lambda, \tag{13c}$$

where $\Phi(x, y, \lambda, t)$ denotes the spectral power distribution of the light hitting position (x, y) at time t, $k' = uv \frac{qg}{C_{FD}} \frac{2^N - 1}{V_{max}} \frac{1}{hc}$, and $SSF_{power}(\lambda) = SSF_{quantal}(\lambda)\lambda$ is the equal-power SSF. We will omit the subscript because it is usually clear what SSF is being used (e.g., from the quantity that is being multiplied with the SSF). Also note that in some literature, the SSF is used interchangeably with QE, so be very careful.

5 Color Sensing

There is one main piece of the on-chip optics we have not discussed: the color filters, which are critical for color sensing and deserve their own section.

5.1 Goal of Color Sensing

What does it mean for an image sensor to capture color? We know that colors are subjective sensations caused by cone photoreceptor responses of light; a color can be expressed as a point in a 3D space formed by the L, M and S cone responses, i.e., the LMS cone space. Ideally, if we can build an image sensor in such a way that it also possess three kinds of pixels, each of which has a spectral sensitivity matching exactly that of a cone class (i.e., cone fundamental), the sensor would be able to accurately capture and reproduce the color information.

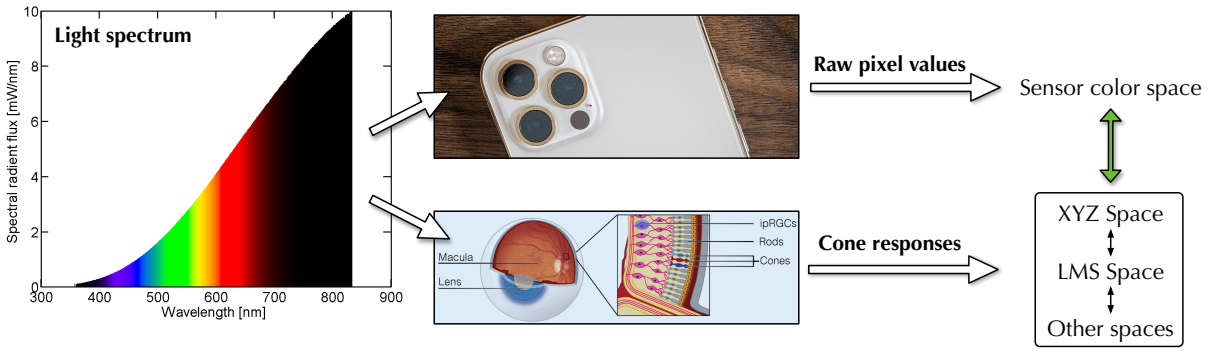


Figure 16: The goal of color sensing is to form a color space from the raw pixel values and for there to exist a (preferably linear) transformation between the sensor color space and a standard color space, typically the CIE XYZ space. Adapted from Blume and Garbazza and Spitschan [2019]; Thorseth [2015]; ajay_suresh [2021].

In fact, it is even sufficient for the sensor responses to be just a (linear) transformation away from the cone responses, as long as we can pre-calibrate the transformation matrix offline. This idea is illustrated in Figure 16. We emphasize linear transformation here simply because it is computationally cheaper; nothing prevents you from designing a sensor sensitivity profile that requires a sophisticated transformation from the cone space.

Where do the three classes of spectral sensitivities come? Examine our monochromatic sensing model in Equation 12b; it appears that all the pixels share the same response function and, thus, have the same spectral sensitivity: every pixel has the same quantum efficiency and the same optical elements sitting above them (so the same spectral transmittance of the optics).

There are a variety of ways to introduce sensitivity differences across pixels, which we will discuss shortly in Chapter 5.2. Assuming, for now, that we have somehow introduced the three classes of SSFs, denoted $SSF_R(\lambda)$, $SSF_G(\lambda)$, and $SSF_B(\lambda)$. Given an incident light with a SPD $\Phi(\lambda)$, the camera responses are:

$$[\int_\lambda \Phi(\lambda) SSF_R(\lambda) \mathrm{d}\lambda, \int_\lambda \Phi(\lambda) SSF_G(\lambda) \mathrm{d}\lambda, \int_\lambda \Phi(\lambda) SSF_B(\lambda) \mathrm{d}\lambda]. \tag{14}$$

This is a directly invocation of Equation 13c with the constant omitted. The color of the light expressed in the LMS cone space is:

$$[\int_\lambda \Phi(\lambda) L(\lambda) \mathrm{d}\lambda, \int_\lambda \Phi(\lambda) M(\lambda) \mathrm{d}\lambda, \int_\lambda \Phi(\lambda) S(\lambda) \mathrm{d}\lambda]. \tag{15}$$

If the cone responses form a 3D cone space, the camera raw responses also form a color space, which is sometimes called the camera's native color space. We provide an interactive tutorial that allows you to interactively explore and compare the native color spaces of various

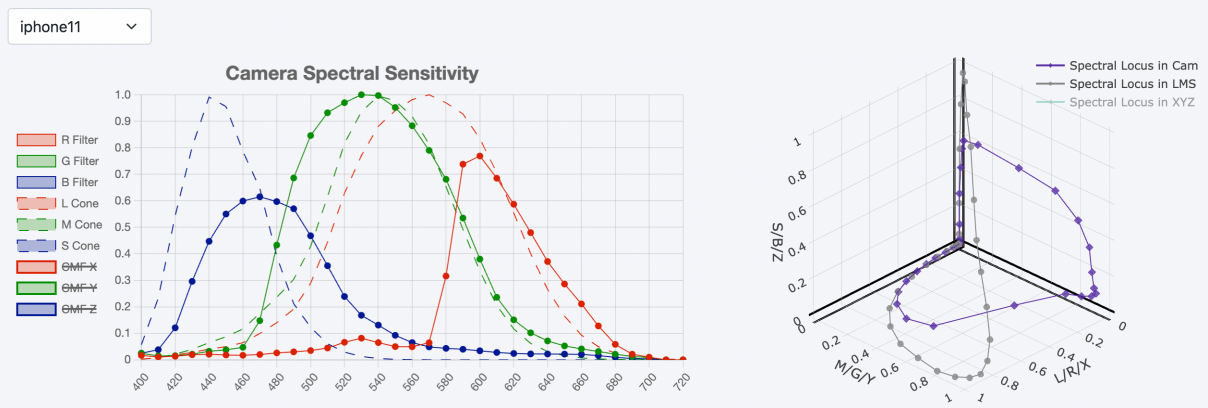


Figure 17: Left: Spectral sensitivity functions of iPhone 11 (the RGB filters; solid lines) in comparison with the LMS cone fundamentals (dashed lines). Right: the spectral locus in the LMS space and in the camera's native color space. Adapted from Zhu [2022].

cameras and the LMS cone space. Figure 17 (left) shows the SSFs of iPhone 11 (solid lines) and the cone fundamentals. The SSFs are normalized so that SSF_G is peaked at unity, and the cone fundamentals are each normalized to peak at unity, so you could compare the relatively sensitivity between the three SSFs in iPhone 11 but could not between the cone classes. Usually the SSF of a camera depends on a variety of factors such as the materials of the optical elements and the photodiodes as well as the pixel design, so it is almost impossible for the three SSFs to match exactly the cone fundamentals. Figure 17 (right) shows the spectral locus in iPhone 11's native color space and in the cone space; they evidently do not overlap.

A major task in sensor calibration is to identify a transformation matrix M such that the following (approximately) holds:

$$\begin{bmatrix} \int_\lambda \Phi(\lambda) SSF_R(\lambda) \mathrm{d}\lambda \\ \int_\lambda \Phi(\lambda) SSF_G(\lambda) \mathrm{d}\lambda \\ \int_\lambda \Phi(\lambda) SSF_B(\lambda) \mathrm{d}\lambda \end{bmatrix} \times M = \begin{bmatrix} \int_\lambda \Phi(\lambda) L(\lambda) \mathrm{d}\lambda \\ \int_\lambda \Phi(\lambda) M(\lambda) \mathrm{d}\lambda \\ \int_\lambda \Phi(\lambda) S(\lambda) \mathrm{d}\lambda \end{bmatrix} \tag{16}$$

The transformation matrix is then applied in the post-processing pipeline of the raw pixels to turn raw pixel responses to a color value. We will discuss the calibration and the post-processing pipeline in greater details later.

5.2 Implementing Three "Classes of Pixels"

Perhaps the most straightforward method to introduce varying SSF is to apply a spectral filter to different pixels. A spectral filter is just a transparent optical element with a wavelength-selective transmittance. We need only three filters to emulate the three cone classes, but ideally each pixel should get all three simultaneously, which is difficult if you think about it, since at any given time you can physically have only one filter sitting on a pixel.

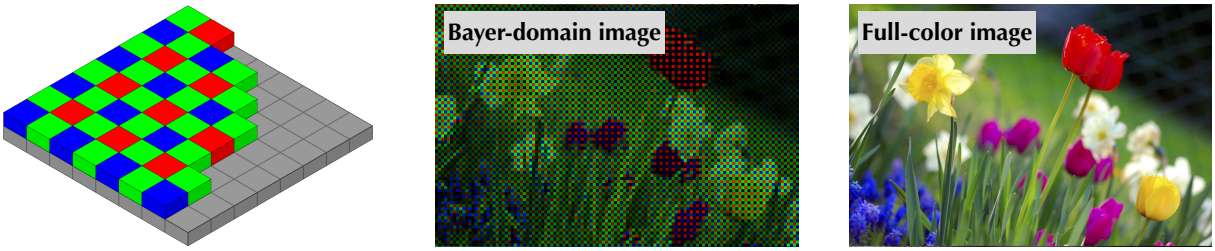


Figure 18: Left: the Bayer color filter array; from Cburnett [2006]. Middle and Right: a bayer-domain image where each pixel generates only one response and a full-color image assuming each pixel generates three responses; adapted from Cmglee [2018].

Three-Shot and Three-Chip Methods

There are two ways to go about addressing this issue. We can take three images of the same scene, each with a different filter, and then combine the together. This approach is believed to be pioneered by Sergey Prokudin-Gorsky, who conducted a breathtaking "photographic survey" of the early 20th-century Russia using this method [Prokudin-Gorsky, 1948]. This is called the "three-shot" approach. Alternatively, one could split the incident lights and send each of them to a different sensor, each with a different filter. This approach would obviously increase the form factor of the camera, but avoids having to register and align the three separate shots, which is subjective to object motion. These camera are called "three-chip" or "three-CCD/COMS" cameras, which are still very widely used today in broadcasting, film studios, etc.

Color Filter Array (CFA)

Both the three-shot and the three-chip approach allows each incident light to be transformed to three responses needed for color reproduction — at the cost of capturing overhead or bulky system design. A much simpler approach, and the most commonly used approach today, is called Color Filter Array (CFA), which assigns each pixel only *one* filter.

Figure 18 shows the most commonly used CFA, where the three classes of filters are tiled in what is called the Bayer filter mosaic, named after Bryce Bayer, who invented this pattern while working for Eastman Kodak in Rochester, NY [Bayer, 1976]. Each of the three filters has a transmittance spectrum that peaks at, roughly, red-ish, green-ish, and blue-ish wavelengths, similar to the spectra shown in Figure 17 (left).

The three filter classes are organized in 2×2 tiles, where each tiles has two green filters. Bayer did so because he wanted to mimic human vision, where the photopic Luminance Efficiency Function (LEF), as you would recall from the color vision chapter, is most sensitive to green-ish lights [Sharpe et al., 2005, 2011]. We can see that the CFA approach is actually more similar to human color vision that the three-shot or three-chip approach. In human vision, each cone photoreceptor has a particular sensitivity spectrum, and generates one of the three responses needed to form color vision.

A necessary consequence of using the CFA is that each pixel gets only one color channel

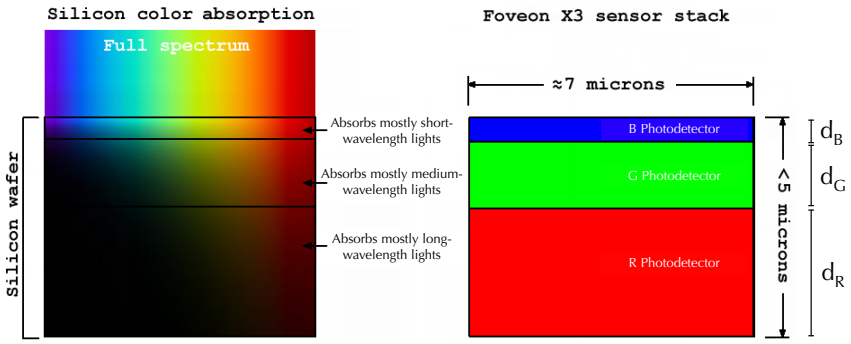


Figure 19: Illustration of the Foveon X3 pixel, which has three PDs made of the same material (silicon) vertically stacked; adapted from Anoneditor [2007]. Each PD receives a different light spectra (due to the depth-varying absorption), effectively creating three different responses of the same light incident on the pixel surface.

information. Figure 18 (middle) shows a raw image captured using a CFA, where each pixel evidently has only one color channel. The overall image looks overwhelmingly green because of the sheer amount of green filters. An important step in the post-processing pipeline is to reconstruct the two other missing channels, a process called "demosaicing", i.e., removing the bayer mosaic artifacts. An example of the reconstructed image is shown in Figure 18 (right).

We will have more to say about the demosaicing process when we get to the post-processing chapter, but for now, let's just observe that demosaicing is nothing more than a signal sampling and reconstruction problem. The CFA allows each pixel to sample only one channel of the three channels of response. So the green-filter response, for instance, is sampled by half of the pixels[5], and the other two responses are sampled by one quarter of the pixels each. The job of demosaicing is then to reconstruct the full signal responses from the samples — a well-established problem in signal processing.

Foveon Approach

The final approach does away with optical color filters altogether. Instead, we will use three photodiodes vertically stacked for each pixel. Figure 19 Foveon X3, perhaps the most famous sensor that uses this architecture.

The idea is that silicon absorption spectrum is wavelength sensitive, shown in the right panel of Figure 3. Blue-ish lights have a much shorter mean free length than green-ish lights, which have shorter mean free length that red-ish lights. This means most short-wavelength lights will be absorbed after the first photodiode, leaving mostly medium to long-wavelength lights. Those lights will go through the second photodiode, which absorbs mostly the medium-wavelength lights, leaving mostly long-wavelength lights to the third photodiode. As a result, each PD

[5]If we want to be pedantic, each green pixel has a small, but non-infinitesimal, area, so it first performs a low-pass filtering using a box filter whose extent is the pixel area, followed by sampling at the center of the pixel.

actually receives a different light spectrum, effectively creating three different responses for the same light incident on the pixel.

Let's assume that the three PDs have a depth of d_B, d_G, and d_R, respectively. The incident light impinging on the pixel (i.e., the first PD surface) has a SPD $\Phi(\lambda)$. The light impinging on the second PD then has a spectrum $\Phi(\lambda)e^{-\sigma(\lambda)d_B}$, where $\sigma(\lambda)$ is the silicon's absorption coefficient spectrum. This is easily derived from, if you recall from the material chapter, the fact that pure absorption (no scattering and emission) leads to an exponential decay of the input signal. Similarly, the light impinging on the third PD then has a spectrum $\Phi(\lambda)e^{-\sigma(\lambda)(d_B+d_G)}$. The responses produced by the three PDs are thus (in the order of R, G, and G):

$$[\int_\lambda \Phi(\lambda)\eta_R(\lambda)e^{-\sigma(\lambda)(d_B+d_G)}, \int_\lambda \Phi(\lambda)\eta_G(\lambda)e^{-\sigma(\lambda)(d_B)}, \int_\lambda \Phi(\lambda)\eta_B(\lambda)], \tag{17}$$

where $\eta_R(\lambda)$, $\eta_G(\lambda)$, and $\eta_B(\lambda)$ are QE spectra of the three PDs (where we consider only photons that reach a PD as the denominator in Equation 3 while ignoring photons that are reflected/absorbed before the photons hit the PD), respectively, and $\Phi(\lambda)$ is the SPD of the light incident on the pixel surface. The three PDs use the exact same material (so they share the same silicon absorption spectrum), but can still have different $\eta(\lambda)$s because of the thickness differences, which is a result of the differences in the lengths of the depletion and neutral regions in the PD p-n junctions. Can you guess why the thickness tends to increase for deeper PDs in Figure 19 (right)?

Compared to using the CFA, the vertical PD stacking approach is much more complicated to fabricate and more costly, so is much less commonly used. It avoids color sampling (and the resulting aliasing) and the need for demosaicing, and in theory could also have a higher overall quantum efficiency (and signal to noise ratio) since there is no color filters, so it might find uses in scientific imaging [Chen et al., 2023].

6 Image Signal Processing

The output of an image sensor is what we usually call the raw pixels. The raw pixels are not the usual RGB images we are used to see. For starters, if we use the common CFA approach for color sensing, each raw pixel has only one color channel response — the two missing channel responses must be recovered. We have also ignored noises, which are introduced every step along the signal transduction chain from incident lights to raw values. Therefore, the raw pixels usually go through a post-processing pipeline to yield a visually pleasing image.

That pipeline in modern cameras is implemented by a special hardware accelerator called the Image Signal Processor (ISP), which is an Intellectual Property (IP) block in a mobile System-on-a-Chip (SoC). Implementing the post-processing algorithms in a dedicated hardware makes a lot of sense from an efficiency perspective: when you press a button to capture an image you certainly do not want to wait for a long time or burn a lot of energy before the image is shown to you. As many mobile vendors do not actually control the optics and the sensor, the ISP

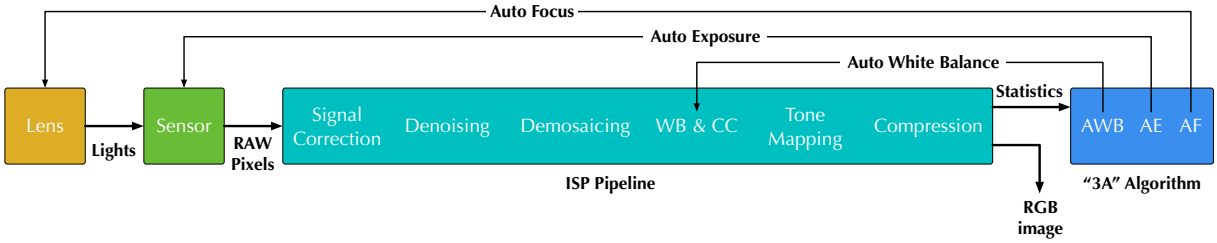


Figure 20: A general ISP pipeline. The exact stages and arrangement of the stages are proprietary. The ISP pipeline outputs both the RGB (or other color spaces) images and image statistics; the latter is used to drive the so-called "3A" algorithms, which are feedback controls over the lens, sensor, and the ISP.

increasingly has become the key product differentiator. As a result, many companies have their custom ISP designs; for instance, Qualcomm's Snapdragon SoC has their own Spectra ISP.

Many texts exist on the general ISP algorithm [Ramanath et al., 2005; Karaimer and Brown, 2016] and the hardware design [Hegarty et al., 2014], which we refer you to. The goal of this section is to walk through the general pipeline and point out main ideas. One thing worth emphasizing here is that the ISP design is strongly influenced by the downstream task that consumes the output of the ISP. The two main consumers are human vision and machine vision. The former cares about visually pleasing images while the latter does not — as long as the key semantics information is retained and can be extracted.

6.1 General Pipeline

Figure 20 shows a general ISP pipeline and how it fits into the entire imaging pipeline. The ISP takes the raw pixels generated by the sensor and generates two types of output: the finished image, usually encoded in sRGB color space and compressed, and statistics of the image that are used to drive the so-called "3A" algorithms, i.e., auto white balance (AWB), auto exposure (AE), and auto focus (AF). We will not have much time to discuss the 3A algorithms, but they can be thought of as feedback controls over the rest of the imaging system: AWB controls the white balancing stage in the ISP, AE control the exposure time of the image sensor, and AF controls the lens movement in the optics. The 3A algorithms usually run on the host CPU or an MCU, because they are relatively simple computationally.

The ISP pipeline shown here is a general architecture that covers roughly what an ISP has to do. Keep in mind that the exact stages and their arrangements are proprietary and vary by vendors. Regardless, all ISPs operate on a set of basic principles.

- Recall that the raw pixel values output by the sensor should ideally be proportional to the scene luminance, but this is hardly the case in reality. The first thing an ISP does is to recover luminance-proportional values from the raw sensor output; this includes correcting some known (calibrated) issues in sensing and denoising. It is important that these steps

are taken at the very beginning of an ISP; if the pixel values are noisy, any subsequent manipulations on the pixels also manipulate, sometimes amplify, the noise.

- After that, we can assume that the raw pixels carry physical meanings: they are proportional to luminance, but of course because of the CFA, the color information is spatially sampled. The raw pixels before demosaicing are usually called pixels in the Bayer domain. The next stage is demosaicing, which essentially reconstructs the color information (all three channels) from the single-channel samples.

- The demosaiced color information is encoded in the sensor's native color space, because the sensor's SSFs almost certainly do not match the cone fundamentals. So the next stage is to transform color from the sensor's native color space to a typical color space such as the CIE XYZ space. White balancing usually is implemented along with color correction, because both involving manipulating color information [Rowlands, 2020].

- Usually there is a tone mapping stage in the ISP. The dynamic range of the raw pixels is usually, not always, different from that of a typical output medium (e.g., a display or a print). Tone mapping operators map signals between the two dynamic ranges so that the output image is visually appealing.

- The final output is usually compressed, either through an image compression algorithm (e.g., JPEG) or, in the case of video capturing, a video compression algorithms (e.g., H264).

6.2 Two Trends

Two trends emerge, which we will explore when appropriate. First, it has become increasingly common to co-design ISP algorithms, along with optics and image sensor design, with the downstream tasks. This is particularly important for machine vision, whose is not concerned with the traditional goal of an ISP, i.e., generating visually pleasing images. A co-design between the ISP and the machine vision algorithms could potentially improve both task quality and efficiency. Second, a huge amount of recent efforts have been spent on exploring the notion of "neural ISP", which is nothing more than replacing part, or the entirety, of the ISP pipeline with deep neural networks (DNNs). The learning paradigm has two main advantages: it replaces some of the heuristics in traditional ISP designs, and it allows the algorithm to be more easily updated without having to wait until the next generation of the product. The latter point is possible because a neural ISP pipeline can run on a DNN accelerator that almost all modern mobile SoCs have, and updating the algorithm is nothing more than updating the model weights.

The key issue with neural ISP is speed and efficiency: a neural ISP model executed on a generic DNN accelerator is likely much slower and more energy hungry than traditional ISPs. So it is more likely that neural ISPs will found their main uses in offline image processing and photo finishing rather than in the real-time imaging pipeline.

References

Adair and Nikitas. The Components of a Digital Camera's Image Sensor. https://www.dummies.com/article/home-auto-hobbies/photography/cameras/general-cameras/the-components-of-a-digital-cameras-image-sensor-202248/, 2017.

ajay_suresh. iPhone 12 cameras; CC BY-SA 2.0 license. https://commons.wikimedia.org/wiki/File:Apple_iPhone_12_Pro_-_Cameras_(50535314721).jpg, 2021.

AlmaPhoto. A Simple Model for Sharpness in Digital Cameras – AA. https://www.strollswithmydog.com/resolution-model-digital-cameras-aa/, 2016.

Anoneditor. Illustration of the Foveon X3 sensor; CC BY-SA 3.0. https://commons.wikimedia.org/wiki/File:Absorption-X3.png, 2007.

Masakazu Aoki, Haruhisa Ando, Shinya Ohba, Iwao Takemoto, Shusaku Nagahara, Toshio Nakano, Masaharu Kubo, and Tsutomu Fujita. 2/3-inch format mos single-chip color imager. *IEEE Transactions on Electron Devices*, 29(4):745–750, 1982.

APN MJM. A calcite crystal displays the double refractive properties while sitting on a sheet of graph paper; CC BY-SA 3.0 license. https://commons.wikimedia.org/wiki/File:Crystal_on_graph_paper.jpg, 2011.

Arno / Coen. Thermogram of a snake wrapped around a human arm; CC BY-SA 3.0 license. https://commons.wikimedia.org/wiki/File:Wiki_stranglesnake.jpg, 2006.

Bryce E Bayer. Color imaging array, July 20 1976. US Patent 3,971,065.

John A Biretta and Matt McMaster. *Wide Field and Planetary Camera 2 Instrument Handbook v. 10.0*. Space Telescope Science Institute, 2008.

Blume and Garbazza and Spitschan. Schematic overview of photorecetors; CC BY-SA 4.0 license. https://commons.wikimedia.org/wiki/File:Overview_of_the_retina_photoreceptors_(a).png, 2019.

Kyeongryeol Bong, Sungpill Choi, Changhyeon Kim, Donghyeon Han, and Hoi-Jun Yoo. A low-power convolutional neural network face recognition processor and a cis integrated with always-on face detector. *IEEE Journal of Solid-State Circuits*, 53(1):115–123, 2017a.

Kyeongryeol Bong, Sungpill Choi, Changhyeon Kim, Sanghoon Kang, Youchang Kim, and Hoi-Jun Yoo. 14.6 a 0.62 mw ultra-low-power convolutional-neural-network face-recognition processor and a cis integrated with always-on haar-like face detector. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 248–249. IEEE, 2017b.

Willard S Boyle and George E Smith. Charge coupled semiconductor devices. *Bell System Technical Journal*, 49(4):587–593, 1970.

BrayLockBoy. An example of the Rolling shutter effect in action at Afton Down, Isle of Wight, taken by a camera on a car travelling at approximately 50 miles per hour. CC BY-SA 4.0 license. https://commons.wikimedia.org/wiki/File:Rolling_Shutter_Effect_at_Afton_Down,_21_August_2018.jpg, 2018.

Cardinal. New 36MP Nikon D800e: Is it too sharp for you? https://www.extremetech.com/extreme/117627-new-36mp-nikon-d800e-is-it-too-sharp-for-you, 2012.

Cburnett. A Bayer pattern on a sensor; CC BY-SA 3.0. https://commons.wikimedia.org/wiki/File:Bayer_pattern_on_sensor.svg, 2006.

Cheng Chen, Ziwen Wang, Jiajing Wu, Zhengtao Deng, Tao Zhang, Zhongmin Zhu, Yifei Jin, Benjamin Lew, Indrajit Srivastava, Zuodong Liang, et al. Bioinspired, vertically stacked, and perovskite nanocrystal–enhanced cmos imaging sensors for resolving uv spectral signatures. *Science Advances*, 9(44):eadk3860, 2023.

Cmglee. Images of a garden with some tulips and narcissus; CC BY-SA 3.0. https://commons.wikimedia.org/wiki/File:Colorful_spring_garden_Bayer_%2B_RGB.png, 2018.

Cmglee. Comparison of front- vs. back-illuminated sensors; CC BY-SA 4.0 license. https://commons.wikimedia.org/wiki/File:Comparison_backside_illumination.svg, 2019.

Diallo and Britton. The same but different: Nikon D800E. https://www.dpreview.com/reviews/nikon-d800-d800e/3, 2012.

Rudolph H Dyck and Gene P Weckler. Integrated arrays of silicon photodetectors for image sensing. *IEEE Transactions on Electron Devices*, 15(4):196–201, 1968.

Albert Einstein. Über einen die erzeugung und verwandlung des lichtes betreffenden heuristischen gesichtspunkt, 1905a.

Albert Einstein. On a heuristic point of view about the creation and conversion of light. *Annalen der Physik*, 17(6):132–148, 1905b.

Ryoji Eki, Satoshi Yamada, Hiroyuki Ozawa, Hitoshi Kai, Kazuyuki Okuike, Hareesh Gowtham, Hidetomo Nakanishi, Edan Almog, Yoel Livne, Gadi Yuval, et al. 9.6 a 1/2.3 inch 12.3 mpixel with on-chip 4.97 tops/w cnn processor back-illuminated stacked cmos image sensor. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 64, pages 154–156. IEEE, 2021.

Abbas El Gamal and Helmy Eltoukhy. Cmos image sensors. *IEEE Circuits and Devices Magazine*, 21(3):6–20, 2005.

EMVA. EMVA Standard 1288 Standard for Characterization of Image Sensors and Cameras. https://www.emva.org/wp-content/uploads/EMVA1288General_4.0Release.pdf, 2021.

Eric Bajart. Quantum efficiency of the CCD sensor "PC1" in the Hubble Space Telescope's Wide Field and Planetary Camera WFPC2; CC BY-SA 3.0. `https://commons.wikimedia.org/wiki/File:Quantum_efficiency_graph_for_WFPC2-en.svg`, 2010.

Yu Feng, Tianrui Ma, Yuhao Zhu, and Xuan Zhang. Blisscam: Boosting eye tracking efficiency with learned in-sensor sparse sampling. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pages 1262–1277. IEEE, 2024.

Eric R Fossum. Active pixel sensors: Are ccds dinosaurs? In *Charge-Coupled Devices and Solid State Optical Sensors III*, volume 1900, pages 2–14. SPIE, 1993.

Eric R Fossum. Cmos image sensors: Electronic camera-on-a-chip. *IEEE transactions on electron devices*, 44(10):1689–1698, 1997.

Eric R Fossum and Donald B Hondongwa. A review of the pinned photodiode for ccd and cmos image sensors. *IEEE Journal of the electron devices society*, 2014.

Boyd Fowler, Abbas El Gamal, and David XD Yang. A cmos area image sensor with pixel-level a/d conversion. In *Proceedings of IEEE International Solid-State Circuits Conference-ISSCC'94*, pages 226–227. IEEE, 1994.

Andrew S Glassner. *Principles of digital image synthesis*. Elsevier, 1995.

Martin A Green and Mark J Keevers. Optical properties of intrinsic silicon at 300 k. *Progress in Photovoltaics: Research and applications*, 3(3):189–192, 1995.

Tsutomu Haruta, Tsutomu Nakajima, Jun Hashizume, Taku Umebayashi, Hiroshi Takahashi, Kazuo Taniguchi, Masami Kuroda, Hiroshi Sumihiro, Koji Enoki, Takatsugu Yamasaki, et al. 4.6 a 1/2.3 inch 20mpixel 3-layer stacked cmos image sensor with dram. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 76–77. IEEE, 2017.

James Hegarty, John S Brunhaver, Zachary DeVito, Jonathan Ragan-Kelley, Noy Cohen, Steven Bell, Artem Vasilyev, Mark Horowitz, and Pat Hanrahan. Darkroom: compiling high-level image processing code into hardware pipelines. *ACM Trans. Graph.*, 33(4):144–1, 2014.

Tomoki Hirata, Hironobu Murata, Hideaki Matsuda, Yojiro Tezuka, and Shiro Tsunai. 7.8 a 1-inch 17mpixel 1000fps block-controlled coded-exposure back-illuminated stacked cmos image sensor for computational imaging and adaptive dynamic range control. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 64, pages 120–122. IEEE, 2021.

Tzu-Hsiang Hsu, Yi-Ren Chen, Ren-Shuo Liu, Chung-Chuan Lo, Kea-Tiong Tang, Meng-Fan Chang, and Chih-Cheng Hsieh. A 0.5-v real-time computational cmos image sensor with programmable kernel for feature extraction. *IEEE Journal of Solid-State Circuits*, 56(5): 1588–1596, 2020.

Chenming Hu. *Modern semiconductor devices for integrated circuits*. Prentice Hall, 2009.

IRDS. International roadmap for devices and systems. `https://irds.ieee.org/`, 2024.

Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 429–444. Springer, 2016.

Seong-Jin Kim, Kwang-Hyun Lee, Sang-Wook Han, and Euisik Yoon. A 200/spl times/160 pixel cmos fingerprint recognition soc with adaptable column-parallel processors. In *ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005.*, pages 250–596. IEEE, 2005.

Masahiro Kobayashi, Yusuke Onuki, Kazunari Kawabata, Hiroshi Sekine, Toshiki Tsuboi, Takashi Muto, Takeshi Akiyama, Yasushi Matsuno, Hidekazu Takahashi, Toru Koizumi, et al. 4.5a 1.8e-rms temporal noise over 110 db dynamic range 3.4μm pixel pitch global-shutter cmos image sensor with dual-gain amplifiers ss-adc, light guide structure, and multiple-accumulation shutter. *IEEE Journal of Solid-State Circuits*, 53(1):219–228, 2017.

Toru Kondo, Yoshiaki Takemoto, Kenji Kobayashi, Mitsuhiro Tsukimura, Naohiro Takazawa, Hideki Kato, Shunsuke Suzuki, Jun Aoki, Haruhisa Saito, Yuichi Gomi, et al. A 3d stacked cmos image sensor with 16mpixel global-shutter mode and 2mpixel 10000fps mode using 4 million interconnections. In *2015 Symposium on VLSI Circuits (VLSI Circuits)*, pages C90–C91. IEEE, 2015.

Lester J Kozlowski, J Luo, WE Kleinhans, and T Liu. Comparison of passive and active pixel schemes for cmos visible imagers. In *Infrared Readout Electronics IV*, volume 3360, pages 101–110. SPIE, 1998.

Oichi Kumagai, Atsumi Niwa, Katsuhiko Hanzawa, Hidetaka Kato, Shinichiro Futami, Toshio Ohyama, Tsutomu Imoto, Masahiko Nakamizo, Hirotaka Murakami, Tatsuki Nishino, et al. A 1/4-inch 3.9 mpixel low-power event-driven back-illuminated stacked cmos image sensor. In *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*, pages 86–88. IEEE, 2018a.

Y Kumagai, R Yoshita, N Osawa, H Ikeda, K Yamashita, T Abe, S Kudo, J Yamane, T Idekoba, S Noudo, et al. Back-illuminated 2.74μm-pixel-pitch global shutter cmos image sensor with charge-domain memory achieving 10k e-saturation signal. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 10–6. IEEE, 2018b.

Minho Kwon, Seunghyun Lim, Hyeokjong Lee, Il-Seon Ha, Moo-Young Kim, Il-Jin Seo, Suho Lee, Yongsuk Choi, Kyunghoon Kim, Hansoo Lee, et al. A low-power 65/14nm stacked cmos image sensor. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4. IEEE, 2020.

Chiao Liu, Andrew Berkovich, Song Chen, Hans Reyserhove, Syed Shakib Sarwar, and Tsung-Hsun Tsai. Intelligent vision systems–bringing human-machine interface to ar/vr. In *2019 IEEE International Electron Devices Meeting (IEDM)*, pages 10–5. IEEE, 2019.

Chiao Liu, Lyle Bainbridge, Andrew Berkovich, Song Chen, Wei Gao, Tsung-Hsun Tsai, Kazuya Mori, Rimon Ikeno, Masayuki Uno, Toshiyuki Isozaki, et al. A 4.6 μm, 512$\times$ 512, ultra-low power stacked digital pixel sensor with triple quantization and 127db dynamic range. In *2020 IEEE International Electron Devices Meeting (IEDM)*, pages 16–1. IEEE, 2020.

Chiao Liu, Song Chen, Tsung-Hsun Tsai, Barbara De Salvo, and Jorge Gomez. Augmented reality-the next frontier of image sensors and compute systems. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 65, pages 426–428. IEEE, 2022.

Tianrui Ma. *Efficient Data-Driven Machine Vision: A Co-Design of Circuit, Algorithm, and Architecture for Edge Vision Sensors*. PhD thesis, Washington University in St. Louis, 2024.

Tianrui Ma, Yu Feng, Xuan Zhang, and Yuhao Zhu. Camj: Enabling system-level energy modeling and architectural exploration for in-sensor visual computing. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–14, 2023.

Melentijevic. DSLR Internal Cut Filter / Lowpass Filter / Hot Mirror Transmission Curves. `https://kolarivision.com/articles/internal-cut-filter-transmission/`, 2015.

Ken Miyauchi, Kazuya Mori, Toshinori Otaka, Toshiyuki Isozaki, Naoto Yasuda, Alex Tsai, Yusuke Sawai, Hideki Owada, Isao Takayanagi, and Junichi Nakamura. A stacked back side-illuminated voltage domain global shutter cmos image sensor with a 4.0 μm multiple gain readout pixel. *Sensors*, 20(2):486, 2020.

Hirotaka Murakami, Eric Bohannon, John Childs, Grace Gui, Eric Moule, Katsuhiko Hanzawa, Tomofumi Koda, Chiaki Takano, Toshimasa Shimizu, Yuki Takizawa, et al. A 4.9 mpixel programmable-resolution multi-purpose cmos image sensor for computer vision. In *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 65, pages 104–106. IEEE, 2022.

Boris Murmann. ADC Performance Survey 1997-2024. `https://github.com/bmurmann/ADC-survey`, 2014.

Junichi Nakamura. *Image sensors and signal processing for digital still cameras*. CRC press, 2006.

Yoshikazu Nitta, Yoshinori Muramatsu, Kiyotaka Amano, Takayuki Toyama, K Mishina, Atsushi Suzuki, Tadayuki Taura, Akihiko Kato, Masaru Kikuchi, Yukihiro Yasui, et al. High-speed digital double sampling with analog cds on column parallel adc architecture for low-noise active pixel sensor. In *2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers*, pages 2024–2031. IEEE, 2006.

Peter JW Noble. Self-scanned silicon image detector arrays. *IEEE Transactions on electron Devices*, 15(4):202–209, 1968.

Jun Ohta. *Smart CMOS image sensors and applications*. CRC press, 2020.

Ommnomnomgulp. A focal plane shutter firing at 1/500 of a second with the "gap" clearly visible. This shutter is on a Nikon film SLR. CC BY-SA 3.0 license. `https://commons.wikimedia.org/wiki/File:1_500_Sec_Focal_P_Shut.jpg`, 2008.

Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. MIT Press, 4 edition, 2023.

Sergey Prokudin-Gorsky. Library of Congress Prokudin-Gorskii Collection. `https://www.loc.gov/collections/prokudin-gorskii/about-this-collection/`, 1948.

Rajeev Ramanath, Wesley E Snyder, Youngjun Yoo, and Mark S Drew. Color image processing pipeline. *IEEE Signal processing magazine*, 22(1):34–43, 2005.

D Andrew Rowlands. Color conversion matrices in digital cameras: a tutorial. *Optical Engineering*, 59(11):110801–110801, 2020.

Masaki Sakakibara, Yusuke Oike, Takafumi Takatsuka, Akihiko Kato, Katsumi Honda, Tadayuki Taura, Takashi Machida, Jun Okuno, Atsuhiro Ando, Taketo Fukuro, et al. An 83db-dynamic-range single-exposure global-shutter cmos image sensor with in-pixel dual storage. In *2012 IEEE International Solid-State Circuits Conference*, pages 380–382. IEEE, 2012.

Lindsay T Sharpe, Andrew Stockman, Wolfgang Jagla, and Herbert Jägle. A luminous efficiency function, v*(λ), for daylight adaptation. *Journal of vision*, 5(11):3–3, 2005.

Lindsay T Sharpe, Andrew Stockman, Wolfgang Jagla, and Herbert Jägle. A luminous efficiency function, vd65*(λ), for daylight adaptation: a correction. *Color Research & Application*, 36 (1):42–46, 2011.

Laurence Stark, Jeffrey M Raynor, Frederic Lalanne, and Robert K Henderson. A back-illuminated voltage-domain global shutter pixel with dual in-pixel storage. *IEEE Transactions on Electron Devices*, 65(10):4394–4400, 2018.

PK Swain and David Cheskis. Back-illuminated image sensors come to the forefront. *Photonics Spectra*, 42(8):46, 2008.

Thorseth. Spectral power distribution of a 25 W incandescent light bulb; CC BY-SA 4.0 license. `https://commons.wikimedia.org/wiki/File:Spectral_power_distribution_of_a_25_W_incandescent_light_bulb.png`, 2015.

Arnaud Tournier, F Roy, Y Cazaux, F Lalanne, P Malinge, M Mcdonald, G Monnot, and N Roux. A hdr 98db 3.2μm charge domain global shutter cmos image sensor. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 10–4. IEEE, 2018.

H Tsugawa, H Takahashi, R Nakamura, T Umebayashi, T Ogita, H Okano, K Iwase, H Kawashima, T Yamasaki, D Yoneyama, et al. Pixel/dram/logic 3-layer stacked cmos image sensor technology. In *2017 IEEE International Electron Devices Meeting (IEDM)*, pages 3–2. IEEE, 2017.

Gene P Weckler. Operation of pn junction photodetectors in a photon flux integrating mode. *IEEE Journal of solid-state circuits*, 2(3):65–73, 1967.

Han Xu, Ningchao Lin, Li Luo, Qi Wei, Runsheng Wang, Cheng Zhuo, Xunzhao Yin, Fei Qiao, and Huazhong Yang. Senputing: An ultra-low-power always-on vision perception chip featuring the deep fusion of sensing and computing. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 69(1):232–243, 2021.

Keita Yasutomi, Shinya Itoh, and Shoji Kawahito. A two-stage charge transfer active pixel cmos image sensor with low-noise global shuttering and a dual-shuttering mode. *IEEE transactions on electron devices*, 58(3):740–747, 2011.

Toshifumi Yokoyama, Masafumi Tsutsui, Yoshiaki Nishi, Ikuo Mizuno, Veinger Dmitry, and Assaf Lahav. High performance 2.5μm global shutter pixel with new designed light-pipe structure. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 10–5. IEEE, 2018.

Christopher Young, Alex Omid-Zohoor, Pedram Lajevardi, and Boris Murmann. A data-compressive 1.5/2.75-bit log-gradient qvga image sensor with multi-scale readout for always-on object detection. *IEEE Journal of Solid-State Circuits*, 54(11):2932–2946, 2019.

Yuhao Zhu. Exploring Camera Color Space and Color Correction. `https://horizon-lab.org/colorvis/camcolor.html`, 2022.