Understanding the in-camera rendering pipeline & the role of AI and deep learning

Michael S. Brown

Professor and Canada Research Chair

York University – Toronto

Senior Research Director

Samsung AI Center – Toronto

Motivation for this tutorial

Scientist's view of a photograph

Photo by Uwe Hermann

Scientist's view of a photograph

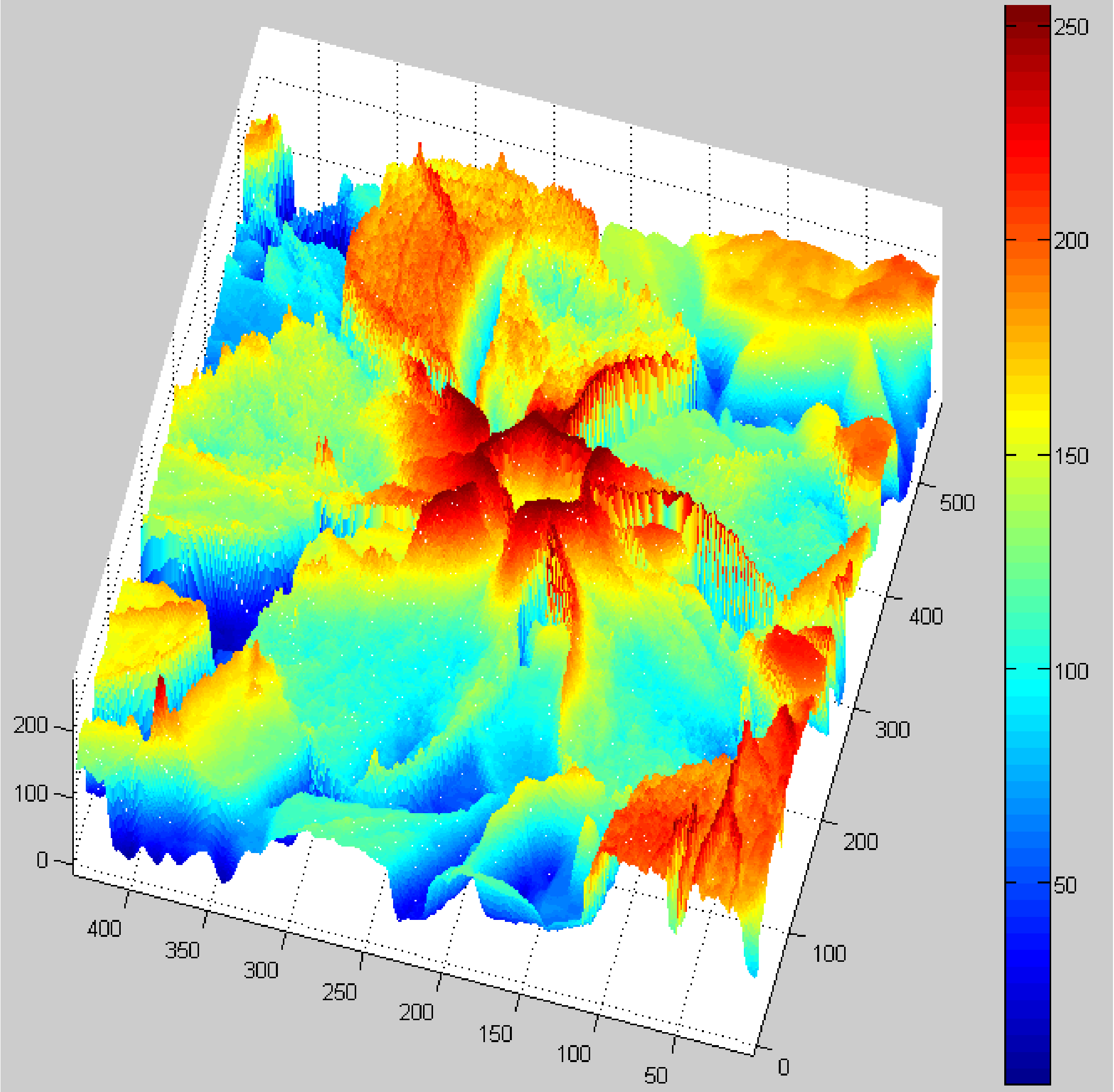
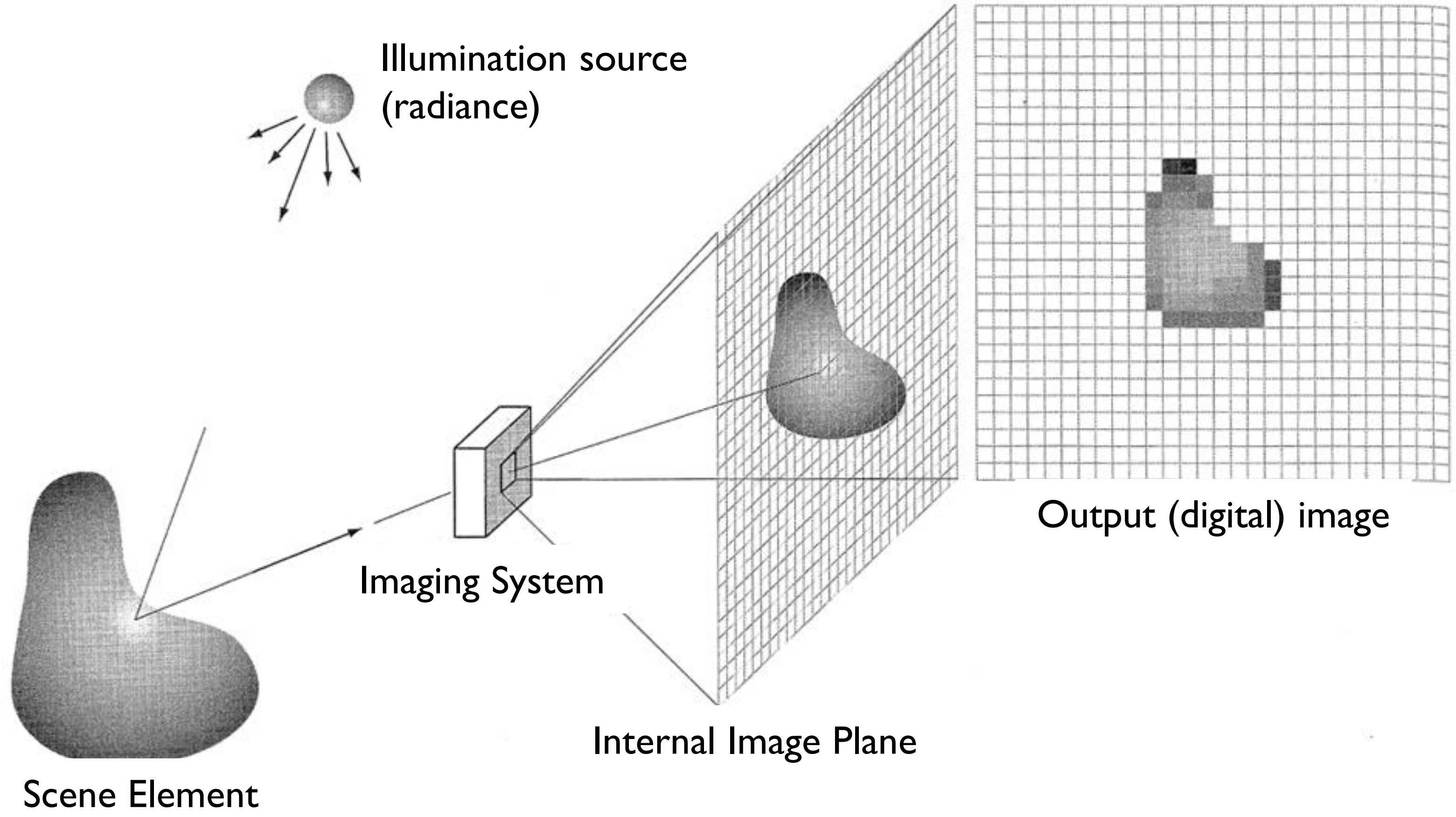


Photo by Uwe Hermann

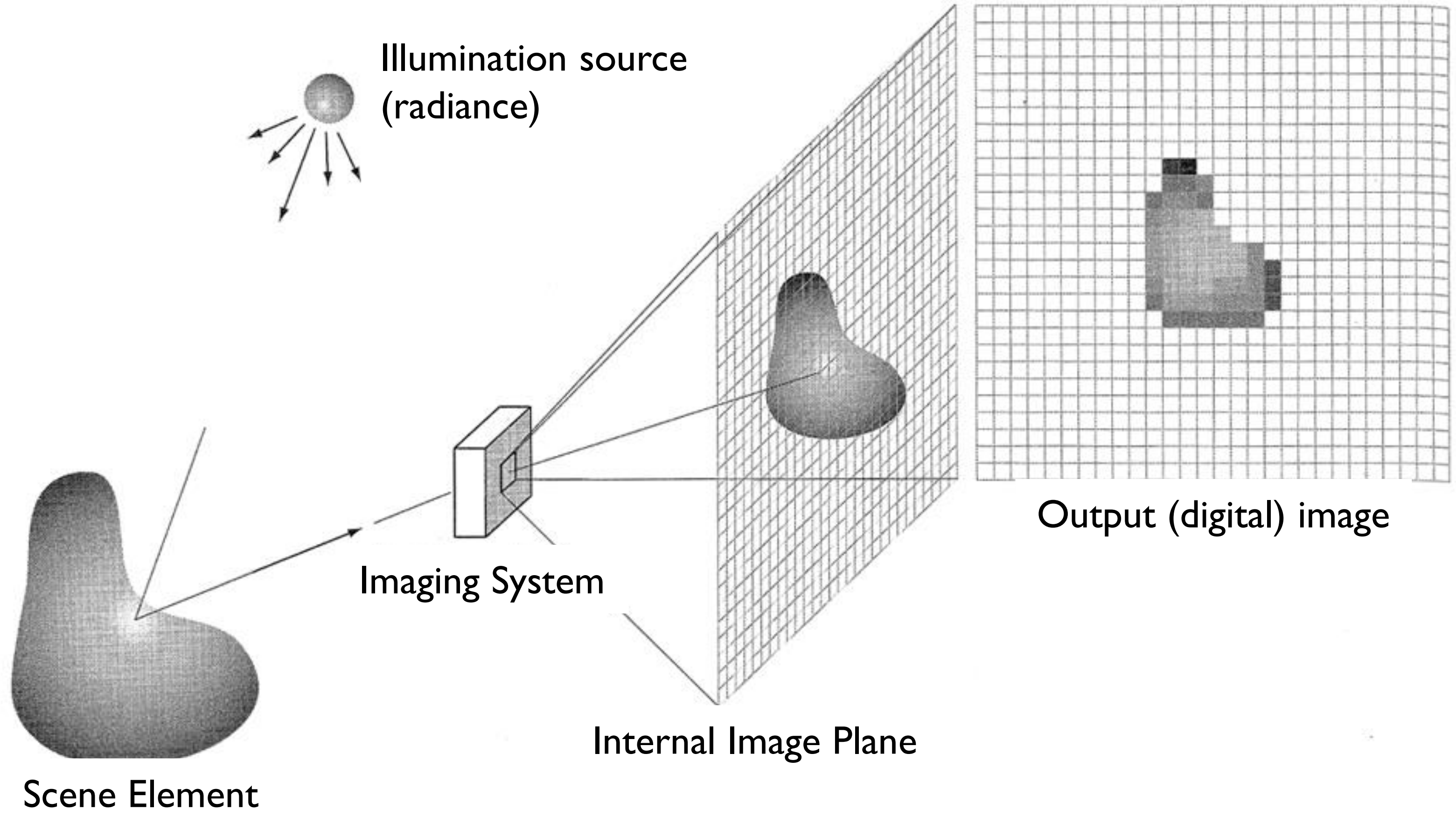
Image = radiant energy measurement



Simple models of a camera assumes an image is a "quantitative measurement" of scene radiance.

Figure from Digital Image Processing, Gonzales/Woods

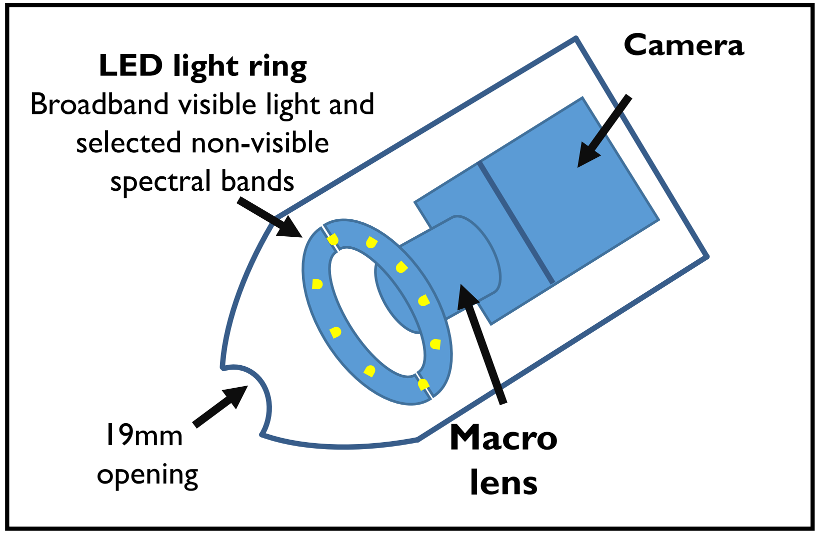
Camera = light measuring device



Simple models of a camera assumes an image is a “quantitative measurement” of scene radiance.

Figure from Digital Image Processing, Gonzales/Woods

Camera = light measuring device



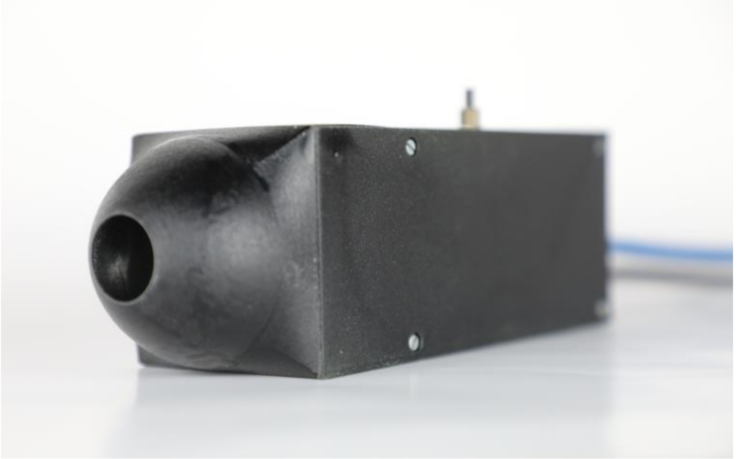
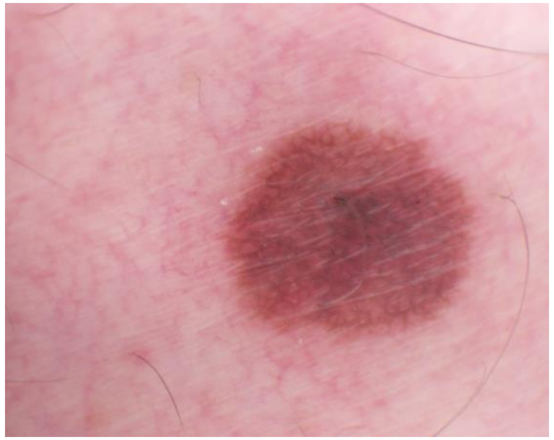


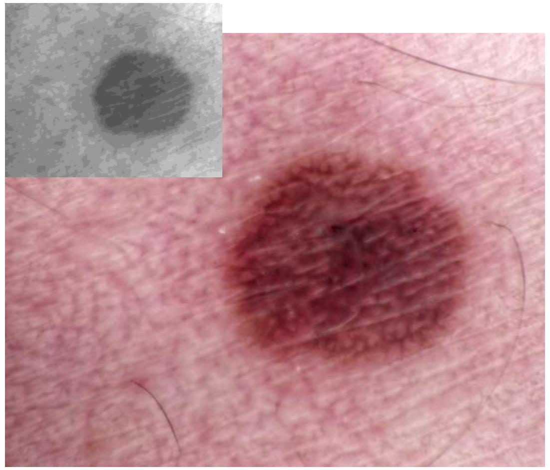
Image courtesy Elucid Labs

Medical imaging explicitly requires accurate measurements.

Visible image



Enhanced RGB image using a UV spectral band

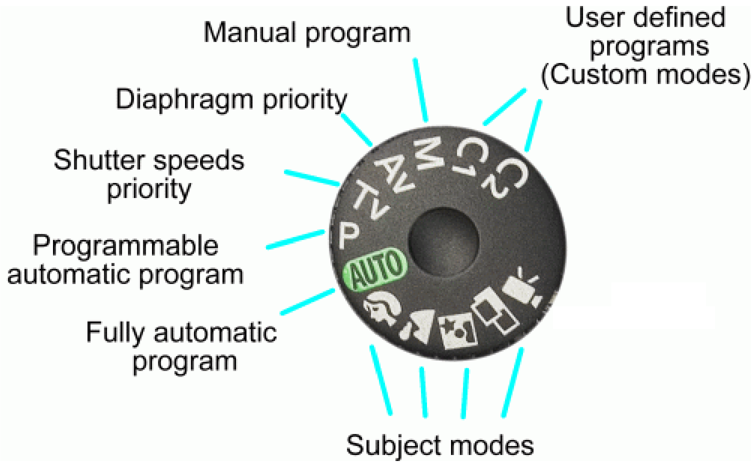


Camera = light measuring device

Image courtesy Scanwell Health

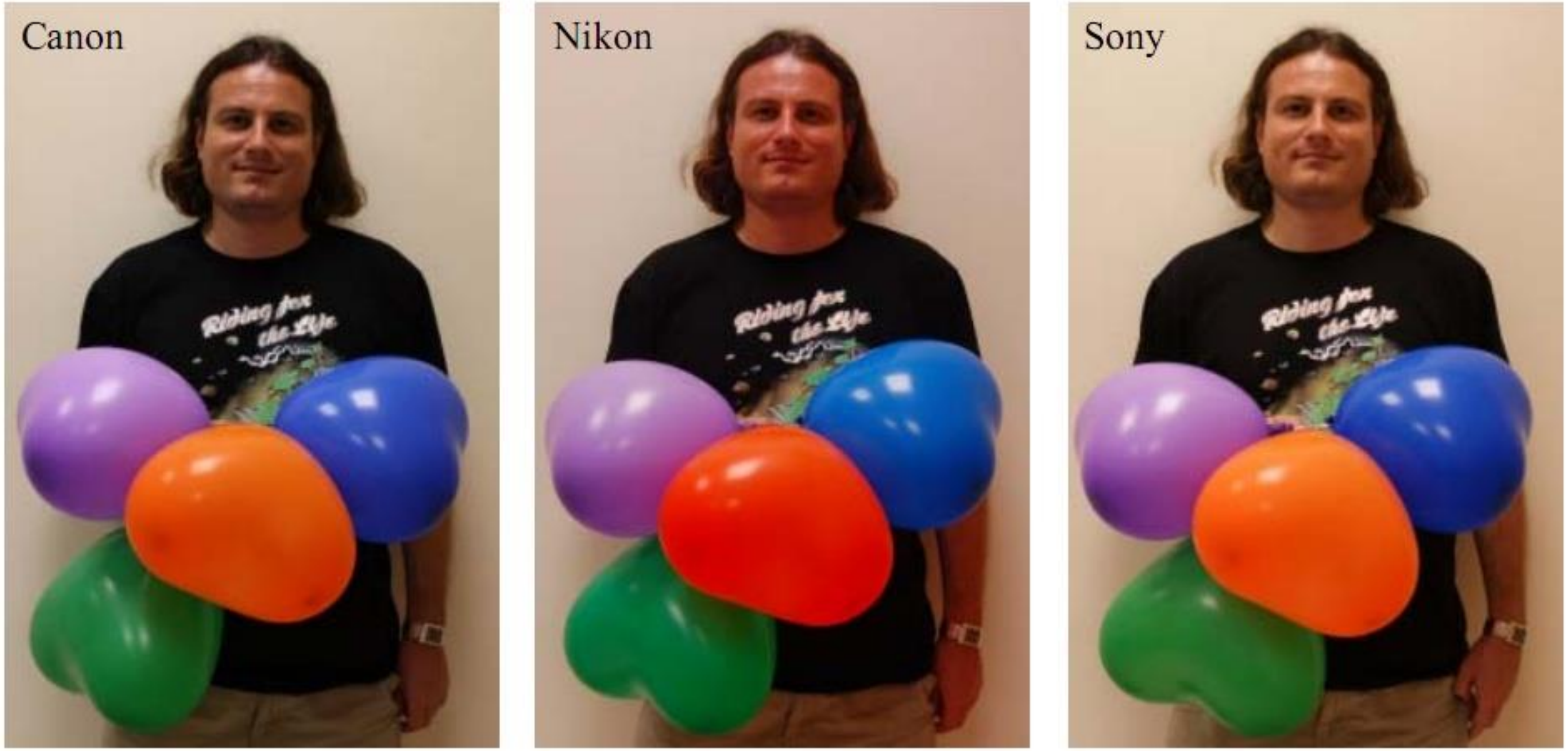
Home diagnostic testing requires accurate measurements across different cameras.

Camera = light measuring device?

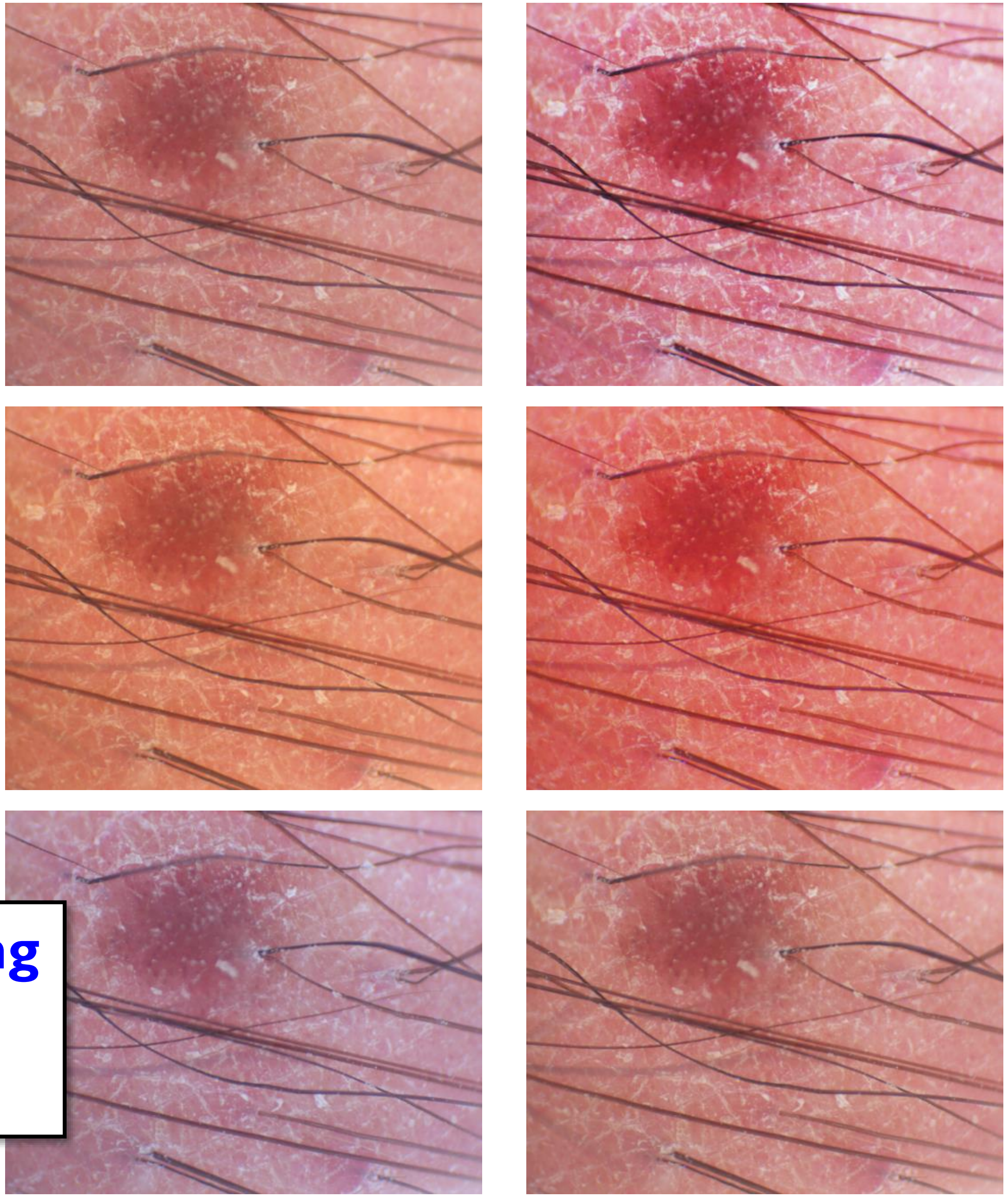


Portrait Mode	Soft Skin Mode	Transform Mode
Self-portrait Mode	Scenery Mode	Panorama Assist Mode
Sports Mode	Night Portrait Mode	Night Scenery Mode
Food Mode	Party Mode	Candle Light Mode
Baby Mode 1/2	Pet Mode	Sunset Mode
High Sensitivity Mode	High-speed Burst Mode	Flash Burst Mode
Starry Sky Mode	Fireworks Mode	Beach Mode
Snow Mode	Aerial Photo Mode	Pin Hole Mode
Film Grain Mode	High Dynamic Mode	Photo Frame Mode

In-camera photo-finishing is the "secret recipe" of a camera



Photographs taken from three different cameras with the same aperture, shutter speed, white-balance , ISO, and picture style.



In-camera photo-finishing may cause problems for scientific applications!

Which one is correct?

Motivation

- Cameras are the primary tool used to capture digital images.
- Digital images are the primary inputs to CV algorithms.
- CV researchers/engineers should have a basic understanding of how cameras work to inform their algorithms.
- This tutorial aims to provide this basic understanding.

A tutorial in three parts

Part 1: Review of color, color constancy, and color spaces

- CIE XYZ, chromatic adaption, color temperature, and output color spaces
- Background on color is necessary to understand Part 2

Part 2: Overview of a typical camera pipeline (ISP)

- Discuss the processing steps used by most ISPs
- Note that some steps are their own research topic

Part 3: Deep-learning/AI and the ISP

- Machine learning for individual ISP components
- Replacing the whole ISP with DNNs

Part 1:
Review of color, color constancy, color temperature, and color spaces

Color and color spaces

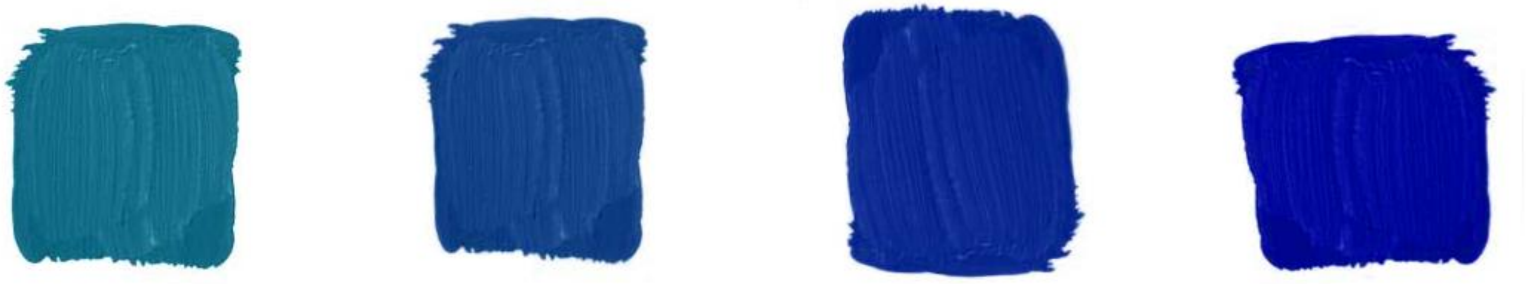
- To understand your camera, it is important to review how humans perceive color in a real environment.

- We must also understand how color is encoded by various models and color spaces.

- One of the main roles of the in-camera hardware is to convert the sensor image into a standard output-referred color space suitable for sharing and display.

Color is perceptual

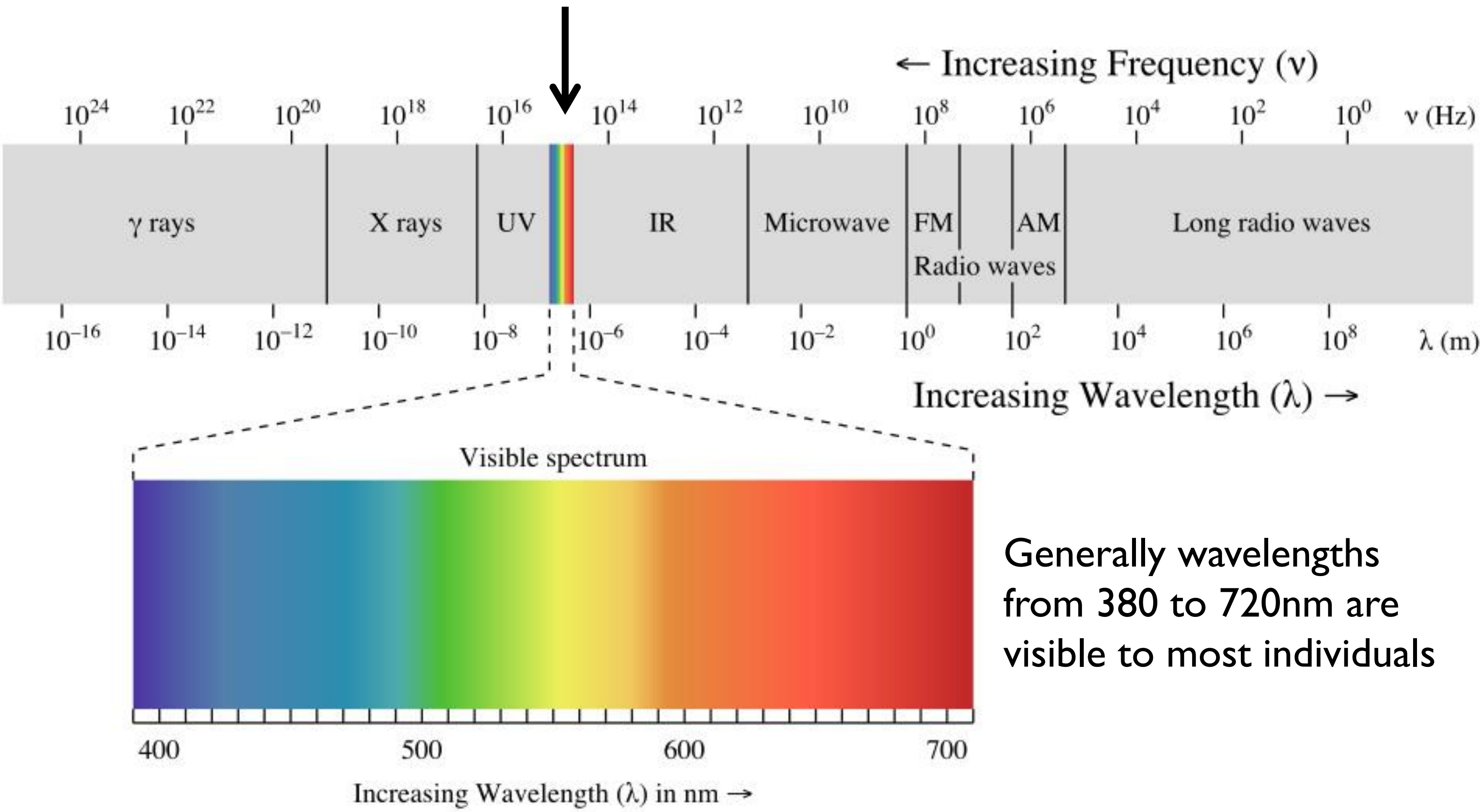
- Color is not a primary *physical* property of an object.
- Red, green, blue, pink, orange, purple, yellow, . . .
  - These are words we assign to visual sensations.
  - The assignment of words can vary among cultures.



Which is the "true blue"?

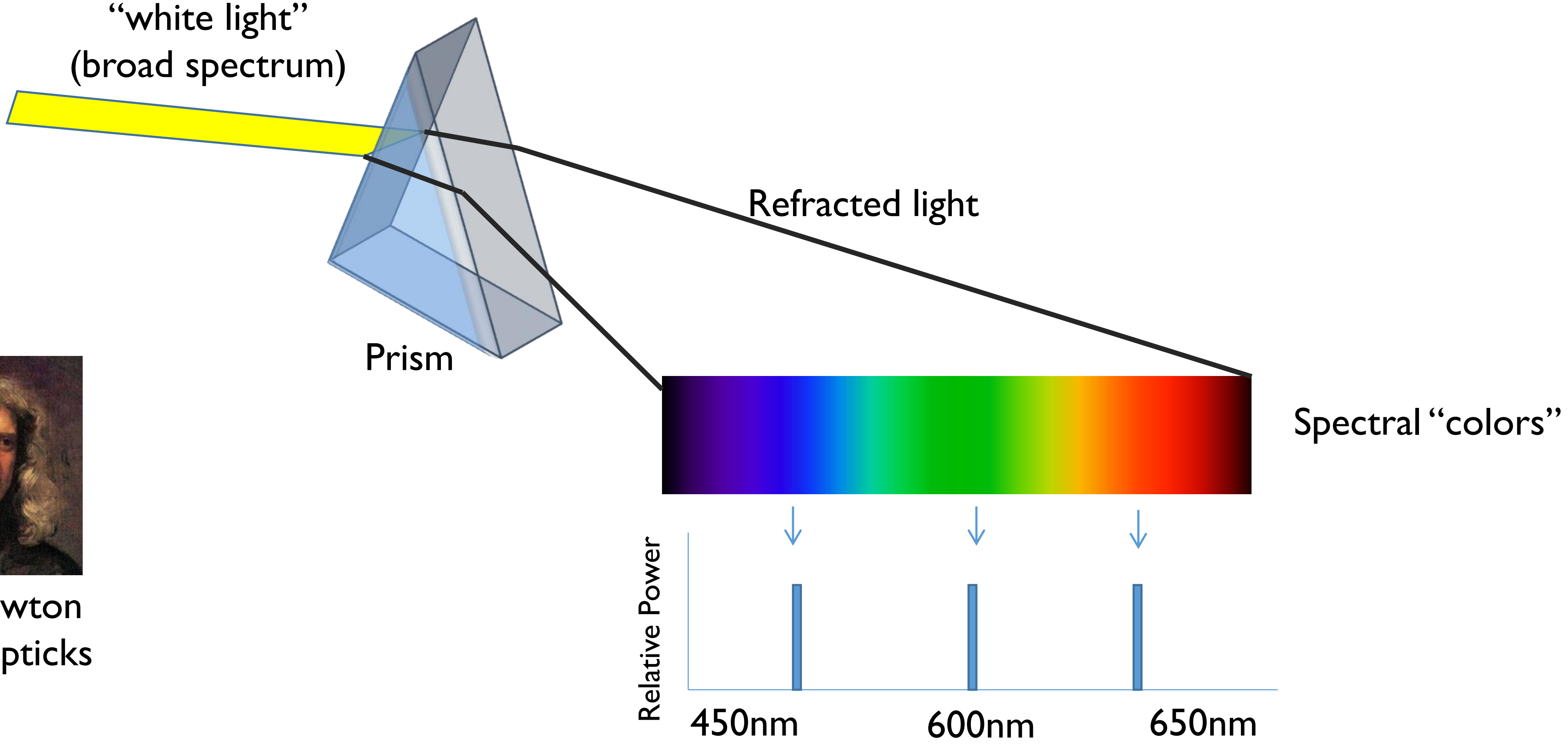
Where do "color sensations" come from?

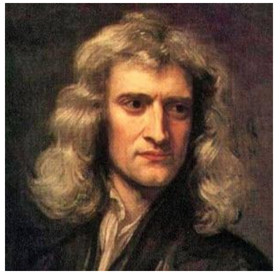
A *very* small range of electromagnetic radiation



Generally wavelengths from 380 to 720nm are visible to most individuals

White light through a prism



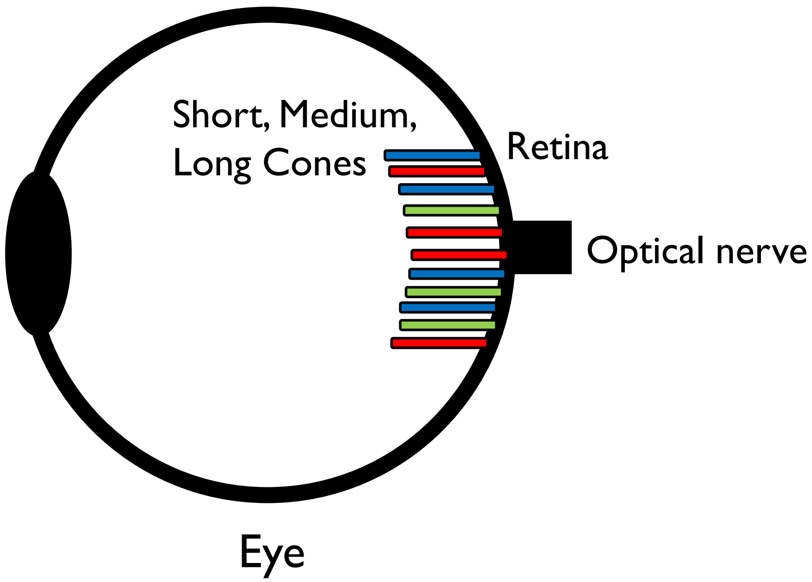


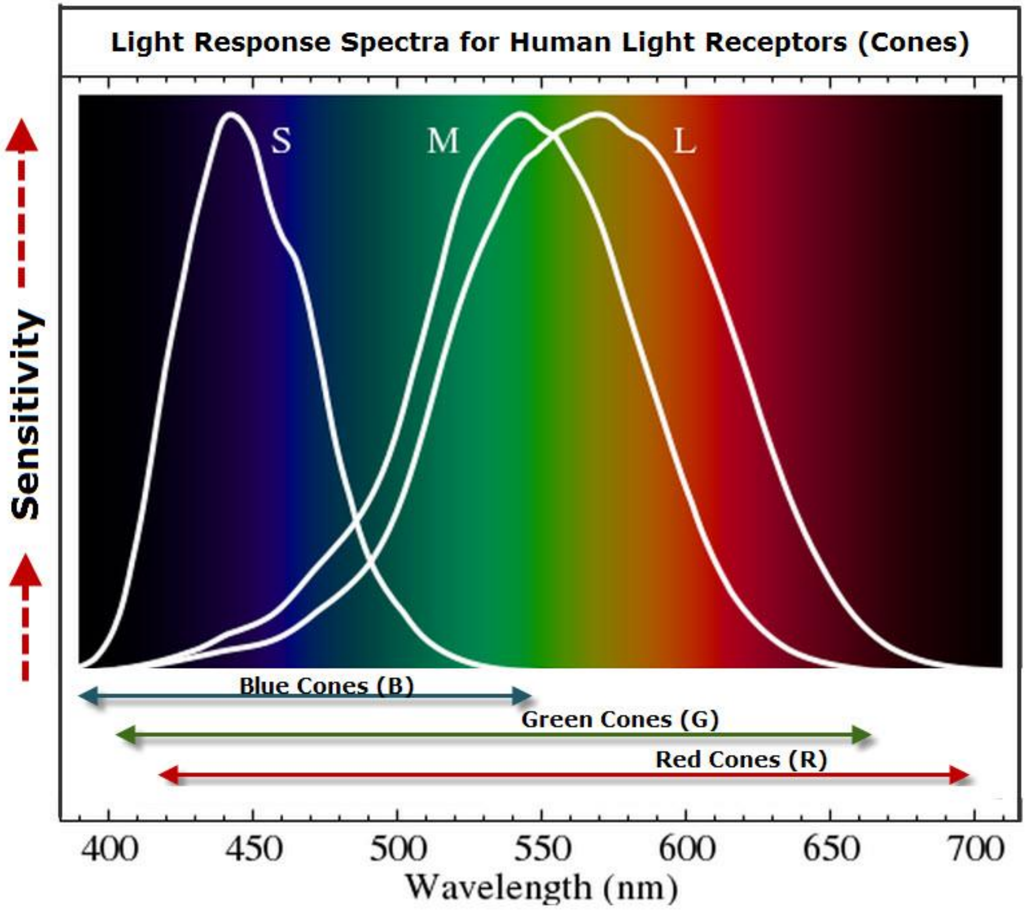
Isaac Newton
1704 - Opticks

Light is separated into "monochromatic" light at different wave lengths.

Biology of color sensations

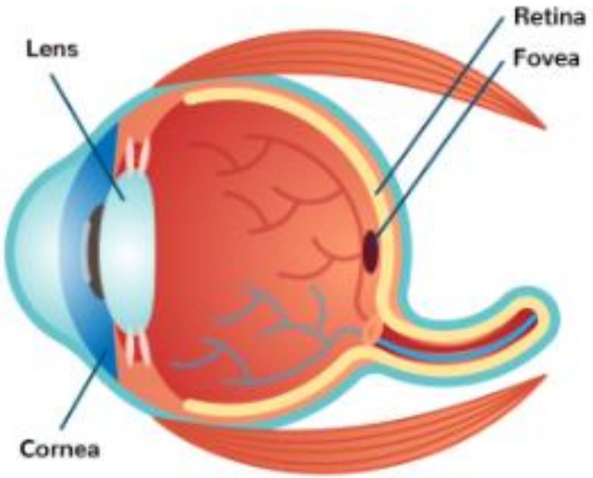
Our eye has three receptors (cone cells). The different cones respond to different ranges of the visible light spectrum.



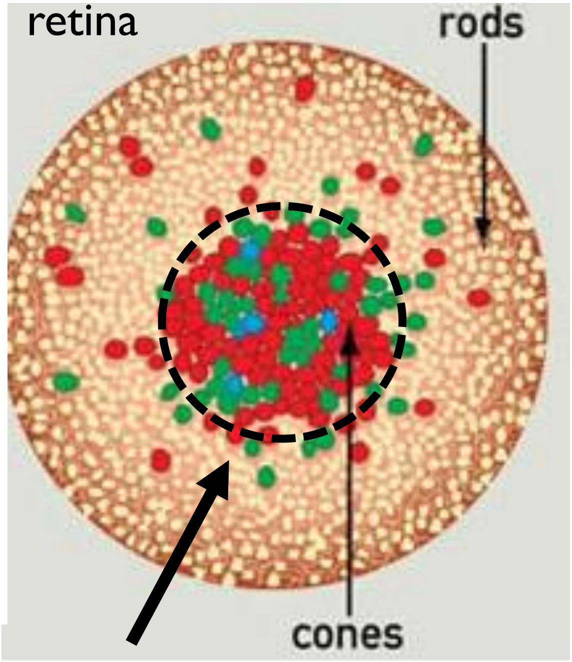


Cones and rods

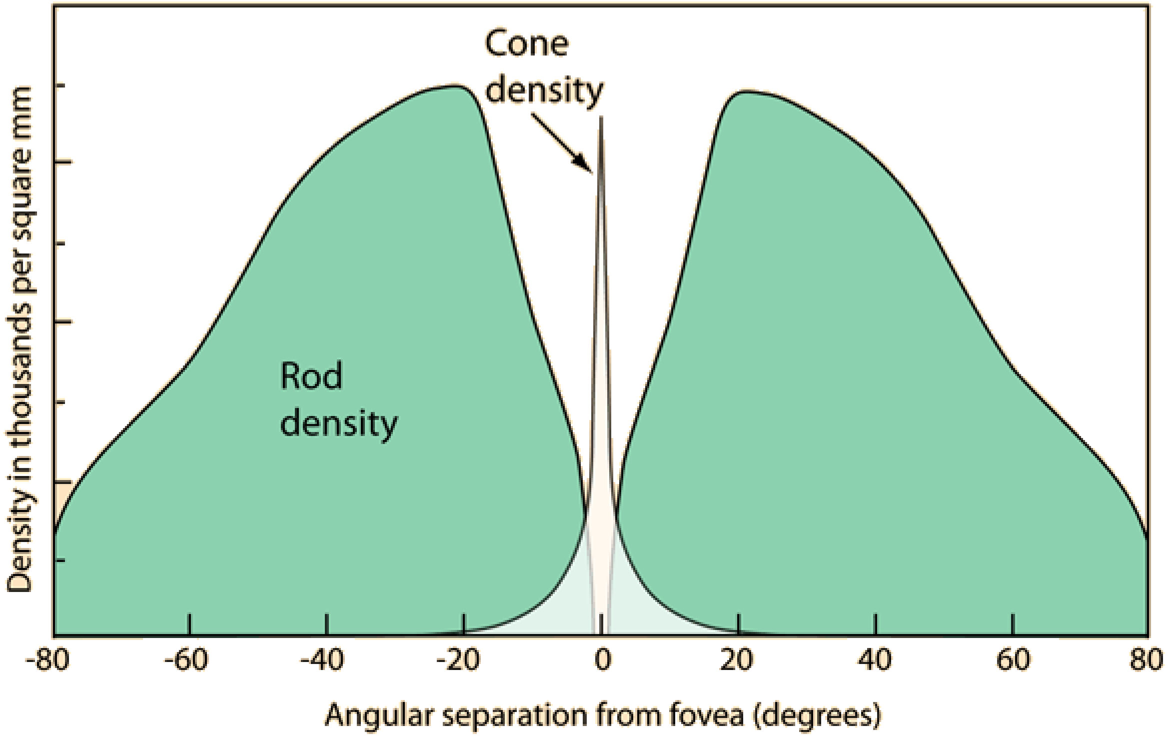
- We have additional light-sensitive cells called *rods* that are not responsible for color. Rods are used in low-light vision.
- Cone cells are most concentrated around the fovea of the eye.



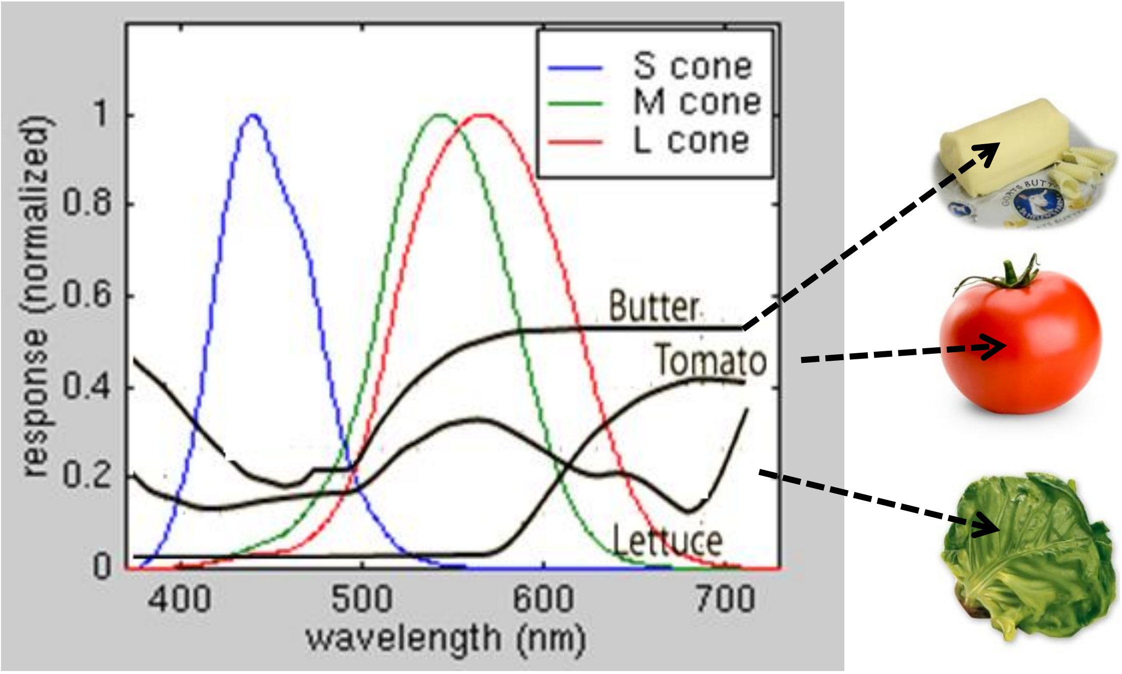
Human eye



Fovea region



Spectral power distribution (SPD)

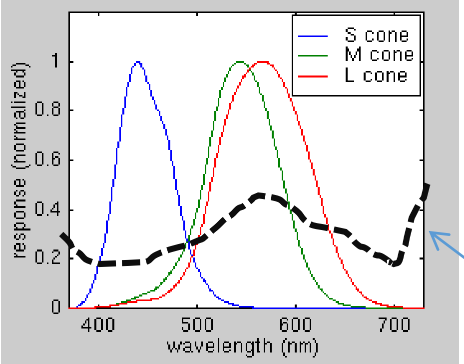


We rarely see monochromatic light in real world scenes. Instead, objects reflect a wide range of wavelengths. This can be described by a ***spectral power distribution*** (SPD) shown above. The SPD plot shows the relative amount of each wavelength reflected over the visible spectrum.

SPD relation to perceived color is not unique

Due to the accumulation effect of the cones, two different SPDs can be perceived as the same color (such SPDs are called "metamers").

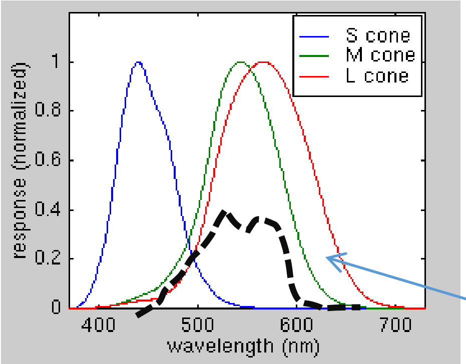
Lettuce SPD



Lettuce SPD stimulating S=0.2, M=0.8, L=0.8

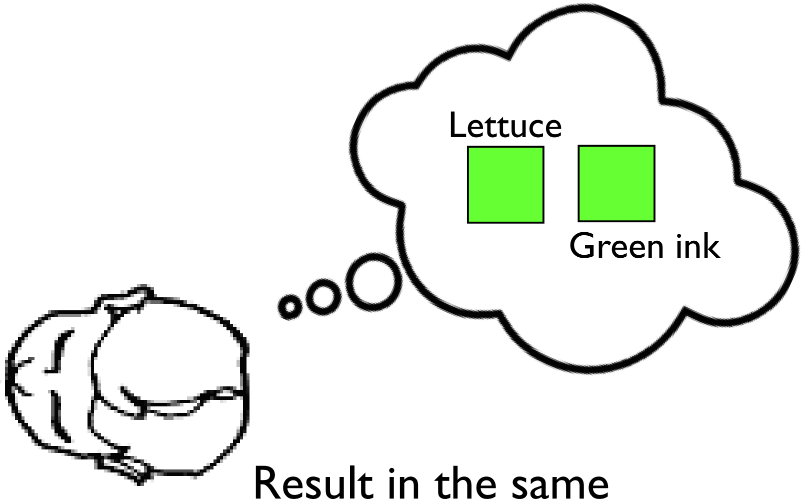
SPD of "real lettuce"

Green Ink SPD



Green ink SPD stimulating S=0.2, M=0.8, L=0.8

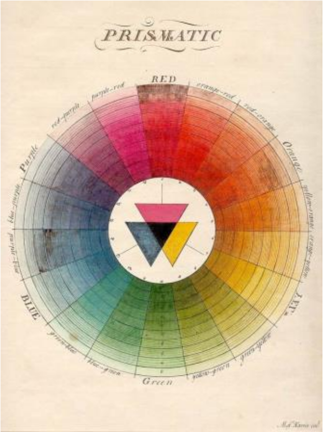
SPD of ink in a "picture of lettuce"



Result in the same color "sensation".

Tristimulus color theory

- Before the biology of cone cells was understood, it was empirically known that only three distinct colors (primaries) could be mixed to produce other colors.
- Moses Harris (1766), Thomas Young (1803), Johann Wolfgang von Goethe (1810), Hermann Grassman (1853), James Maxwell (1856) all explored the theory of trichromacy for human vision.



Harris

Young

von Goethe

Grassman

Maxwell

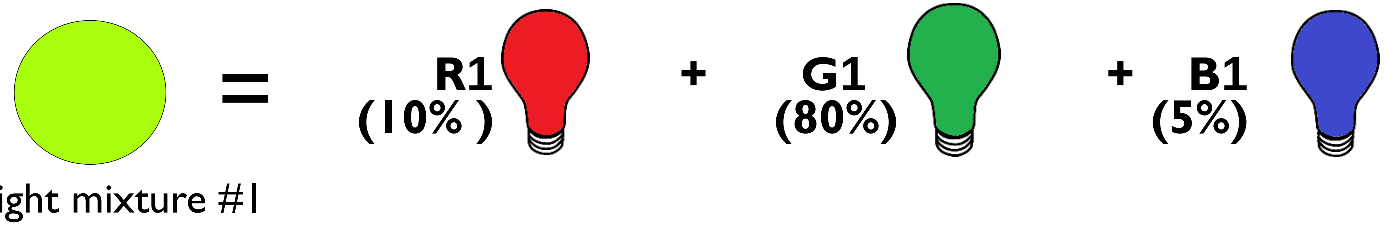
Maxwell’s color disks

Early color photography is attributed to Maxwell.

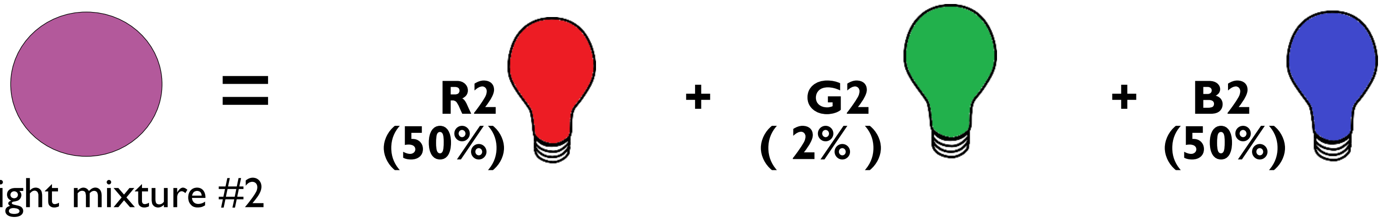
From Harris “The Natural System of Colours"

Tristimulus color theory

Grassman's Law states that a source color can be matched by a **linear** combination of three independent "primaries".



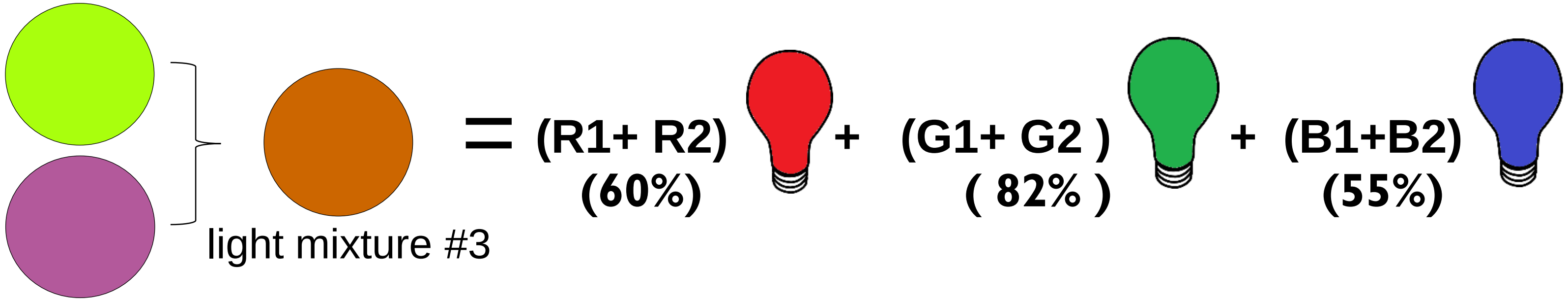
Three lights (shown as lightbulbs) serve as primaries. Each light has intensity, or weights, R1, G1, B1 to match the source light #1 perceived color.



Same three primaries and the weights (R2, G2, B2) of each primary needed to match the source light #2 perceived color

If we combine source lights 1 & 2 to get a new source light 3

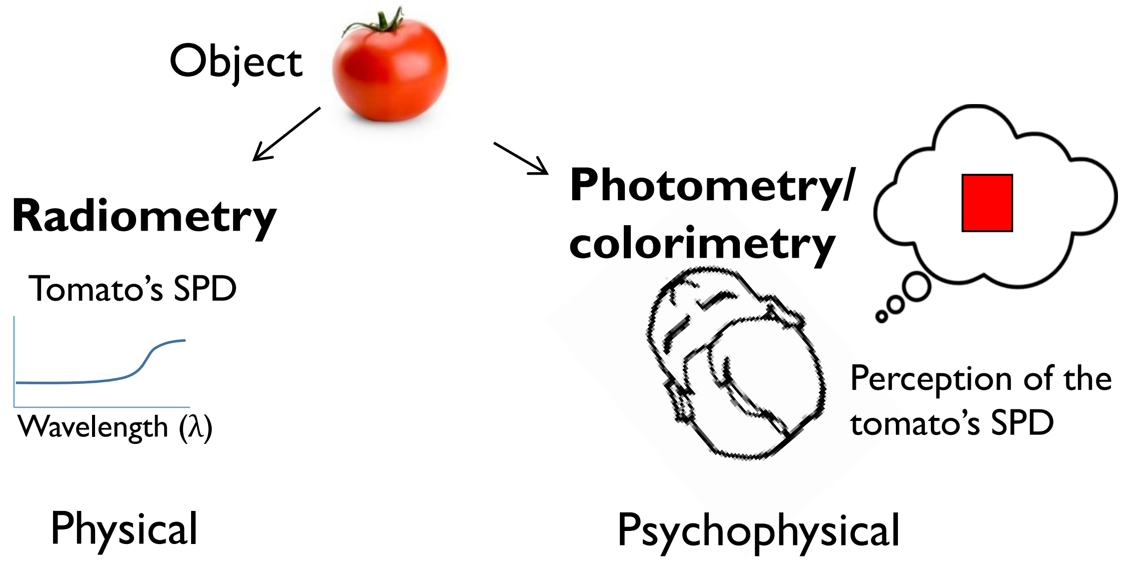
The amount of each primary needed to match the new source light #3 is the sum of the weights that matched lights sources #1 & #2.



This may seem obvious now, but discovering that light obeys the laws of linear algebra was a huge and useful discovery.

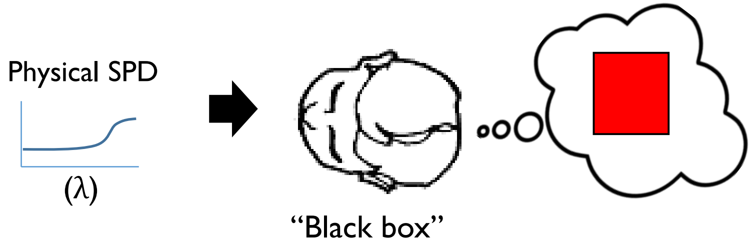
Radiometry vs. photometry/colorimetry

- **Radiometry**
  - Quantitative measurements of radiant energy.
  - Often shown as spectral power distributions (SPD).
  - Measures light coming from a source (radiance) or light falling on a surface (irradiance).
- **Photometry/ colorimetry**
  - Quantitative measurement of **perceived** radiant energy based on human's sensitivity to light.
  - Perceived in terms of "brightness" (photometry) and color (colorimetry).



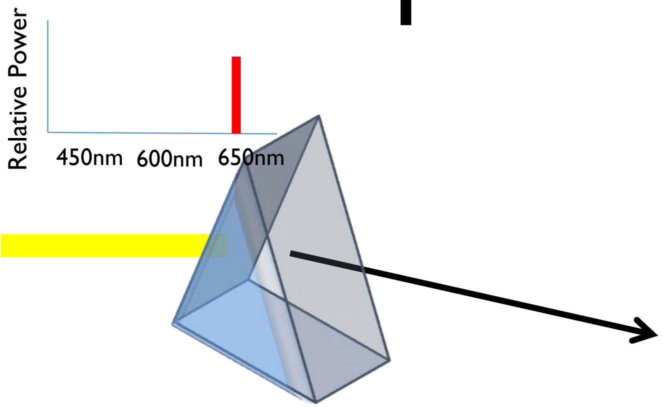
Quantifying color

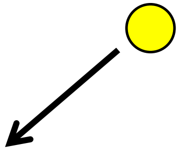
- Human cone photoreceptors (L/M/S) were being characterized well into the 2000s.[1,2]
- The need to quantify color and brightness existed much earlier.
- Since SPDs go through a "black box" (human visual system), the only way to quantify the "black box" is to perform a human study.
- Two key experiments
  - To quantify perceived "brightness" (photometry)
  - To quantify perceived "color" (colorimetry)



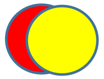
[1]Schnapf et al. "Spectral sensitivity of human cone photoreceptors," Nature 1987
[2]Stockman and Sharpe. "The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype," Vision Research 2000

Experiments for photometry





Reference bright light
with fixed radiant power.

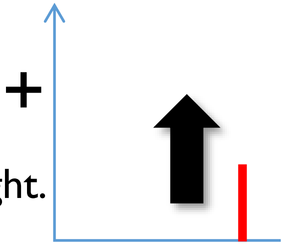


(Alternating between source and reference @ 17Hz)

Chromatic **source** light at a particular wavelength and adjustable radiant power.

Alternate between the source light and reference light 17 times per second (17 hz). A flicker will be noticeable unless the two lights have the same perceived "brightness".

Radiant power of chromatic light.

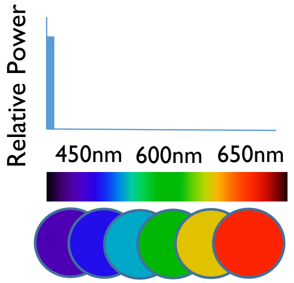


The viewer adjusts the radiant power of the chromatic light until the flicker disappears (i.e. the lights fuse into a constant color). The amount of radiant power needed for this fusion to happen is recorded.

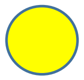
Repeat this flicker fusion test for each wave length in the source light. This allows method can be used to determine the perceived "brightness" of each wavelength.

The "flicker photometry" experiment for photopic sensitivity.

Result of the flicker experiments

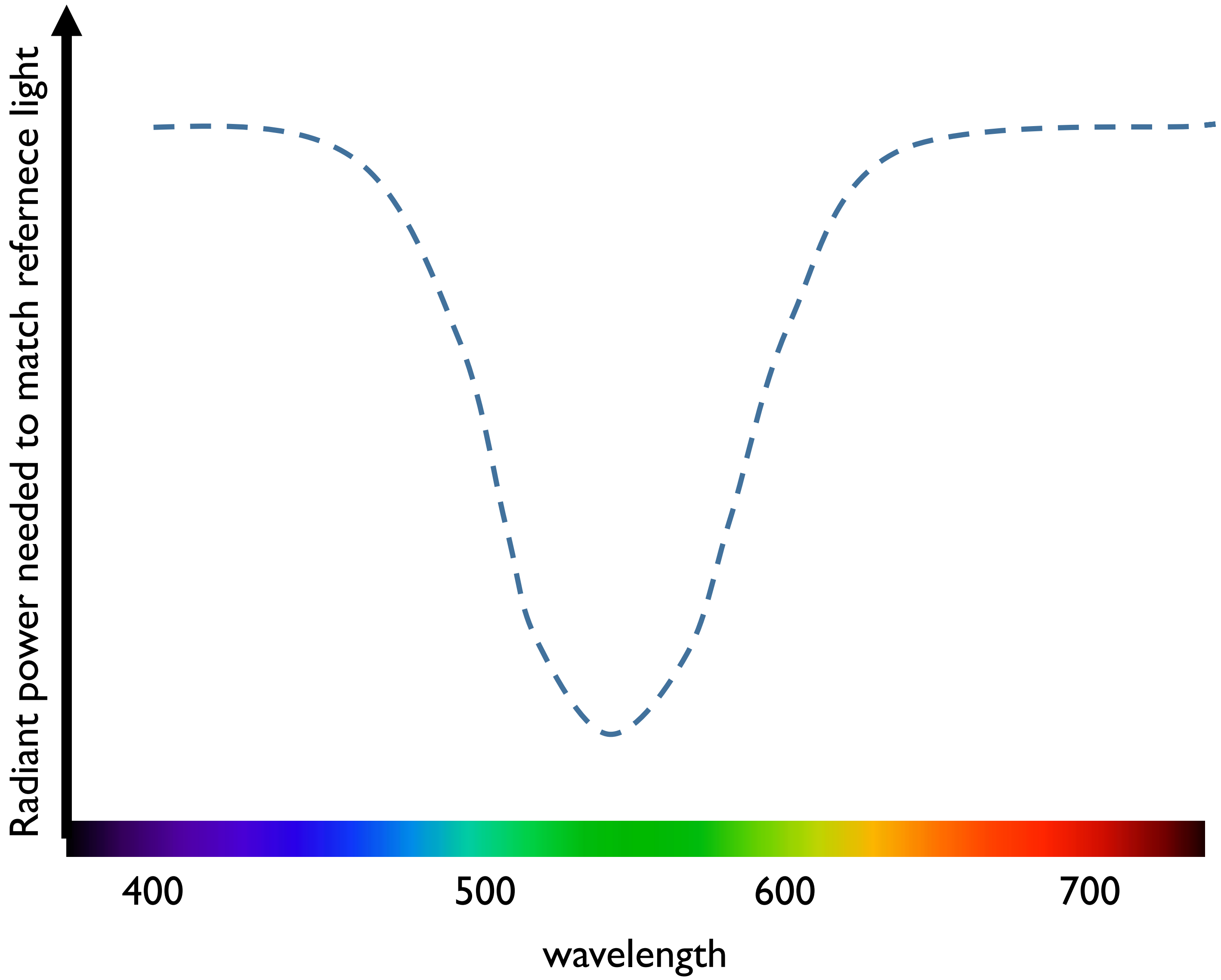


Monochromatic light



Reference light

Perform the flicker experiment for each wavelength.



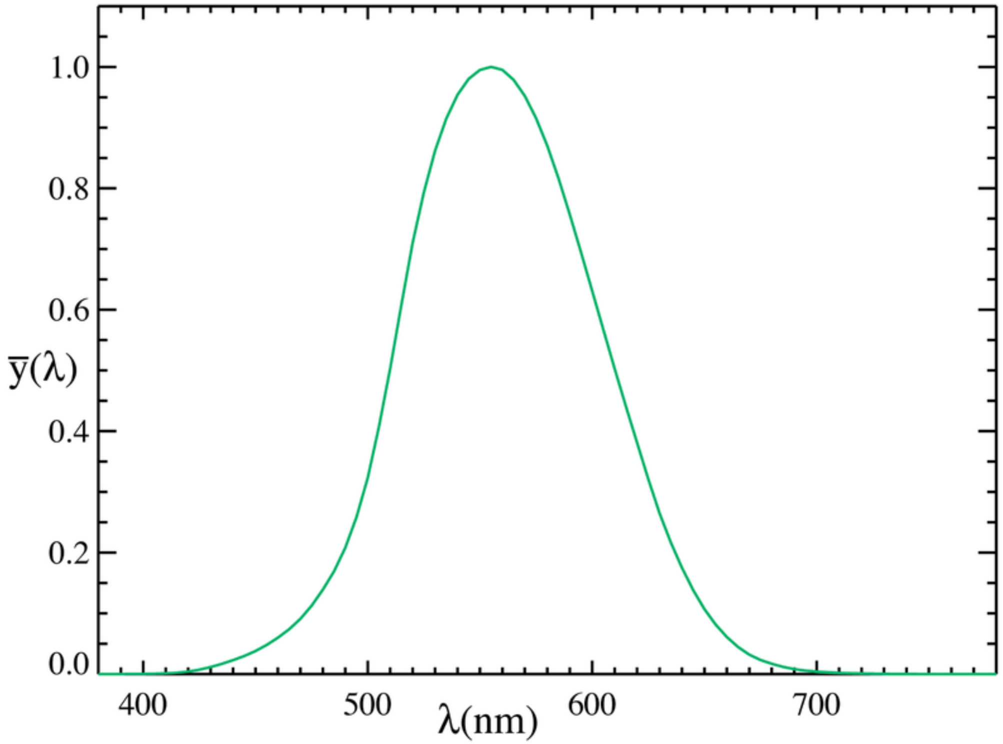
Amount of radiant power needed for each wavelength to make the reference light.

You need a lot more 400nm light to match the reference than you do the 550nm.

This means you perceive 550nm brighter than 400nm.

CIE (1924) Photopic luminosity function

If we invert the
curve on the
previous slide,
we get the
luminosity function.



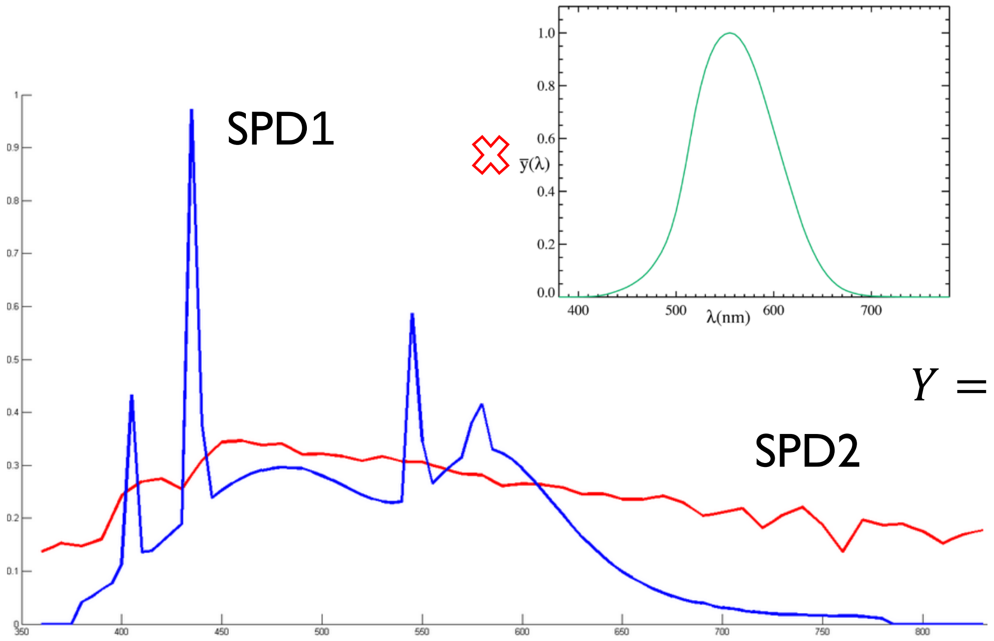
The Luminosity Function (written as $\bar{y}(\lambda)$ or $V(\lambda)$) shows the eye's sensitivity to radiant energy into luminous energy (or perceived radiant energy) based on human experiments (flicker fusion test).

International Commission on Illumination (CIE comes from the French name *Commission internationale e l'éclairage*) was a body established in 1913 as an authority on light, illumination and color . . CIE is still active today -- http://www.cie.co.at

Radiometric to Photometric

How do we use CIEY (or $\bar{y}\,(\lambda)$)?

SPD1 and SPD2 are clearly different. Which one will be perceived brighter (assuming the same overall radiant power.)



Two SPDs

$$Y = \int_{380}^{780} SPD(\lambda)\bar{y}\,(\lambda)\, d\lambda$$

Which SDP is perceived brighter?

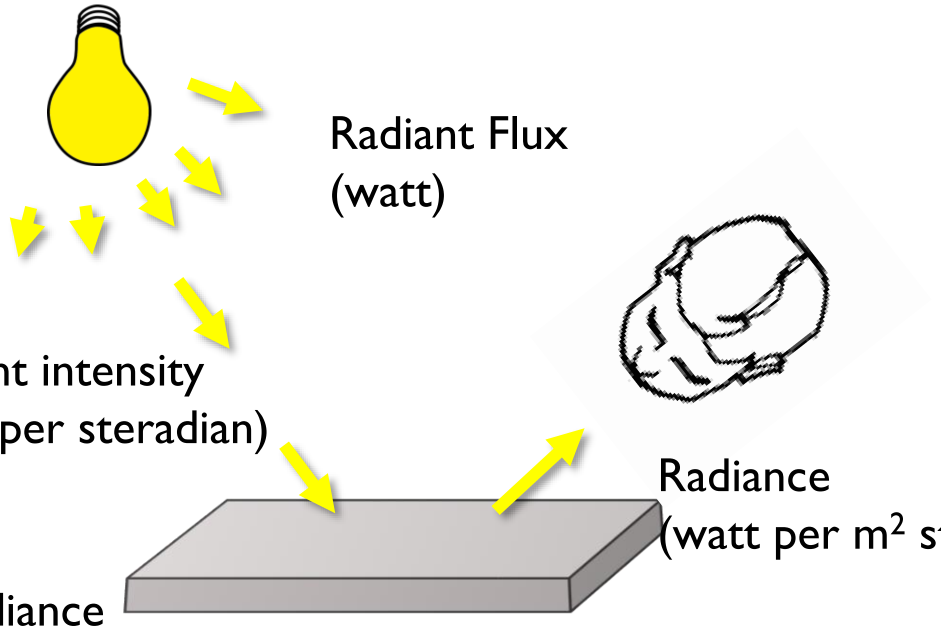
<u>SPD1</u>
Y=0.2989

<u>SPD2</u>
Y=0.2989

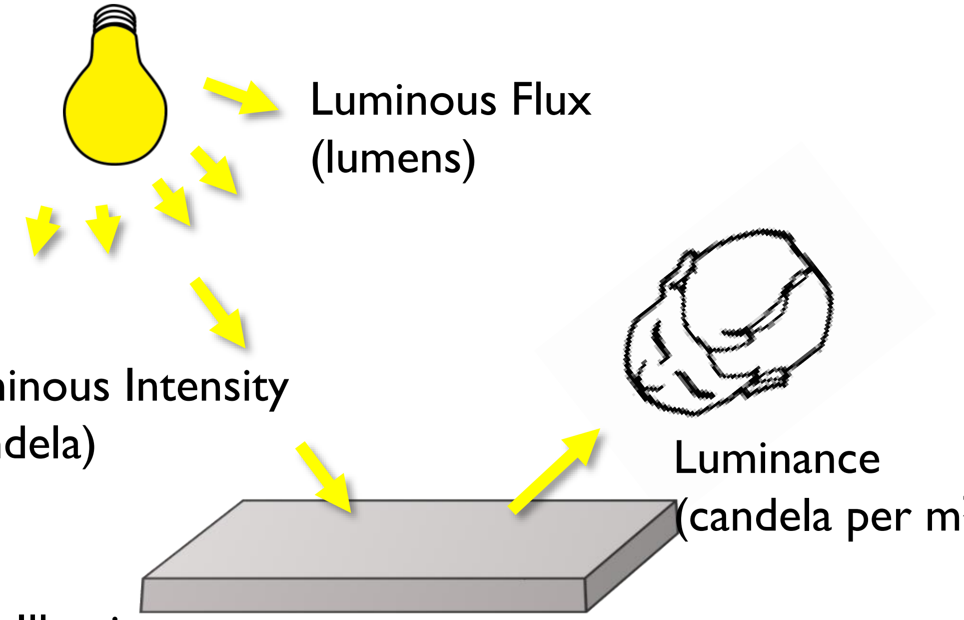
Radiometric **Photometric**

CIEY gives a way to go from radiometric to photometric!
Now can quantify the perceived brightness of different light.

Radiometric vs. photometric units



Radiometric values

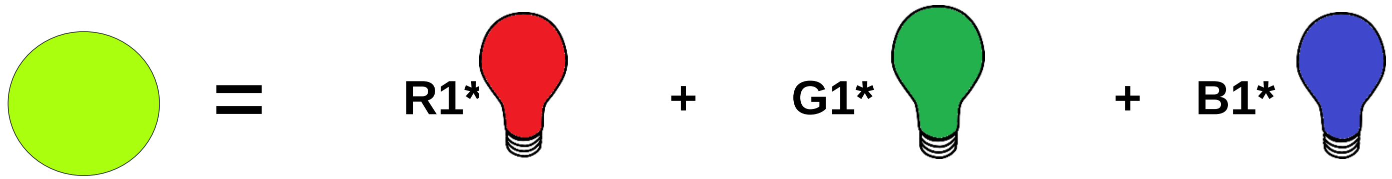


Photometric values
(Radiometric values weighted by the Luminosity Function)

3100 lumens colour brightness

Colorimetry

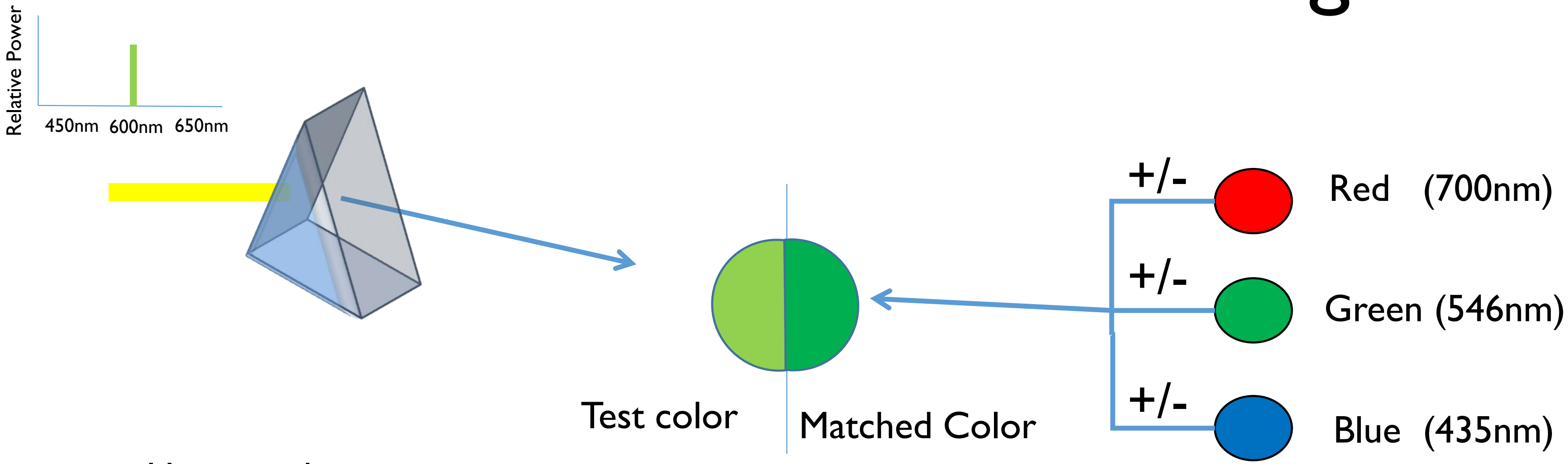
Based on tristimulus color theory, colorimetry attempts to quantify all visible colors in terms of a standard set of primaries



Target color

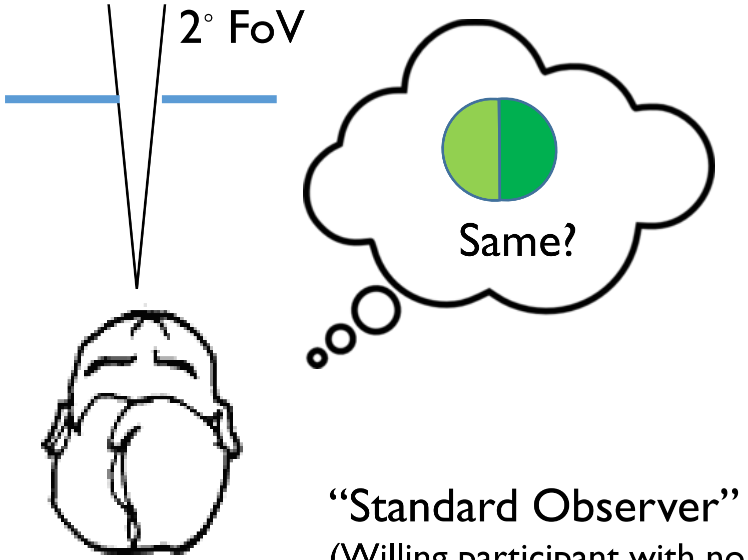
Three fixed primary lights.

CIE RGB color matching

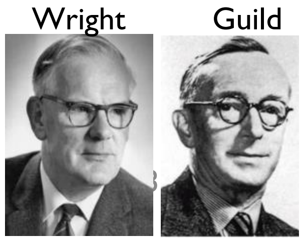


Human subjects matched test colors by add or subtracting three primaries.

Field of view was 2-degrees (where color cones are most concentrated)

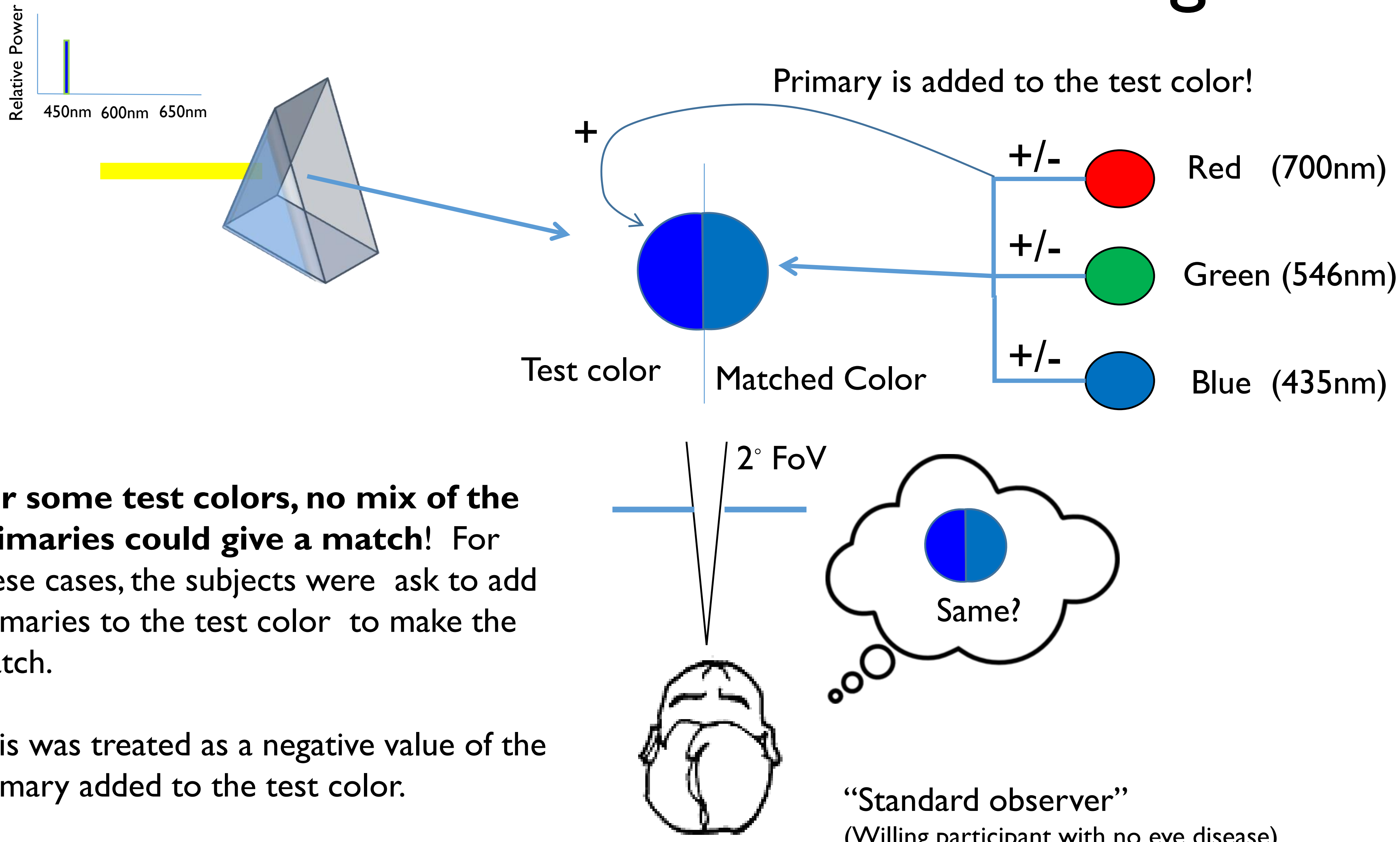


"Standard Observer"
(Willing participant with no eye disease)



Experiments carried out by
W. David Wright (Imperial College) and John Guild (National Physical Laboratory, London) – Late 1920s

CIE RGB color matching

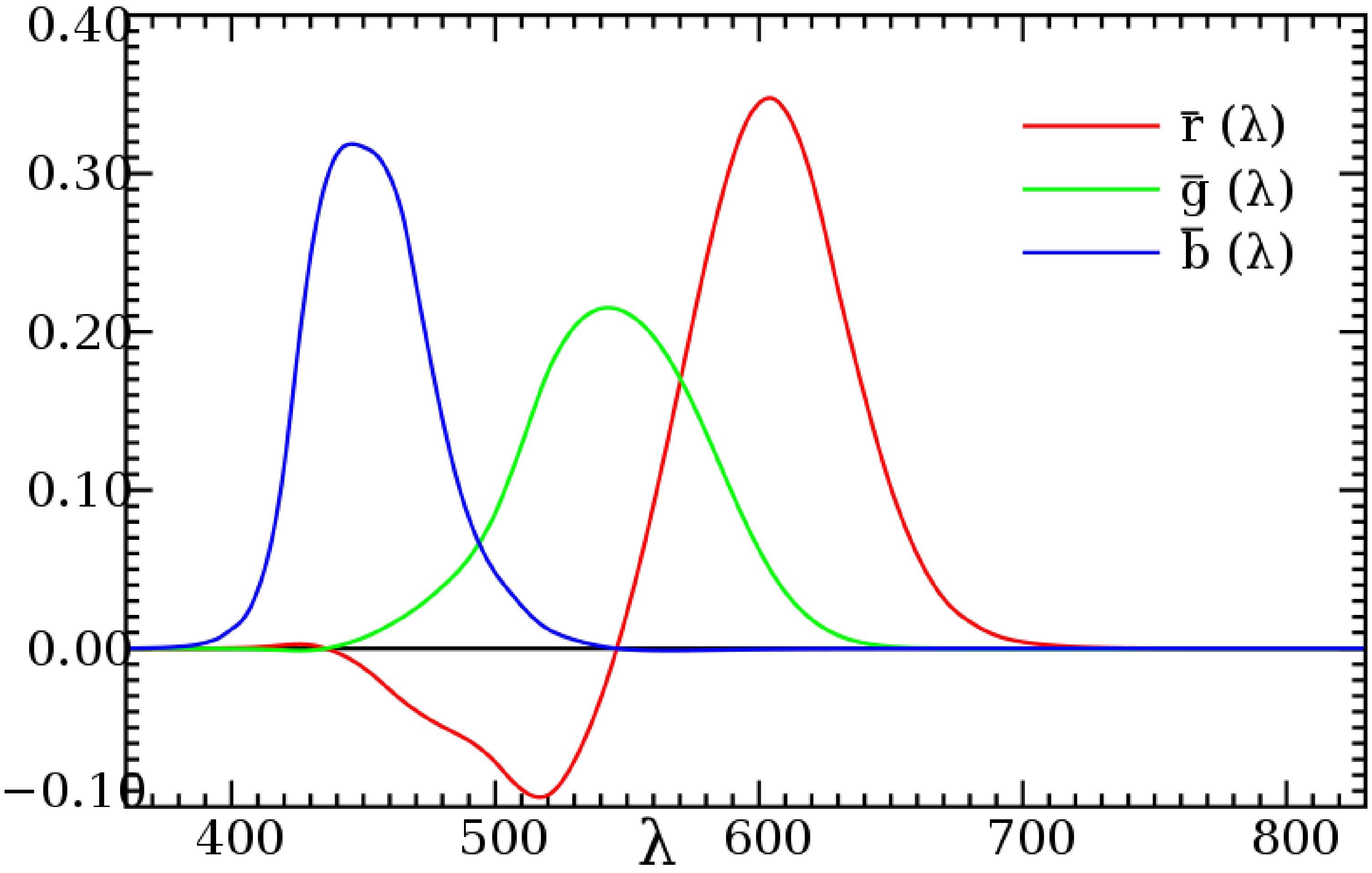


For some test colors, no mix of the primaries could give a match! For these cases, the subjects were ask to add primaries to the test color to make the match.

This was treated as a negative value of the primary added to the test color.

"Standard observer"
(Willing participant with no eye disease)

CIE RGB results

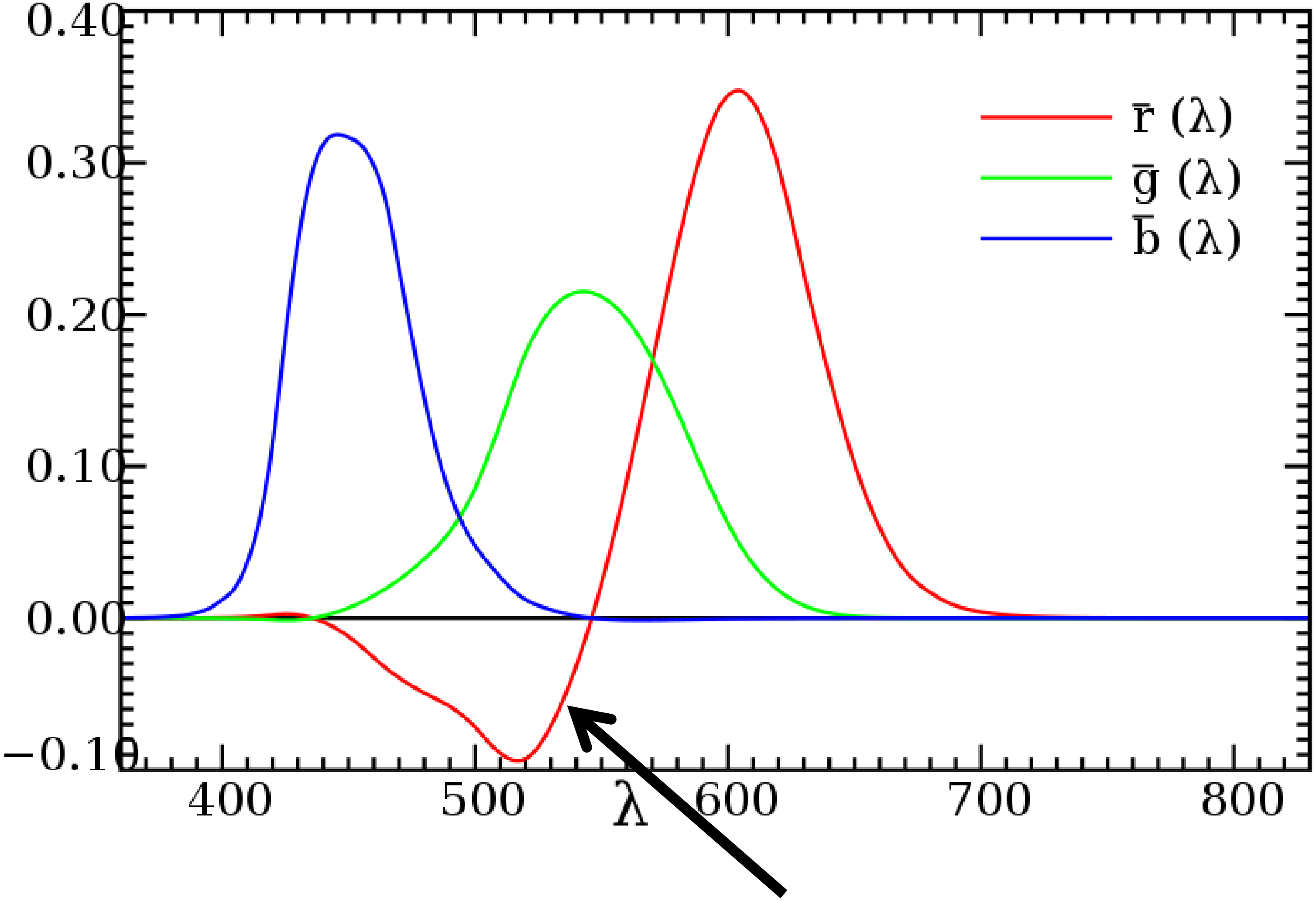


CIE RGB 2-degree Standard Observer
(based on Wright/Guild's data)

Plots are of the mixing coefficients of each primary needed to produce the corresponding monochromatic light at that wavelength.

Note that these functions have been scaled such that area of each curve is equal.

CIE RGB results



Negative values -- the three primaries used did not span the full range of perceptual colors.

CIE 1931 XYZ

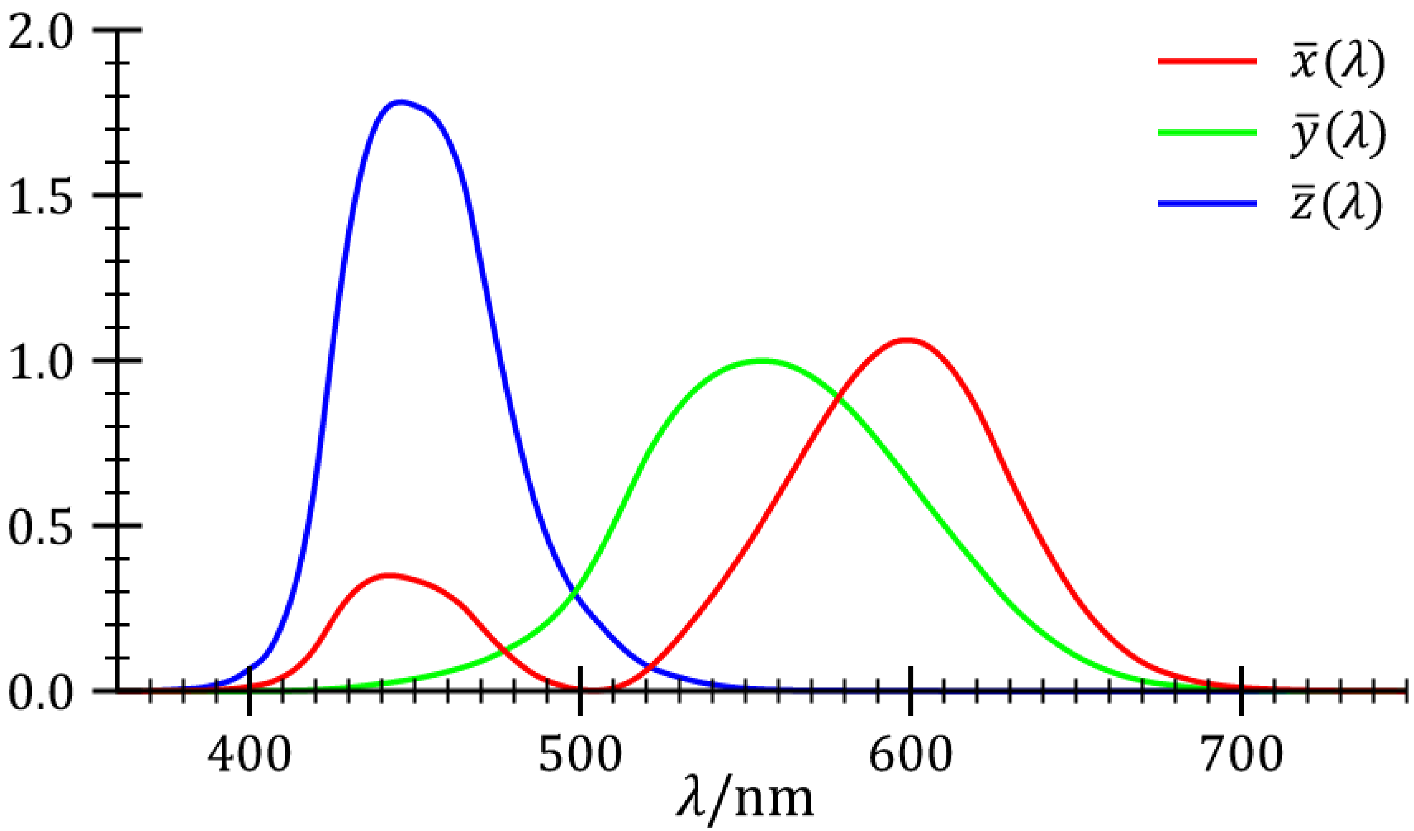
- In 1931, the CIE met and approved defining a new canonical basis, termed XYZ that would be derived from Wright-Guild's CIE RGB data.
- Properties desired in this conversion:
  - Positive values only
  - Pure white light (flat SPD) to lie at `X=1/3, Y=1/3, Z=1/3`
  - Y would be the luminosity function ($V(\lambda)$)
- Quite a bit of freedom in selecting the XYZ basis
  - In the end, the adopted transform was:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4887180 & 0.3106803 & 0.2006017 \\ 0.1762044 & 0.8129847 & 0.0108109 \\ 0.0000000 & 0.0102048 & 0.9897952 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

CIE 1931 RGB

Nice article see: Fairman et al "How the CIE 1931 Color-Matching Functions Were Derived from Wright–Guild Data", Color Research & Application, 1997

CIE 1931 XYZ

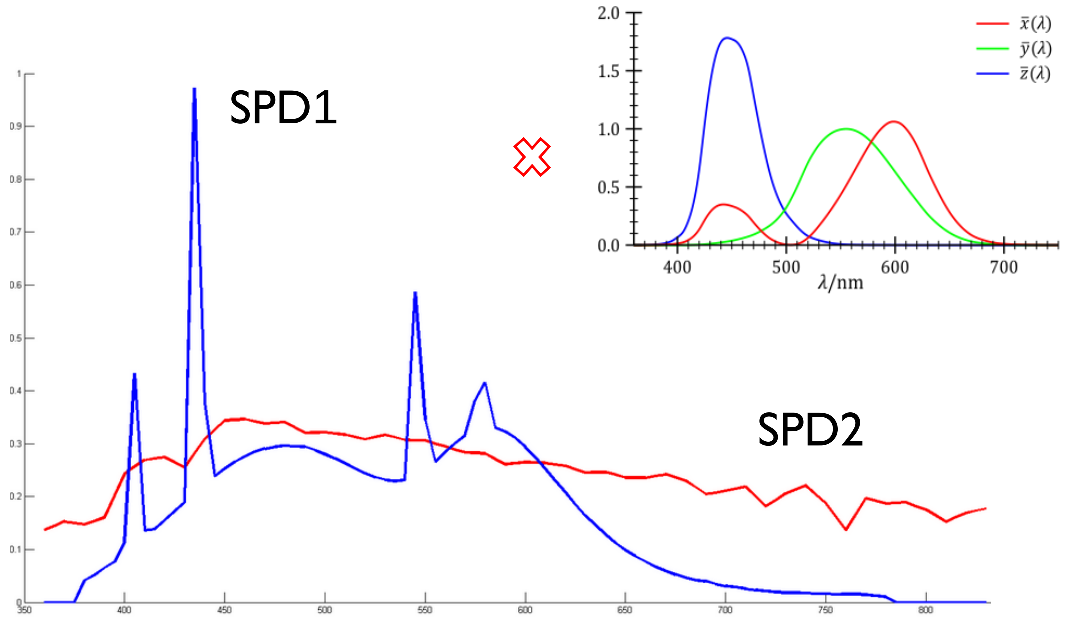


This shows the mixing coefficients $\overline{x}(\lambda)$, $\overline{y}(\lambda)$, $\overline{z}(\lambda)$ for the CIE 1931 2-degree standard observer XYZ basis computed from the CIE RGB data. Coefficients are all now positive. Note that the basis XYZ are not physical SPD like in CIE RGB, but linear combinations defined by the matrix on the previous slide.

SPD to CIE XYZ example

How do we use CIE XYZ?

SPD1 and SPD2 are clearly different. Will they be perceived as the same color?



Two SPDs

CIE XYZ Values

SPD1
X=0.2841
Y=0.2989
Z=0.3254

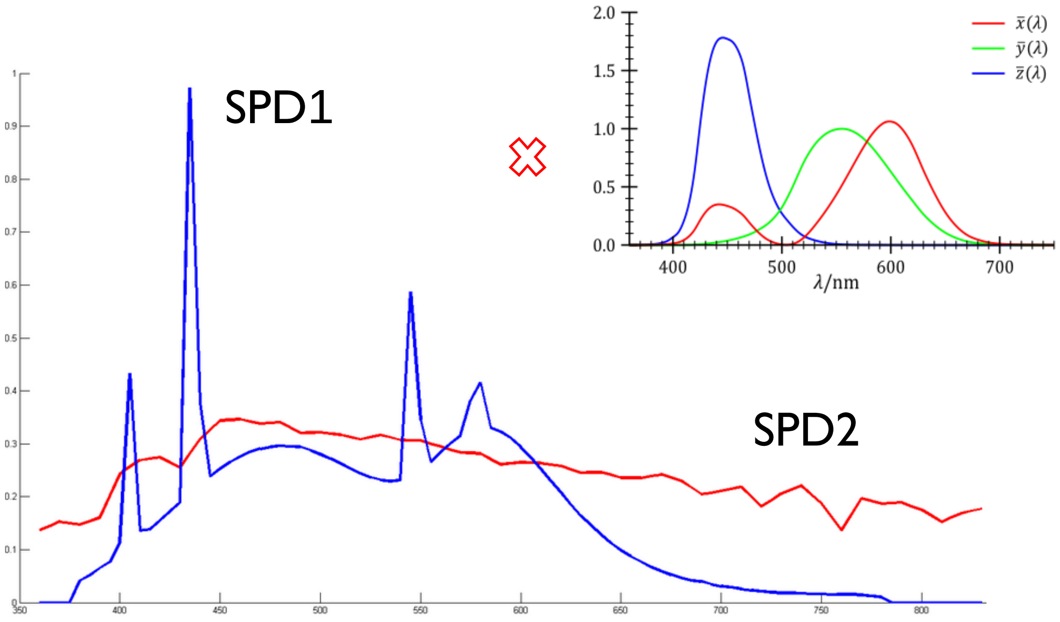
SPD2
X=0.2841
Y=0.2989
Z=0.3254

$$X = \int_{380}^{780} SPD(\lambda)\bar{x}(\lambda)\, d\lambda \qquad Y = \int_{380}^{780} SPD(\lambda)\bar{y}(\lambda)\, d\lambda \qquad Z = \int_{380}^{780} SPD(\lambda)\bar{z}(\lambda)\, d\lambda$$

SPD to CIE XYZ example

How do we use CIE XYZ?

SPD1 and SPD2 are clearly different. Will they be perceived as the same color?



Two SPDs

CIE XYZ Values

<u>SPD1</u>
X=0.2841
Y=0.2989
Z=0.3254

<u>SPD2</u>
X=0.2841
Y=0.2989
Z=0.3254

From their CIE XYZ mappings, we can determine that these two SPDs will be *perceived* as the same color!

Now we can quantify color!

Radiometric **Photometric/Colorimetric**

CIE XYZ gives a way to go from radiometric to colorimetric. Imbedded is also the photometric measurement in the Y value.

CIE XYZ Plot

It is challenging to visualize the 3D CIE XYZ space.
We often don't plot color in this space.

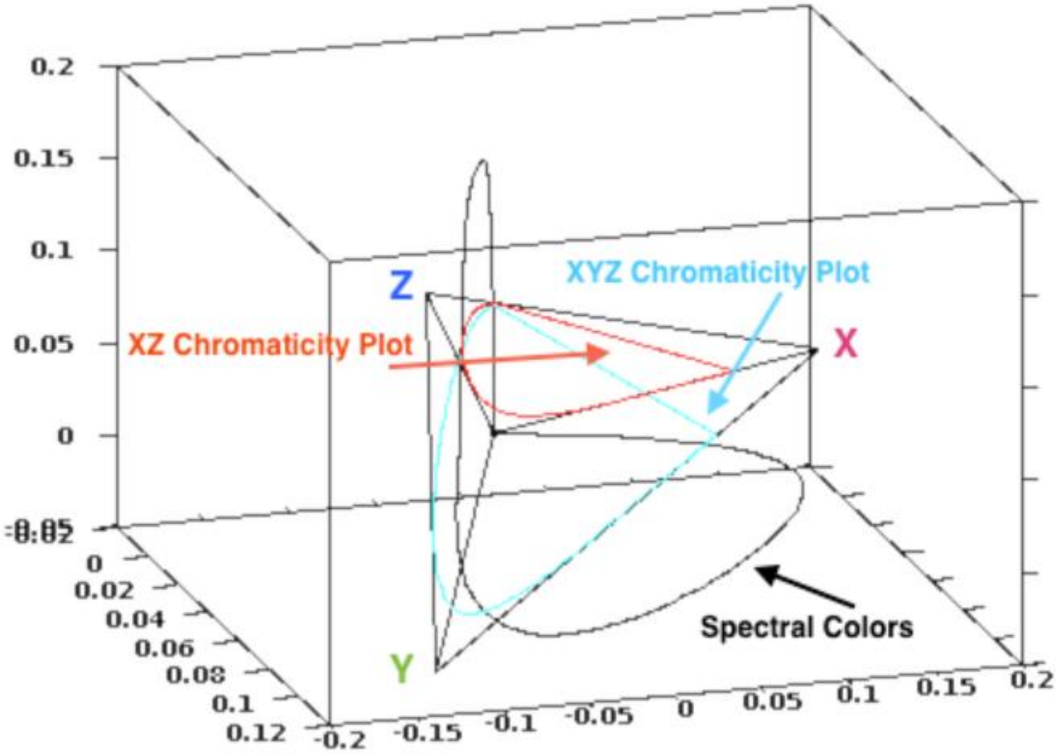


Image: Joffa

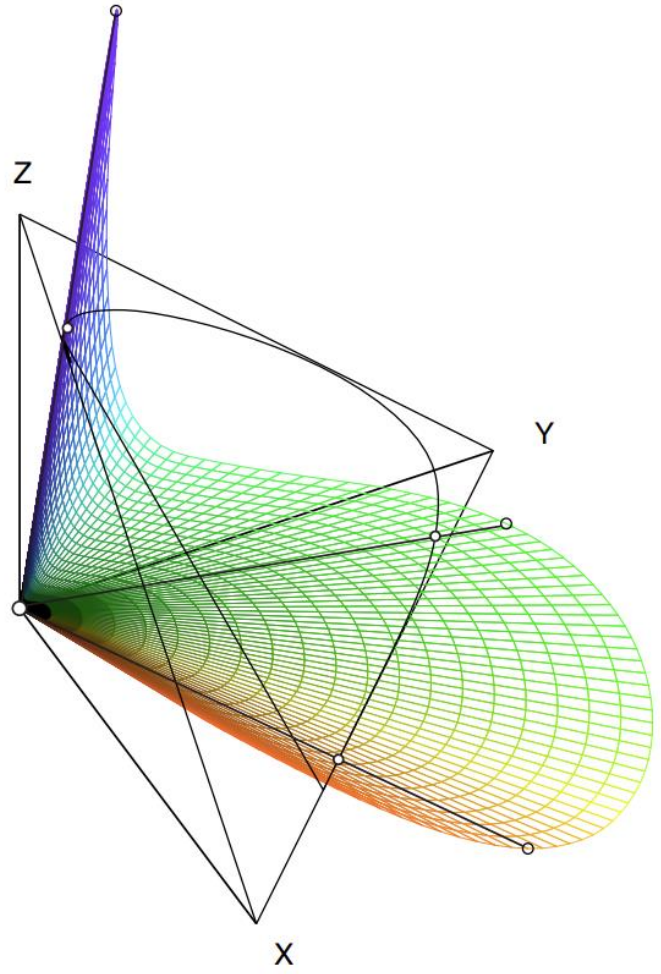


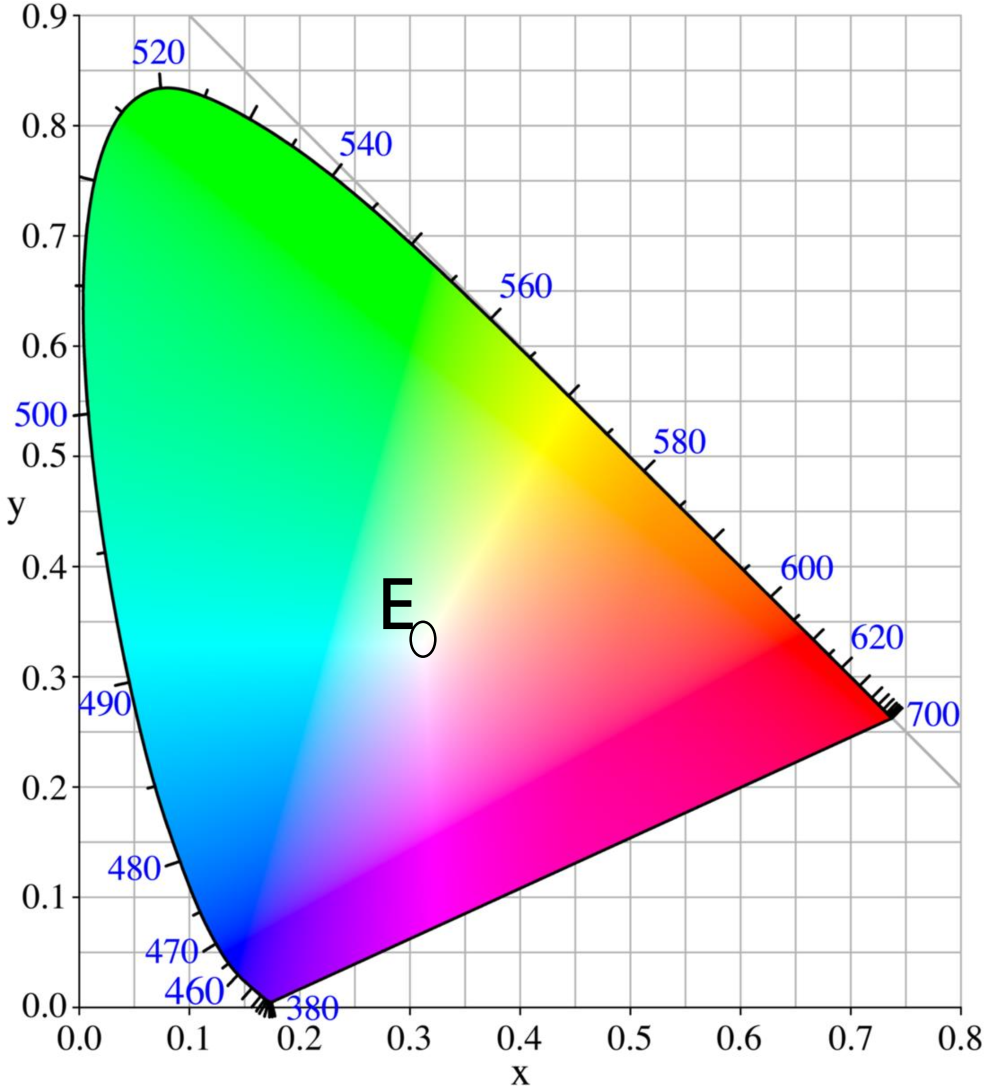
Image: Gernot Hoffmann

Luminance-chromaticity space (CIE xyY)

- CIE XYZ describes a color in terms of linear combination of three primaries (XYZ).
- Sometimes it is useful to discuss color in terms of luminance (perceived brightness) and chromaticity (we can think of as the hue-saturation combined).
- CIE xyY space is used for this purpose.

CIE Yxy chromaticity diagram

Point "E" represents where X=Y=Z have equal energy (X=0.33, Y=0.33, Z=0.33)



In the 1930s, CIE had a bad habit of over using the variables X, Y. Note that x, y are chromaticity coordinates, $\bar{x}, \bar{y}$ (with the bar above) are the matching functions, and X, Y are the imaginary SPDs of CIE XYZ.

Usefulness of CIE 1931 XYZ

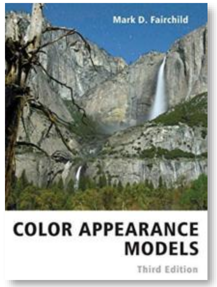
- CIE XYZ space is a “device independent” space – the XYZ values are not specific to any device.
- Electronic devices (e.g. cameras, flatbed, scanners, printers, displays) can compute mappings of their device specific values to the corresponding CIE XYZ values.
- This provides a canonical space to match between devices (at least in theory).

A caution on CIE xy chromaticity

From Mark D. Fairchild's book: "Color Appearance Models"

"The use of chromaticity diagrams should be avoided in most circumstances, particularly when the phenomena being investigated are highly dependent on the three-dimensional nature of color. For example, the display and comparison of the color gamuts of imaging devices in chromaticity diagrams is misleading to the point of being almost completely erroneous."

Fairchild



Fast forward 90+ years

- CIE 1931 XYZ, CIE 1931 xyY (2-degree standard observer) color spaces have stood the test of time.
- Many other studies have followed (most notably - CIE 1965 XYZ 10-degree standard observer), . . .
- But in the literature (and in this tutorial) you'll find CIE 1931 XYZ color space remains the preferred standard.

What is perhaps most amazing?

- 90+ years of CIE XYZ, and it is all based on the experiments by Guild and Wright's "standard observers."

- How many standard observers were used? 100, 500, 1000?

A standard observer

CIE XYZ is based on 17 (male) standard observers

10 by Wright, 7 by Guild

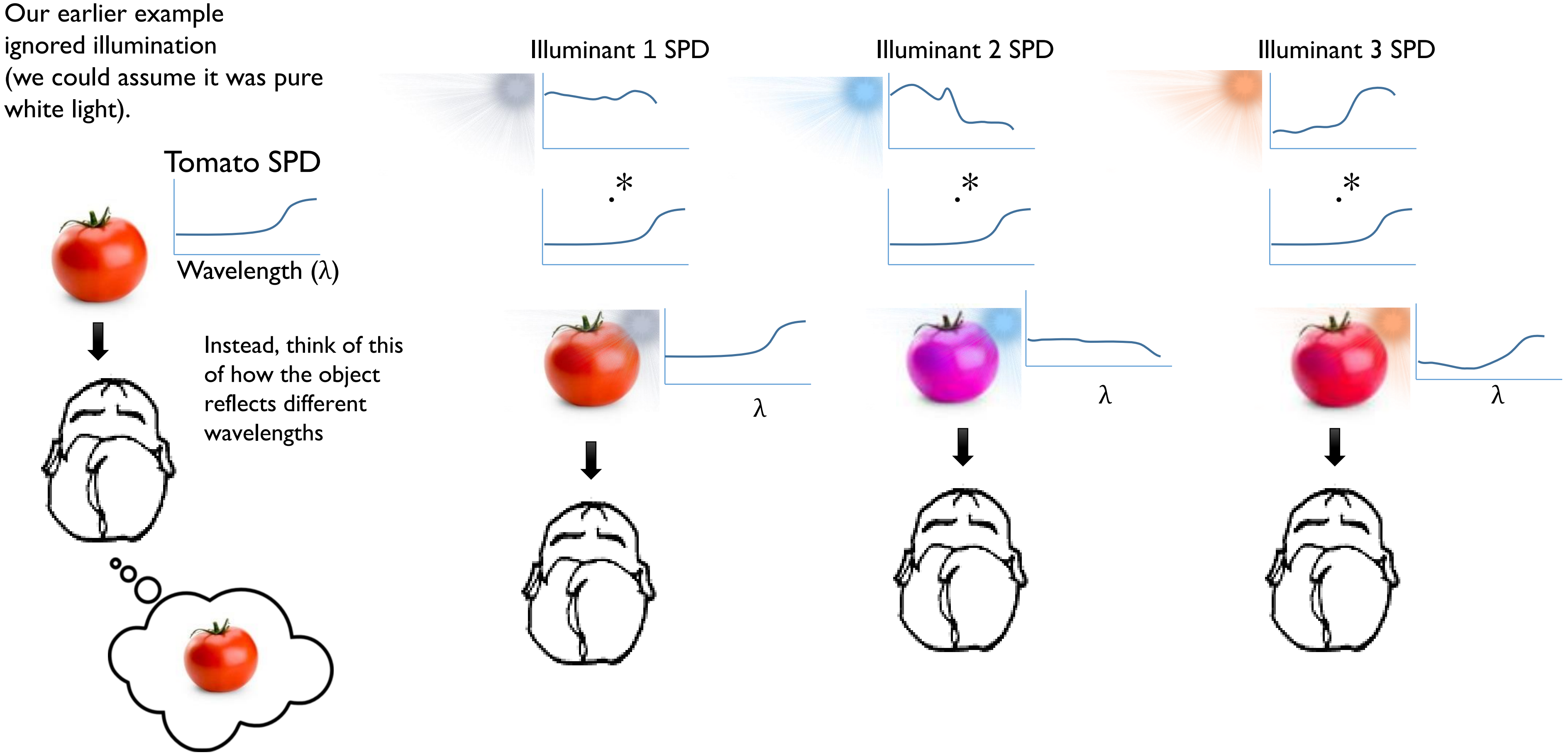
“The Standard Observers”

Can we talk about cameras now?

Sorry, not yet . . .

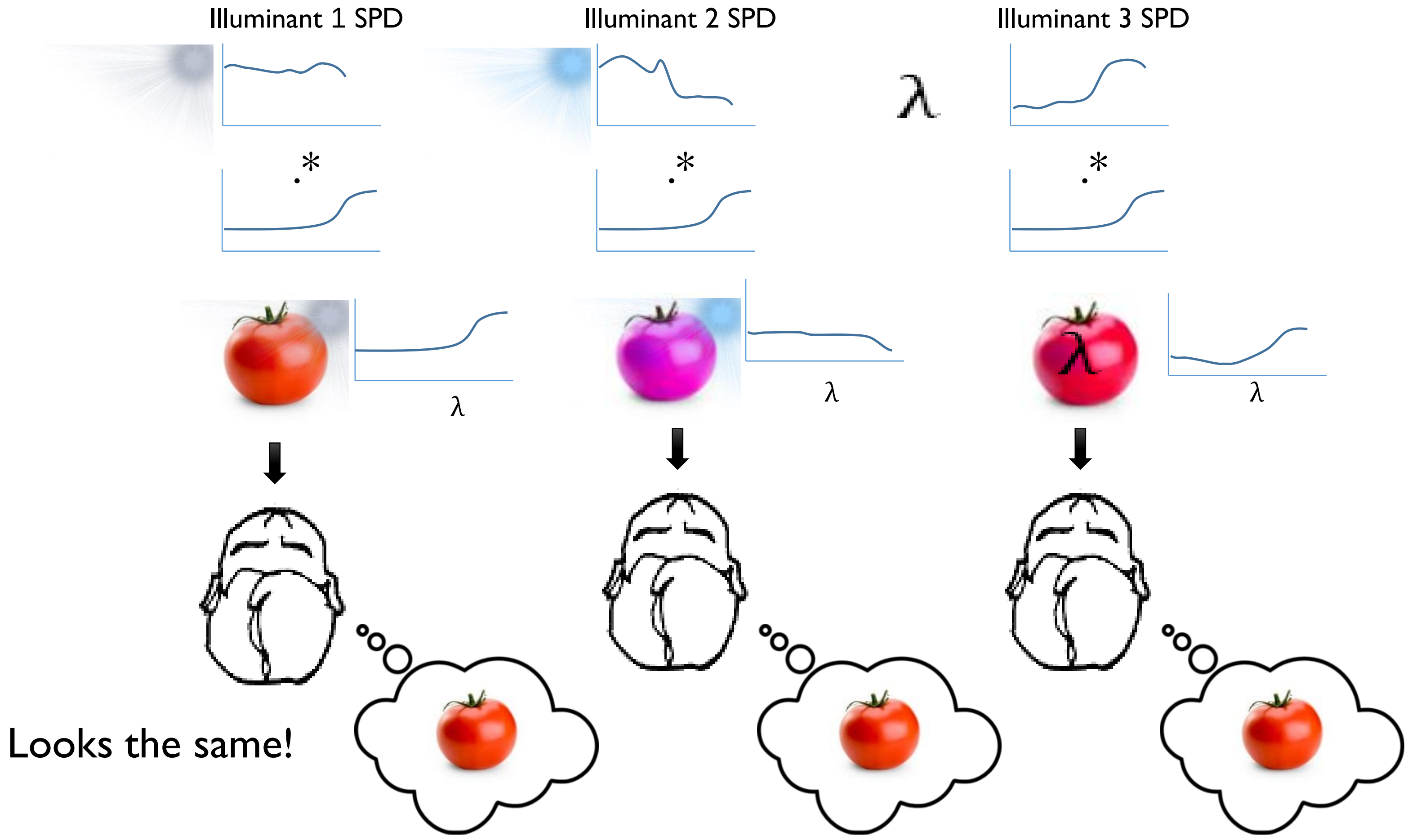
An object's SPD

In the real world, most objects do not emit an SPD, instead, they reflect an SPD.
As a result, an object's SPD depends on the environmental illumination.



Color constancy

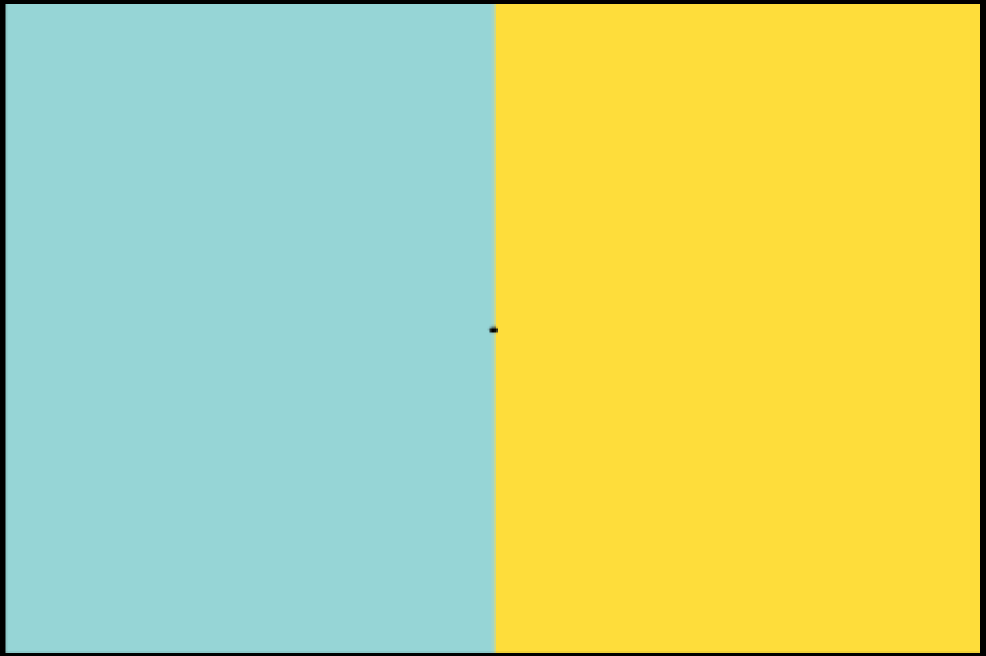
Our visual system has an amazing ability to compensate for environmental illumination such that objects are perceived as the same color.



Chromatic adaptation example

Example from Andrew Stockman (UCL)

Chromatic adaptation example



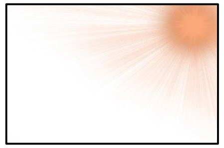
Color constancy/chromatic adaptation

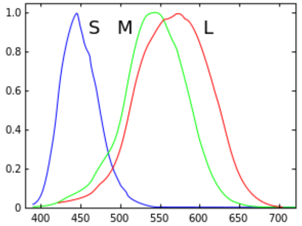
- Color constancy (*chromatic adaptation*) is the ability of the human visual system to adapt to scene illumination.
- This ability is not perfect, but it works fairly well.
- **Image sensors do not have this ability! We will discuss this in part 2 . . this is related to the camera's white-balance module.**

Color constancy (at its simplest)

Johannes von Kries

- The *Von Kries* transform
- Compensate for L/M/S channel corresponding to the L, M, S response to scene illumination.





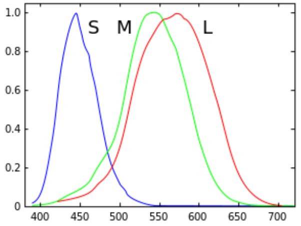
$[L_{illum}, M_{illum}, S_{illum}]$

L/M/S response to the light source.

"Corrected colors"

$[L'_x, M'_X, S'_x]$

$$\begin{bmatrix} L'_x \\ M'_x \\ S'x \end{bmatrix} = \begin{bmatrix} 1/L_{illum} & 0 & 0 \\ 0 & 1/M_{illum} & 0 \\ 0 & 0 & 1/S_{illum} \end{bmatrix} \begin{bmatrix} L_x \\ M_x \\ S_x \end{bmatrix}$$



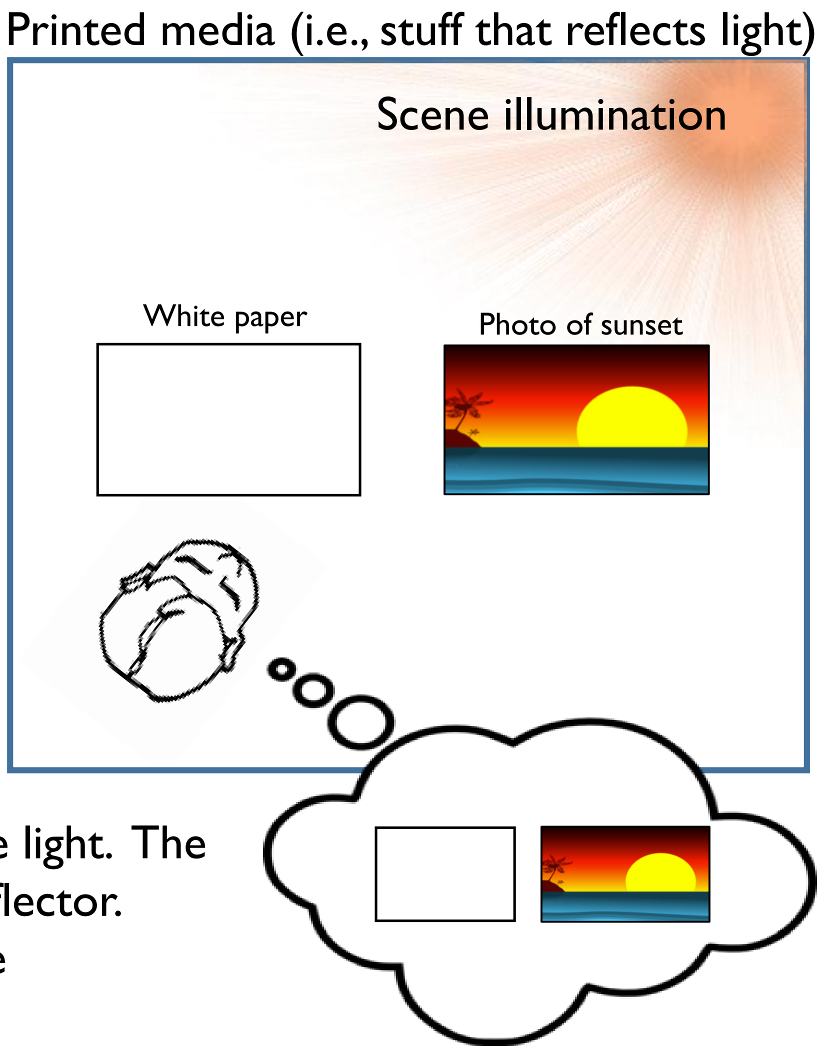
$[L_x, M_x, S_x]$

Long/medium/short cone response with illumination "corrected."

Divide out long/medium/short cone response to the scene's illuminant.

Long/medium/short cone response to scene point x under some illuminant.

Color constancy for printed media

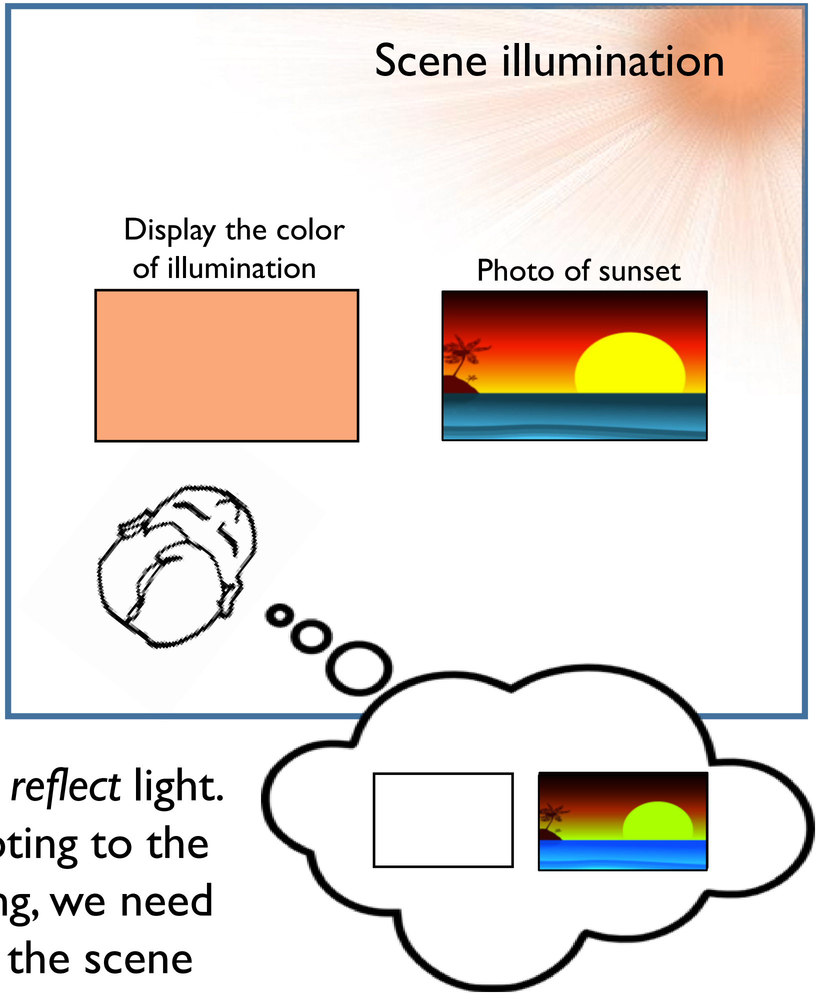


The white paper reflects the light. The paper is almost a perfect reflector. Since we are adapting to the environmental light source, the paper appears white.

The photo also reflects the light, so the colors are perceived correctly.

Color constancy for emissive media

Emissive media (e.g., monitor/tablet, smartphone screen)



The display does *not reflect* light. Because we are adapting to the environmental lighting, we need the display to match the scene illumination. If we match the illumination, the display will appear "white."

The displayed image colors will appear differently than intended, since we are adapting to the environment illumination.

Implications of the previous slides

- ***Color is intimately connected to scene illumination.***

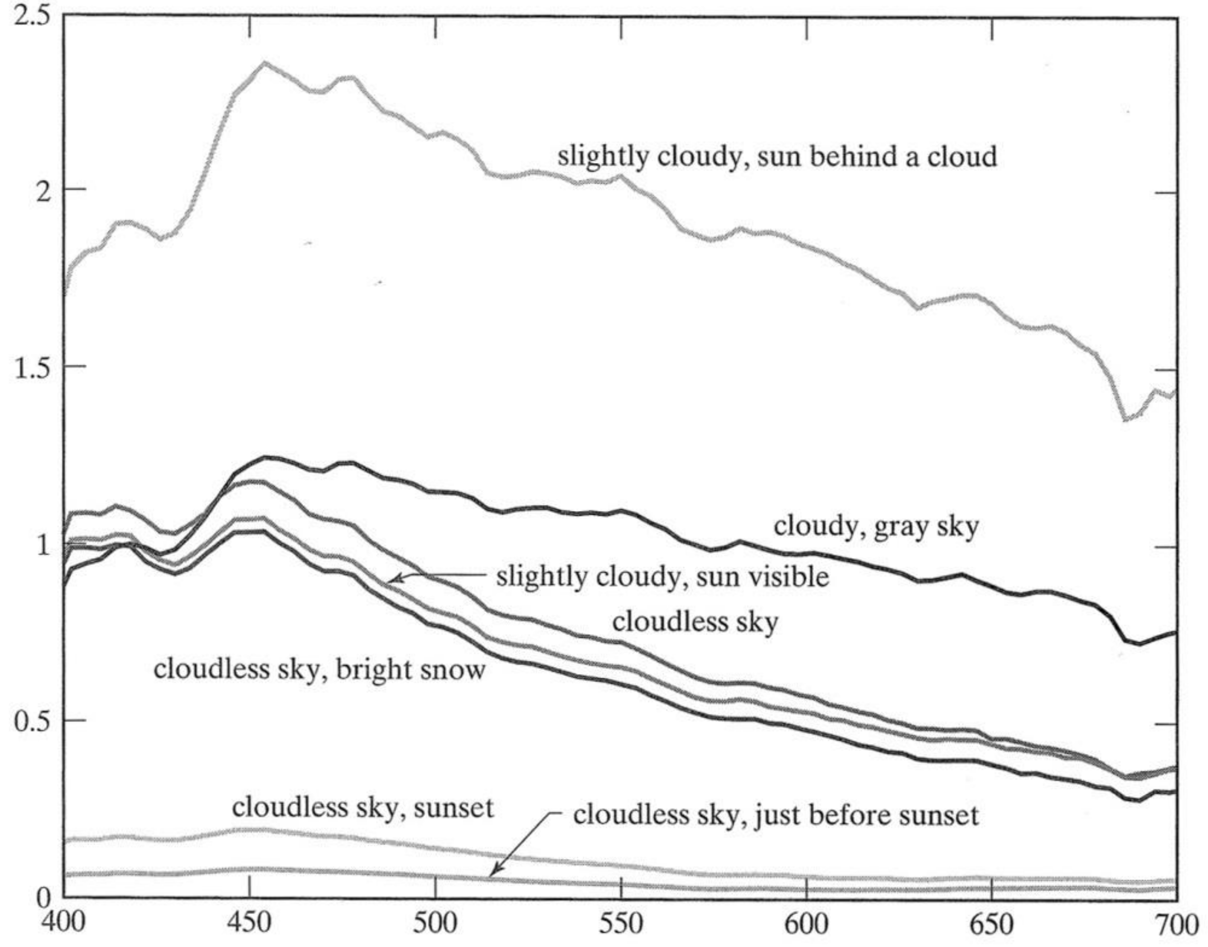
- Even for emissive displays, we have to consider (or make assumptions) about the illumination in the viewing environment of the display.

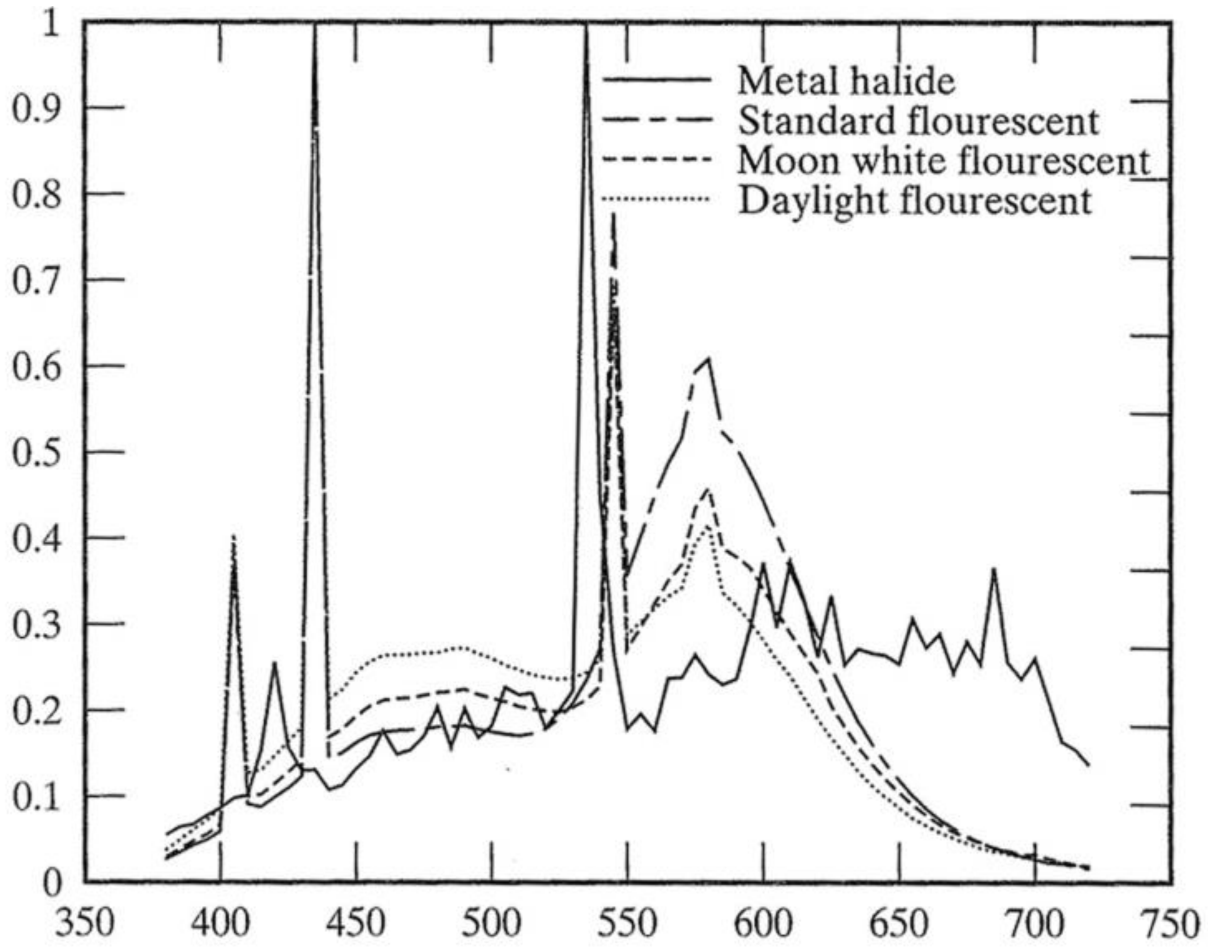
- Keep this in mind because it will play a role when we define color spaces used to encode our images.

Understanding color temperature

- In the photography and display communities, an illumination's "color" is described using a *correlated color temperature* (CCT).
- White balance on cameras also often uses color temperature to describe illumination.
- This is an excellent example of where metamers are used.
  - Recall – a metamer is when two different SPDs appear visually the same color.

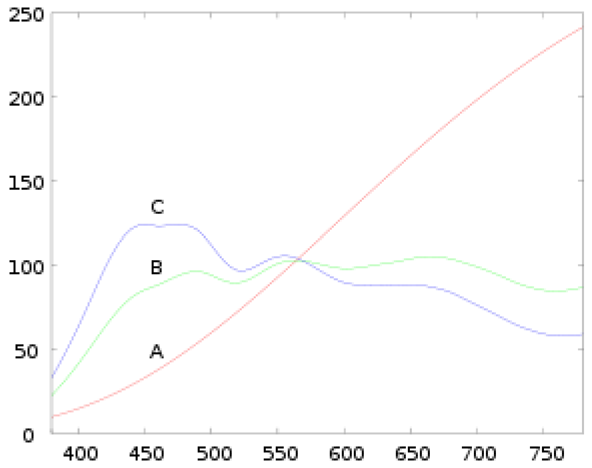
SPDs of common illuminations



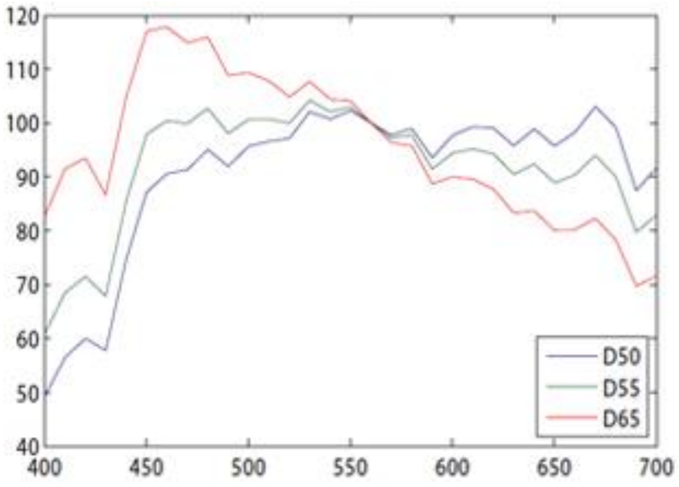


Figures from Ponce and Forsyth

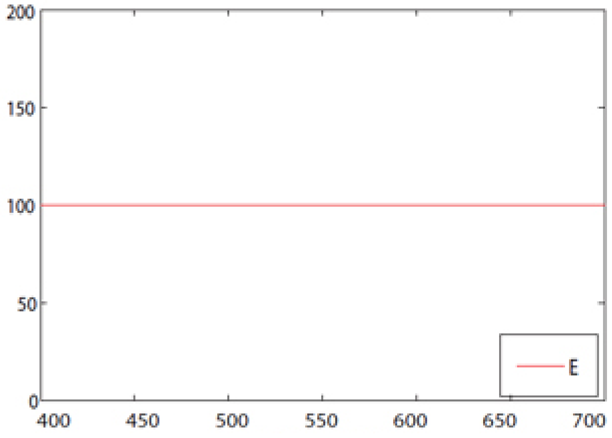
CIE standard illuminants



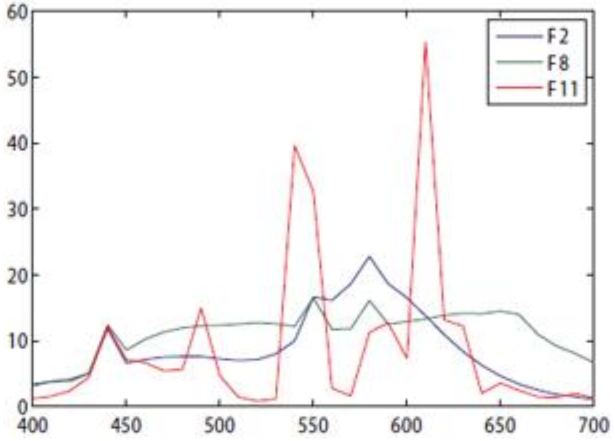
SPDs for CIE standard illuminant A, B, C



SPDs for CIE standard illuminant D50, D55, D65



SPDs for CIE standard illuminant E

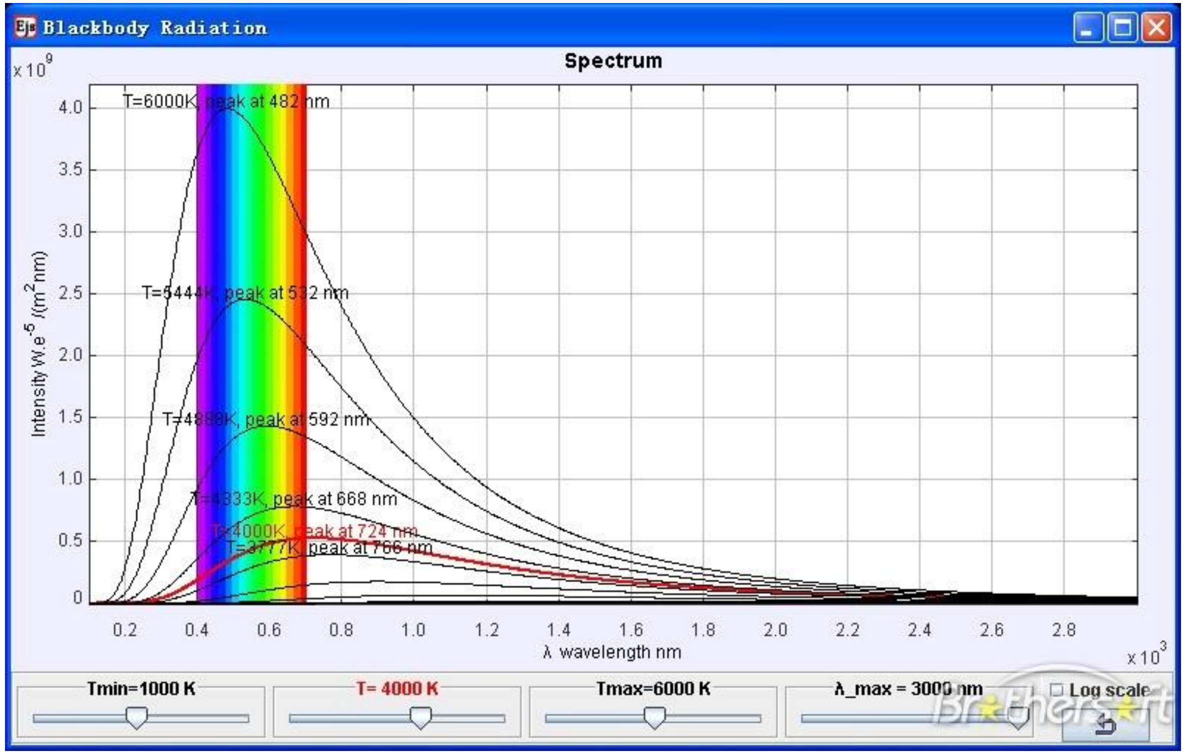


SPDs for CIE standard illuminants F2, F8, F11

D, E, and F series images from http://www.image-engineering.de

Color temperature

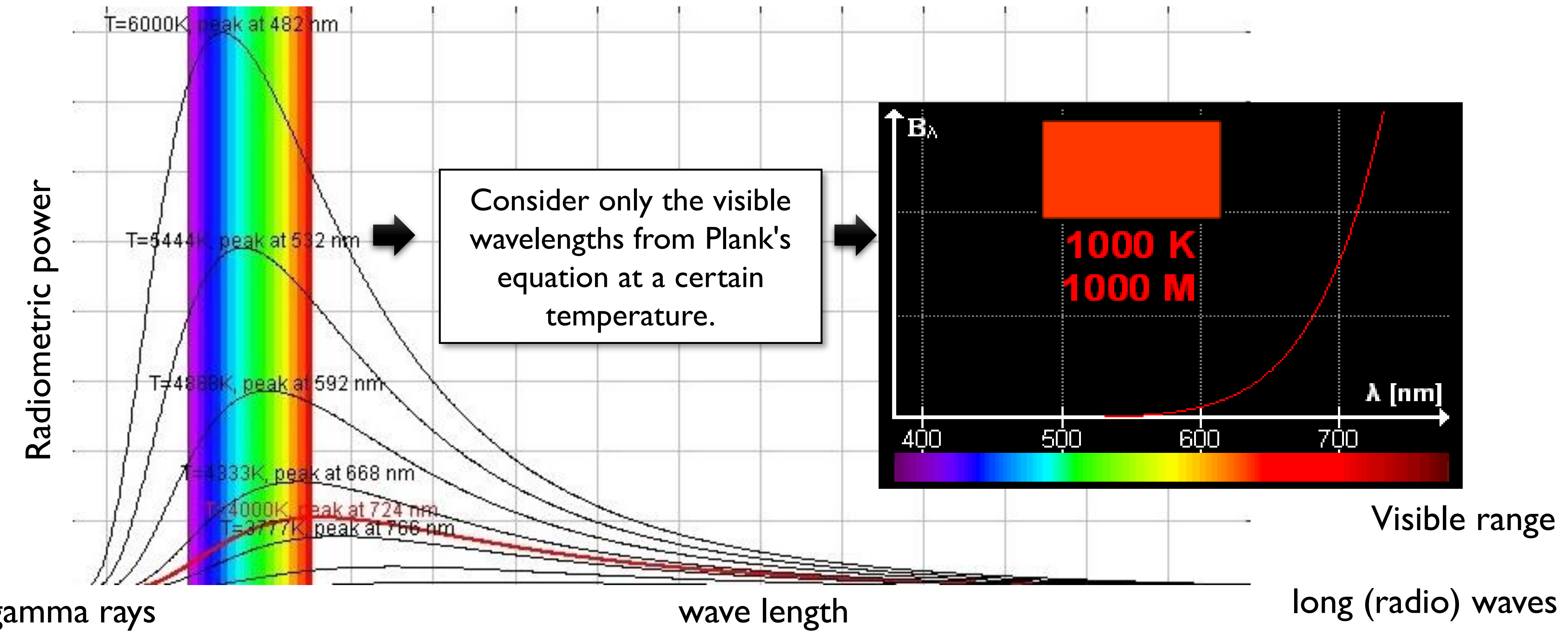
- As mentioned, illuminants are often described by their "color temperature."
- This mapping is based on theoretical *blackbody radiators* that produce SPDs for a given temperature expressed in Kelvin (K).
- We map light sources (both real and synthetic) to their closest color temperature.



$$B_\lambda(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k_B T}} - 1}$$

Plank's law
Spectral density of electromagnetic radiation emitted by a blackbody radiator at a given temperature T.

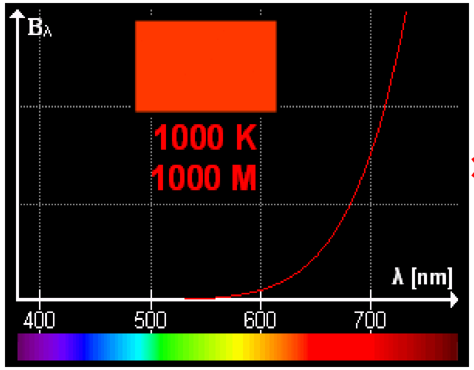
Visible range of a black body radiator SPD

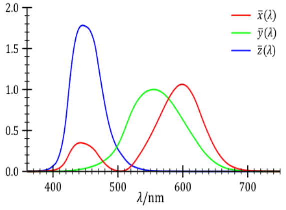


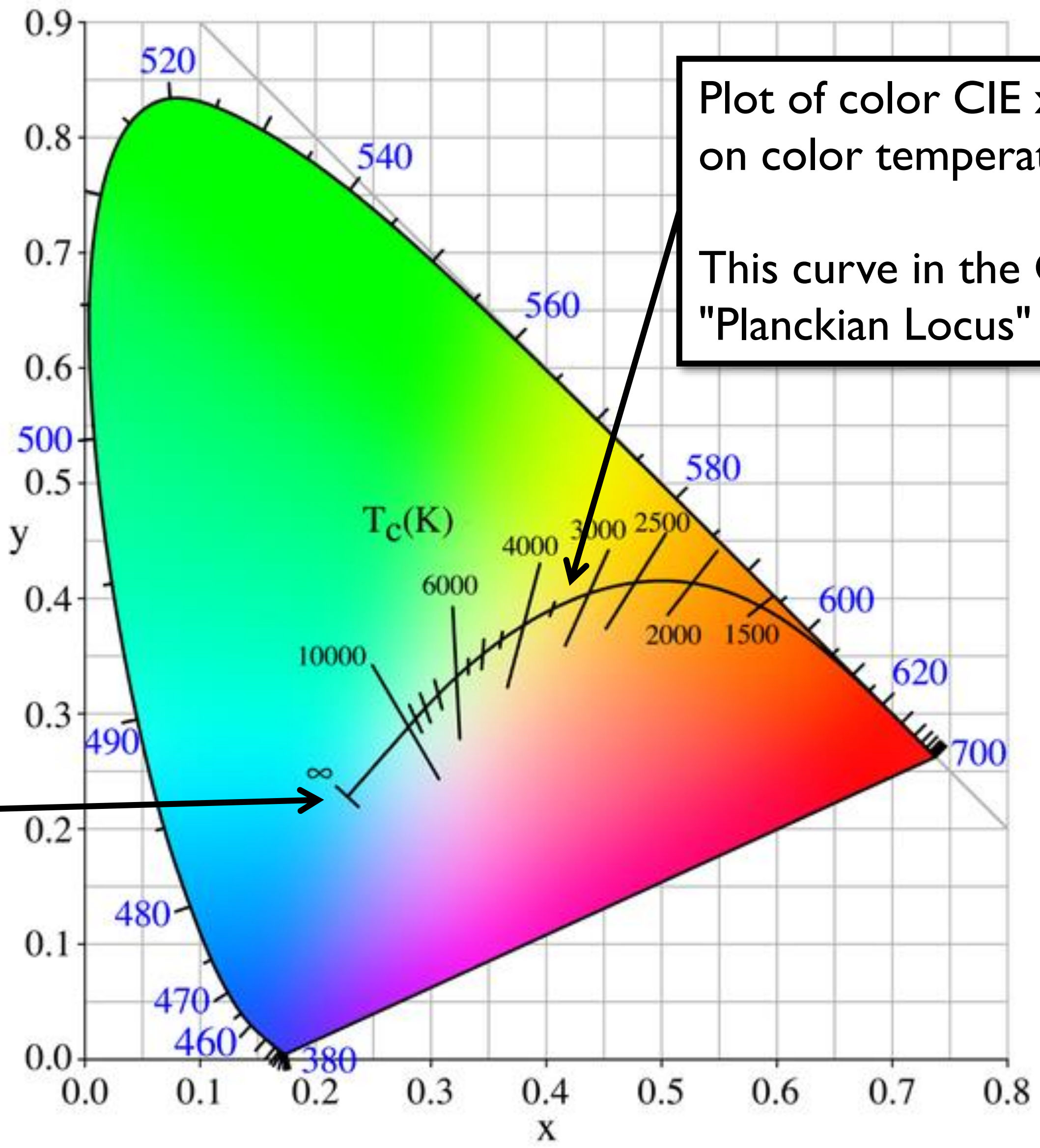
Black body radiator SPD for different color temperatures

Animation credit: Dariusz Kowalczyk

Plot visible SPDs in CIE xy chromaticity





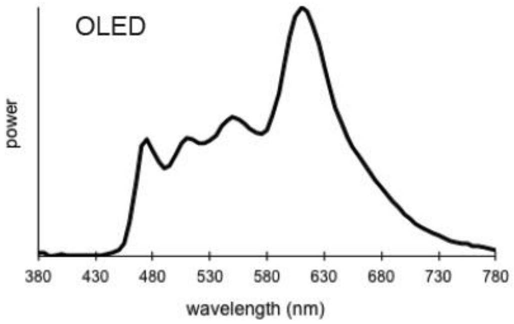


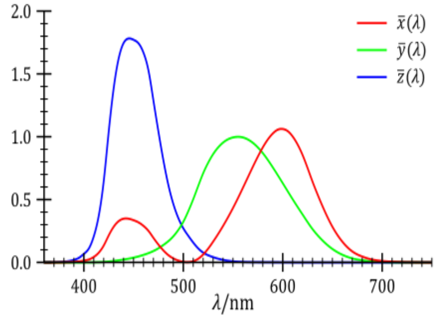
Plot of color CIE xy locations of SPDs based on color temperature.

This curve in the CIE xy plot of the "Planckian Locus" of color temperatures.

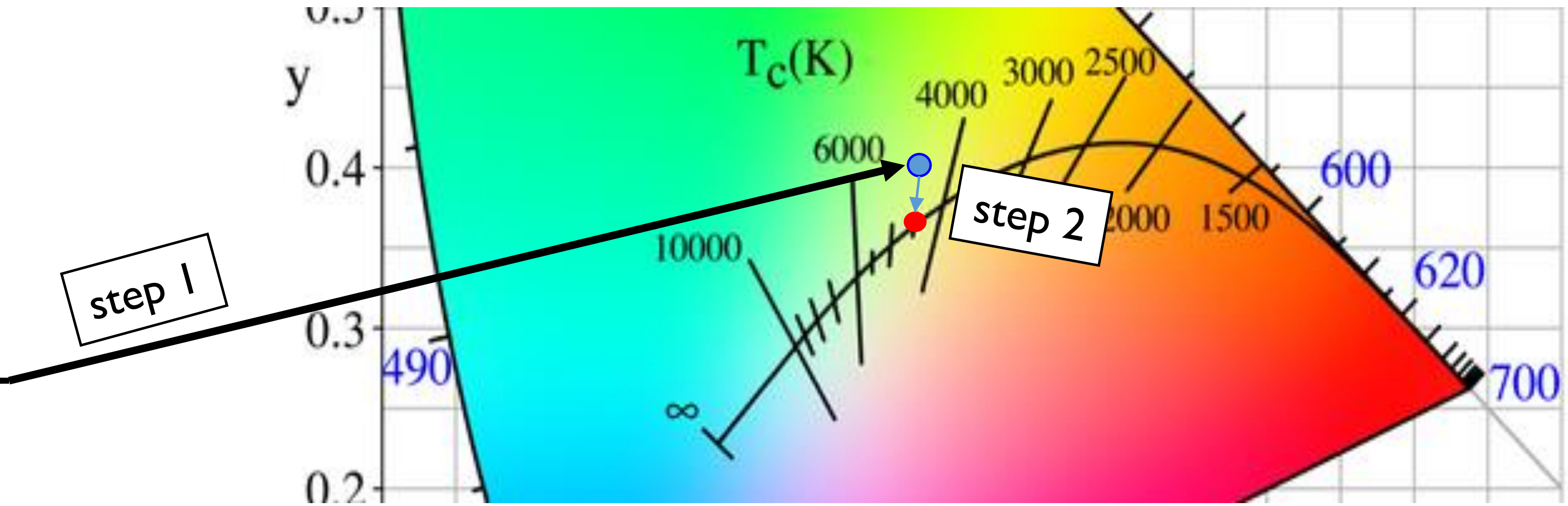
Color temperature of an SPD example

SPD of a light source





CIE 1931 mapping functions

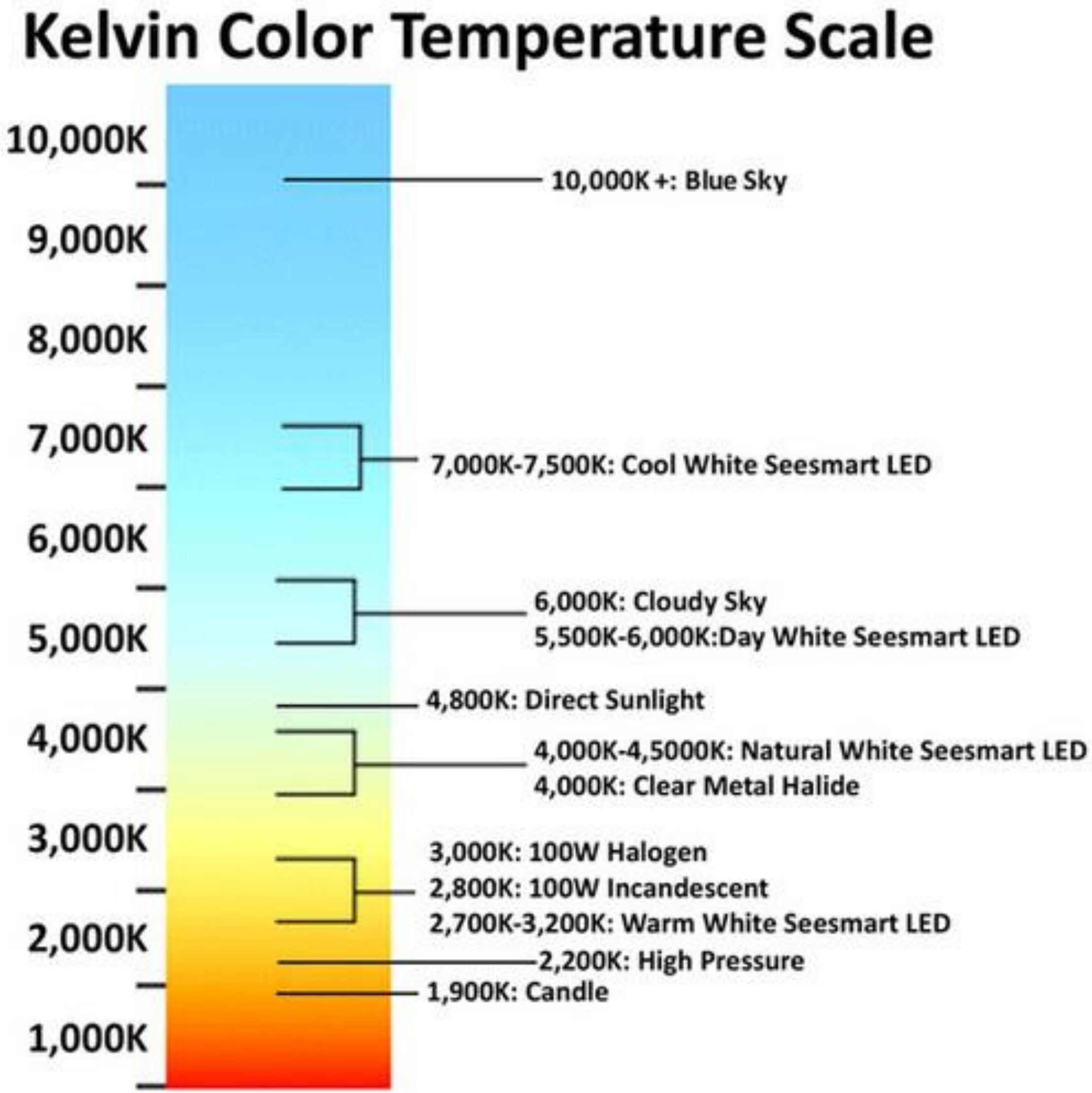


(1) Find the light sources SPD mapping to CIE XYZ using the CIE 1931 mapping functions.
(2) Project the CIE xyY value to the Planckian locus line.

Where the projection falls is the Correlated Color Temperature (CCT) of this light source. So, in this example, the OLED light source is roughly 4500K.

While we often say "color temperature", we should say "correlated color temperature." The concept is *not always* related to the physical temperature of the light source, but its *correlation* with the black body radiator's color temperature.

Color temperature



Typical description of color temperature used in photography & lighting sources.

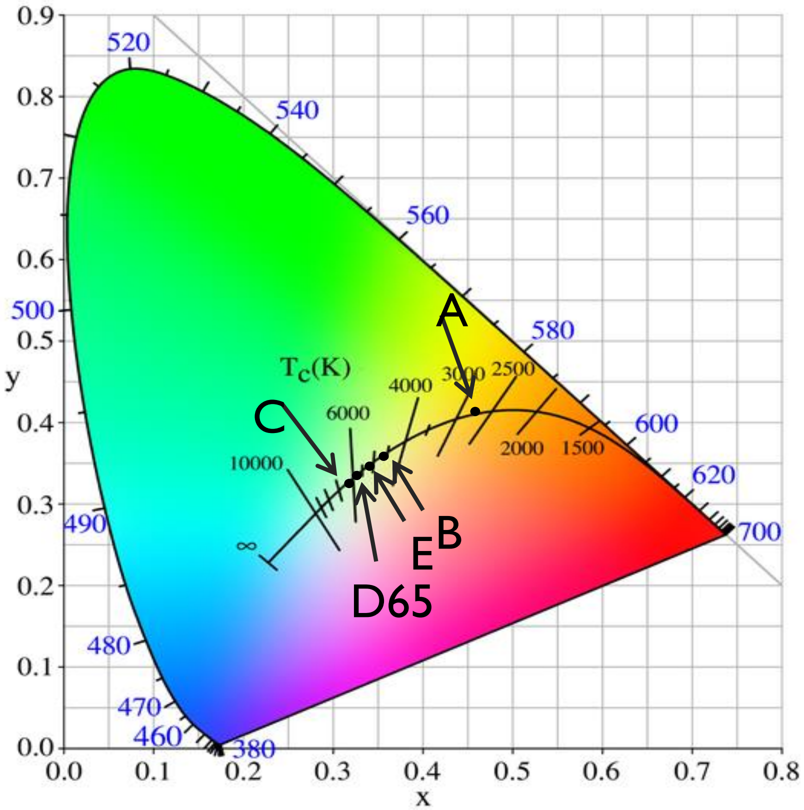
From B&H Photo

Lighting industry uses color temperature

Usage of correlated color temperature in these ads relate to the perceived color of the bulb's light. The heat output of a typical LED bulb is between 60C-100C (~333-373K).

White point

- A white point is a color defined in CIE xyY that we want to be considered "white" (or achromatic/neutral).
- This is essentially an illuminant's SPD in terms of CIE XYZ/CIE xyY
  - Think of it as CIE Yxy value of a white piece of paper under some illumination.



CIE Illuminants

A, B, C, D65, E in terms of CIE xy

CIE	x ,	y
A	0.44757 ,	0.40745
B	0.34842 ,	0.35161
C	0.31006 ,	0.31616
D65	0.31271 ,	0.32902
E	0.33333 ,	0.33333

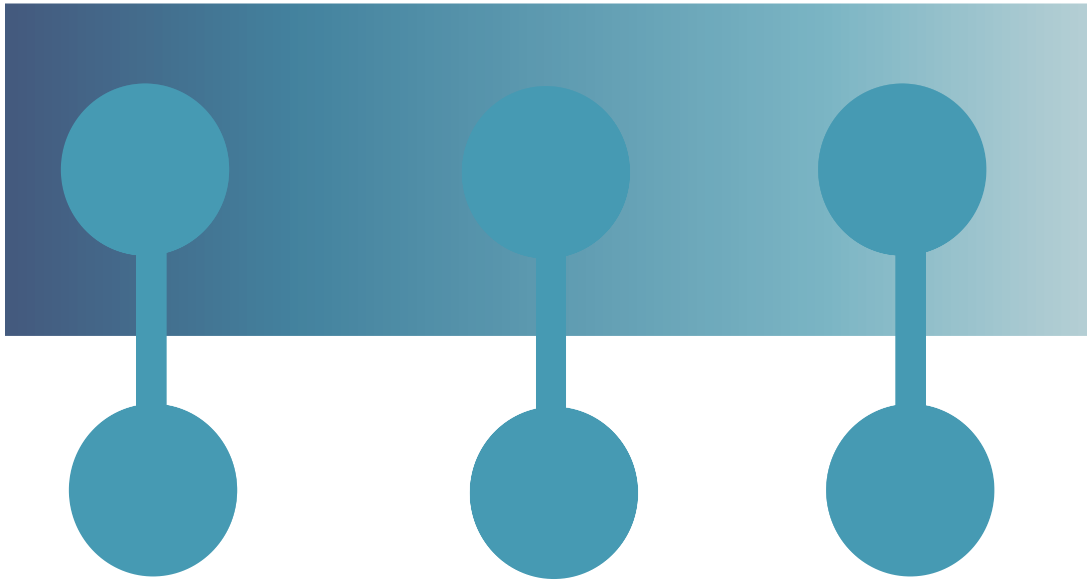
Quick summary on color constancy

- Color constancy is our ability to adapt to illumination in the scene.

- Correlated Color Temperature (CCT) — or just color temperature — is a system used to describe scene illumination.

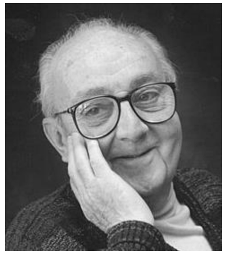
- **Note:** we *must factor in the scene illumination when capturing and displaying color images.*

Color adaptation is not perfect



Mark Fairchild

"True color constancy, almost never. Inconstancy, nearly 100% of the time."



George Box (Statistician pioneer)

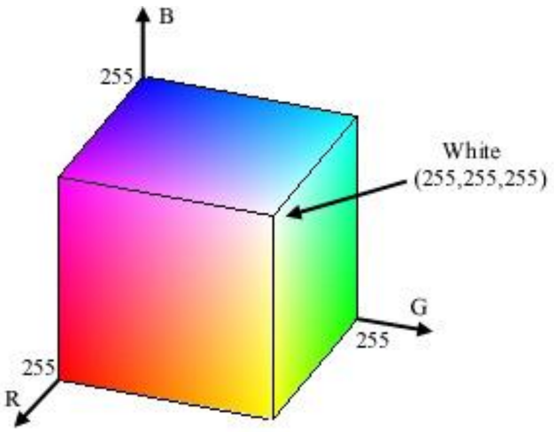
"Remember that all [mathematical] models are wrong; the practical question is how wrong do they have to be to not be useful."

Now we are finally done with color?

Almost . . .

CIE XYZ and RGB

- While CIE XYZ is a canonical color space, images/devices rarely work directly with XYZ.
- RGB primaries dominate the industry, this is because we can produce RGB light sources (LEDs, phosphorus for CRT monitors, filters, etc)
- We are all familiar with the RGB color cube.
- But is the color cube a color space?

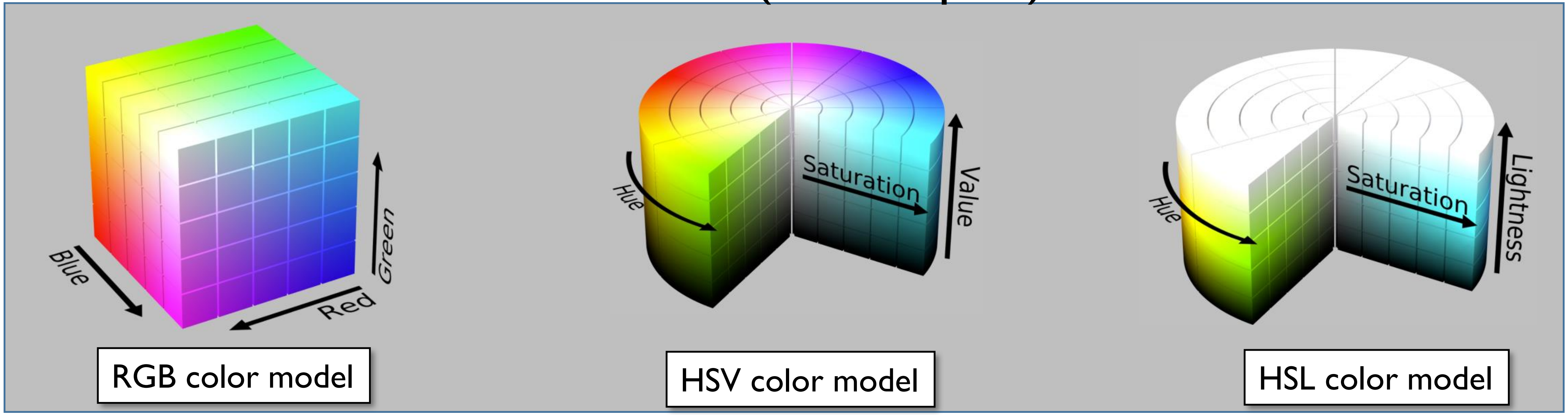


By now, you should realize that "red", "green", and "blue" have no quantitative meaning as words. We need to know their corresponding SPDs or CIE XYZ values.

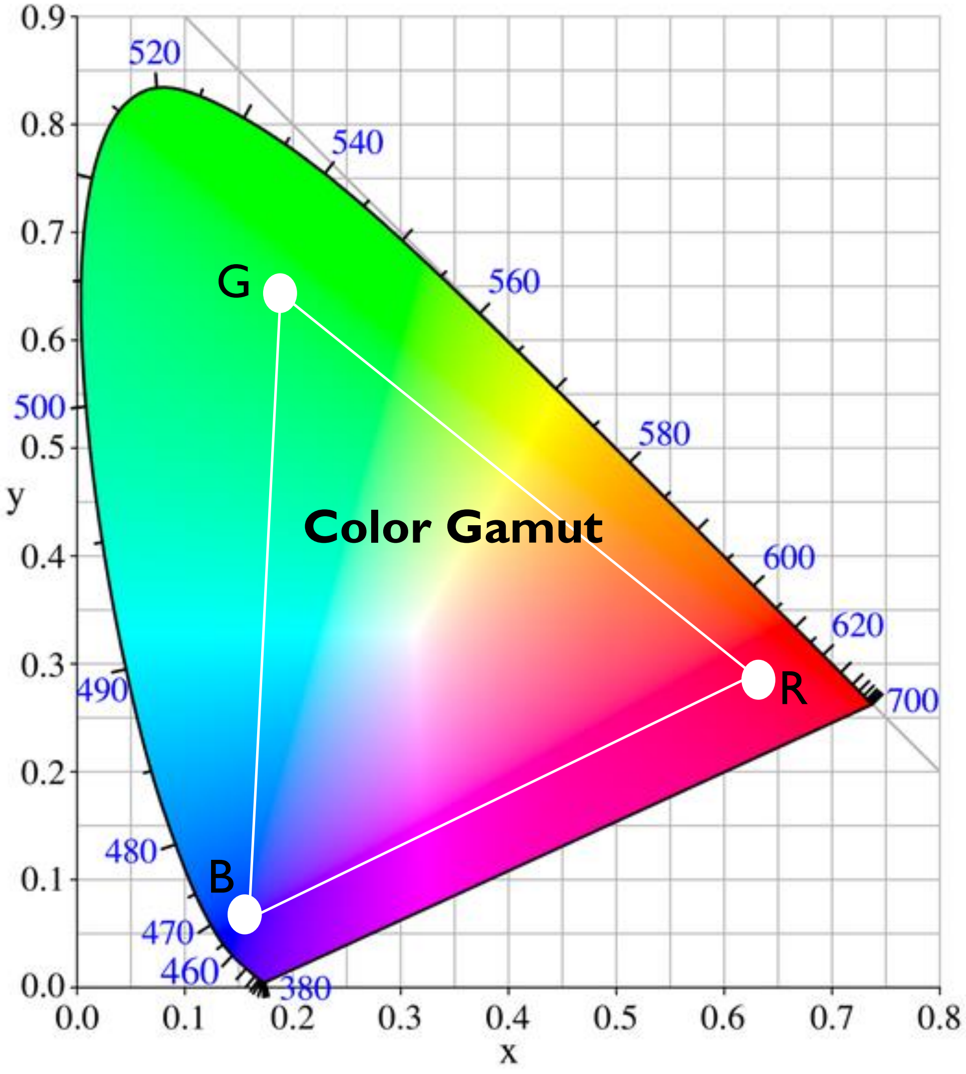
Color model versus color space

- **A color model** is a mathematical system for describing a color as a tuple of numbers (RGB, HSV, HSL, more…)
- **A color space** is a specific range of colors *within a color model*. The range of color (gamut) can be expressed in CIE XYZ. Color spaces typically also define the viewing environment and, therefore, the "white point" of the space.

Color models (not color spaces)

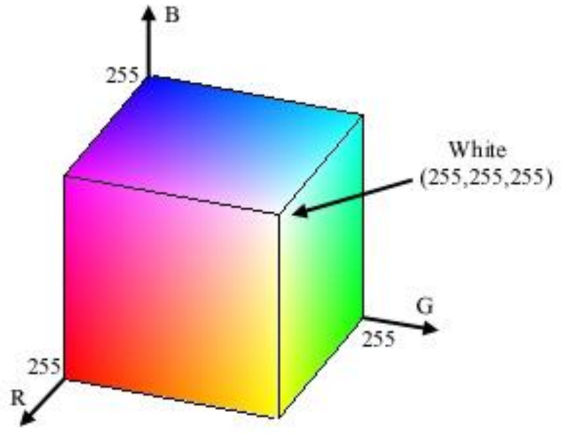


Defining a color space with specific RGB values

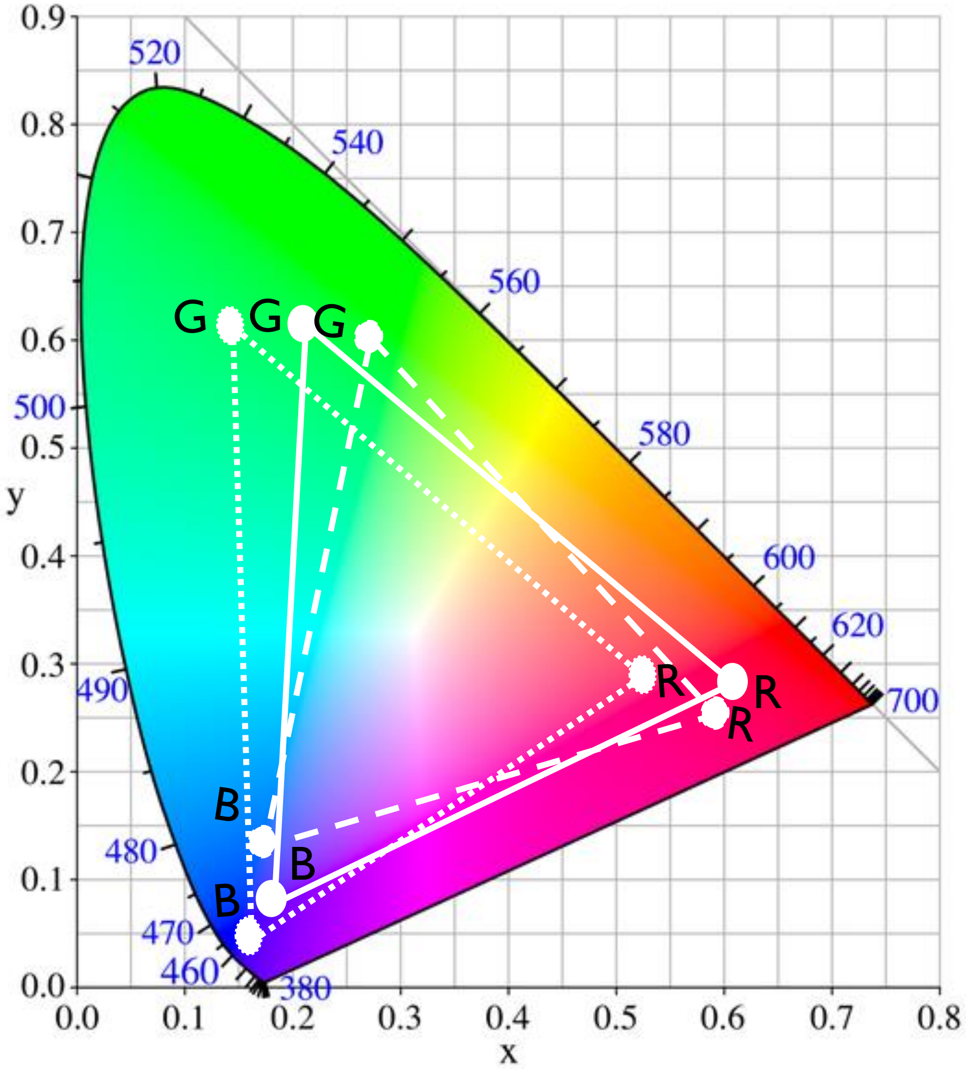


The RGB values span a subspace, of CIE-XYZ to define the devices gamut.

We need to define our RGB values.



Problem with just a color model. .

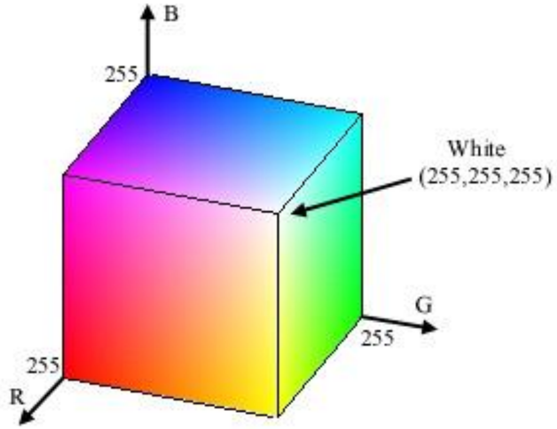


RGB 1 ——

RGB 2 ·············

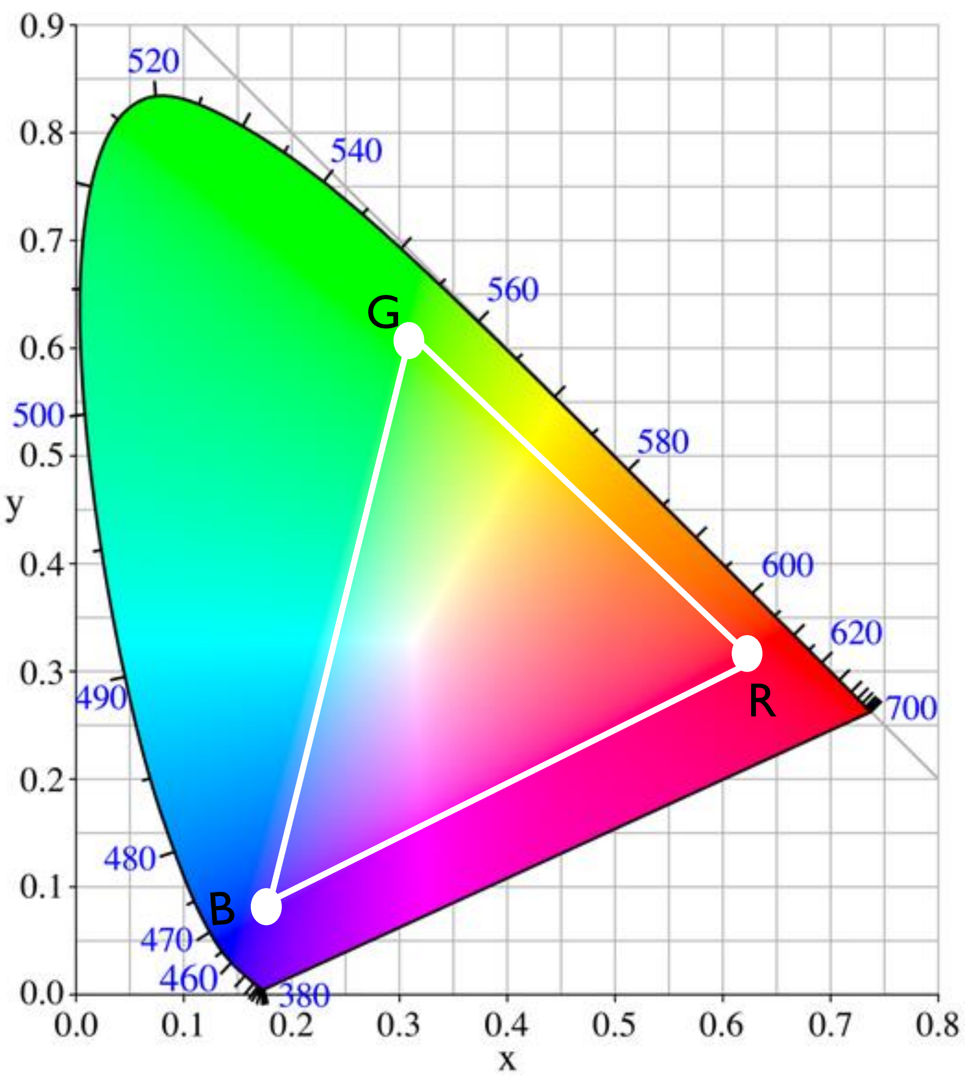
RGB 3 - - - -

Which RGB primaries are the right ones?



RGB values must be specified.
If not, this is a **huge** problem for color reproduction from one device to the next.

Standard RGB (sRGB) – Rec. 709



In 1996, Microsoft and HP defined a set of "standard" RGB primaries.

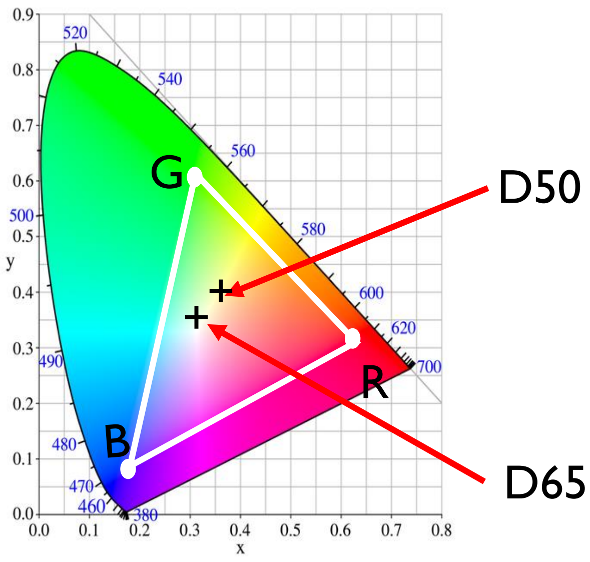
R=CIE xyY (0.64, 0.33, 0.2126)
G=CIE xyY (0.30, 0.60, 0.7153)
B=CIE xyY (0.15, 0.06, 0.0721)

This was considered an RGB space achievable by most devices at the time.

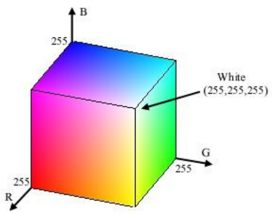
The white point was set to the D65 illuminant**. This is an important to note.** It means sRGB has built in the assumed viewing condition (6500K daylight).

sRGB's white point

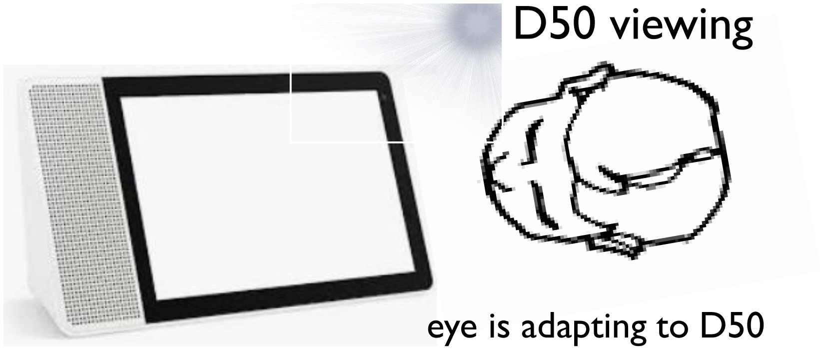
- Color spaces intended for display (called display-referred or output-referred) define a white-point.
- Remember to match the assumed illumination in the viewing environment
  - The "white" of sRGB (i.e., [1,1,1]) is displayed at D65



CIE XYZ

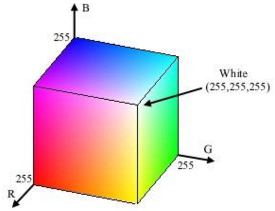


sRGB (D50 whitepoint)

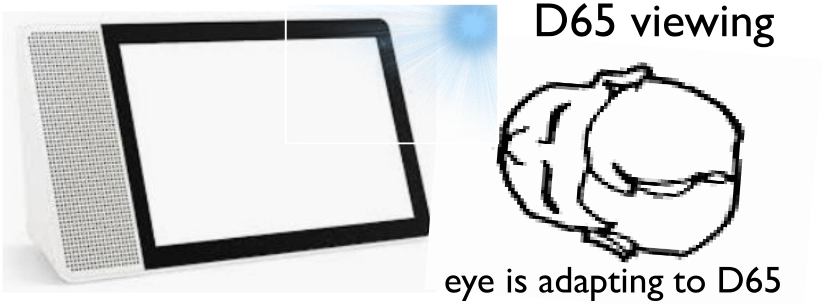


D50 viewing

eye is adapting to D50 environment light.



sRGB (D65 whitepoint)



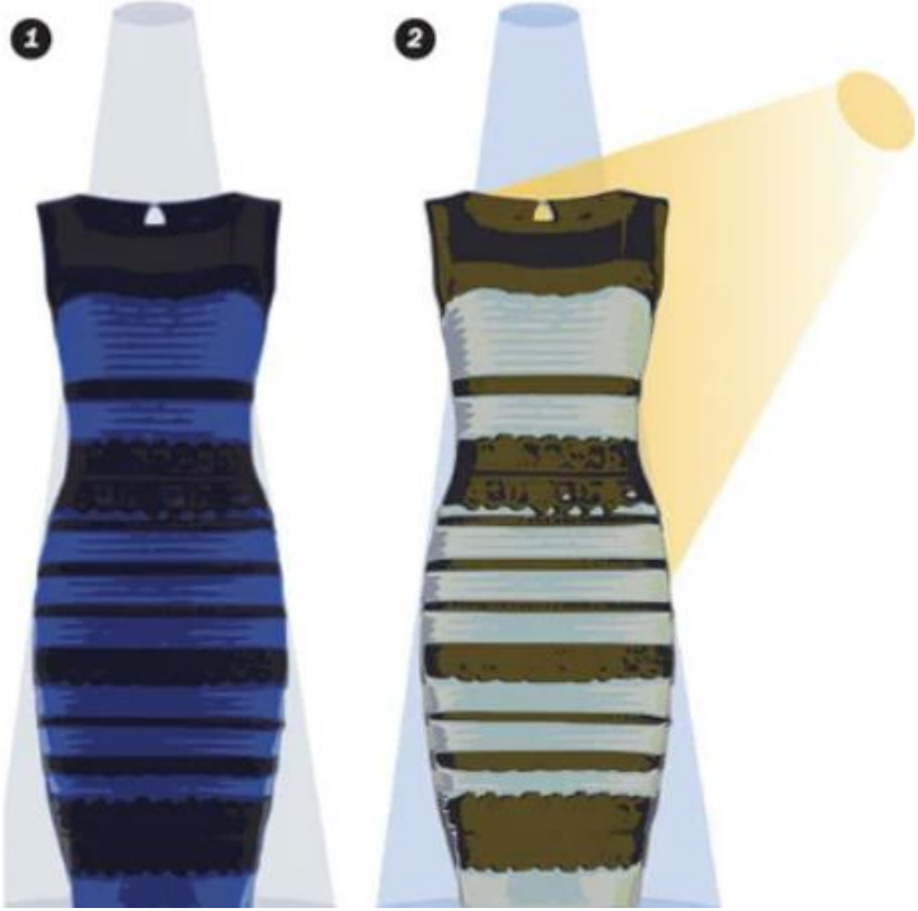
D65 viewing

eye is adapting to D65 environment light.

The positions of the white-point locations are exaggerated here.

Assumed viewing illumination is important

Remember "the dress"?



Viewing illumination

Image: *Scientific America* article explaining how viewing environment lighting impacted our perception of the color.

CIE XYZ to sRGB conversion

Matrix conversion:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 3.2404542 & -1.5371385 & 0.4985314 \\ -0.9692660 & 1.8760108 & 0.0415560 \\ 0.0556434 & -0.2040259 & 1.0572252 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

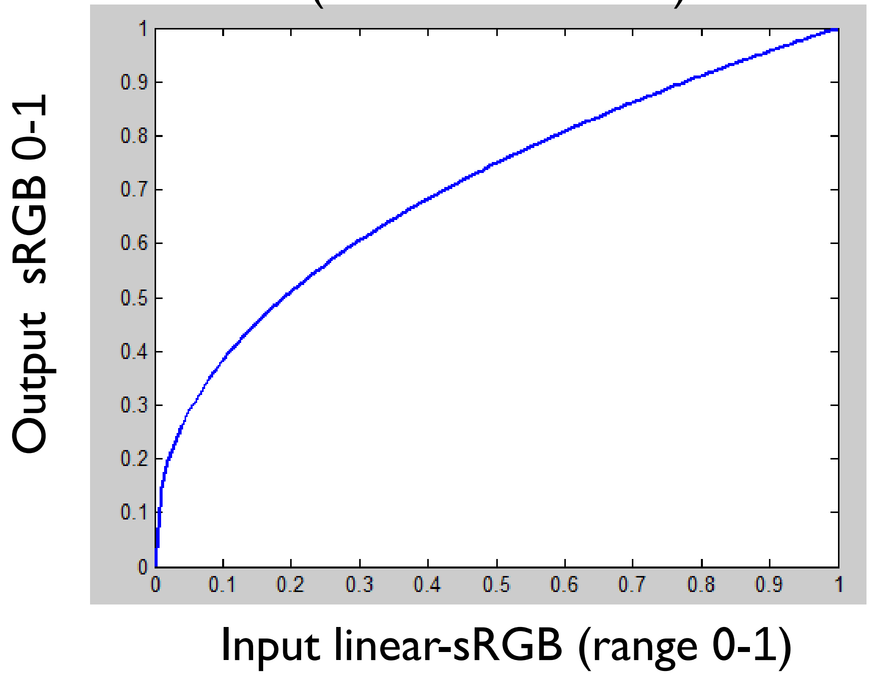
Linearized sRGB (D65)

CIE XYZ

- D65 is set as the white-point.
- This is the linear sRGB space.
- sRGB also specifies a gamma correction of the values (next slide)
- The CIE refers to this as the Recommendation 709 color space – or Rec.709.

sRGB gamma curve

(1D transfer curve)



This is a close approximation of the actual sRGB gamma

The actual formula is a bit complicated, but effectively this is gamma ($I' = 255 \times I^{(1/2.2)}$) where I' is the output intensity and I is the linear sRGB ranged 0-1, with a small linear transfer for linearized sRGB values close to 0 (not shown in this plot). This is known as "perceptual encoding" and is intended to allocate more bits based on our nonlinear response to radiant power.

Stevens' power law

- Physical stimulus vs. perceptual sensation
- Stevens' Power Law

$$S = kI^{a}$$

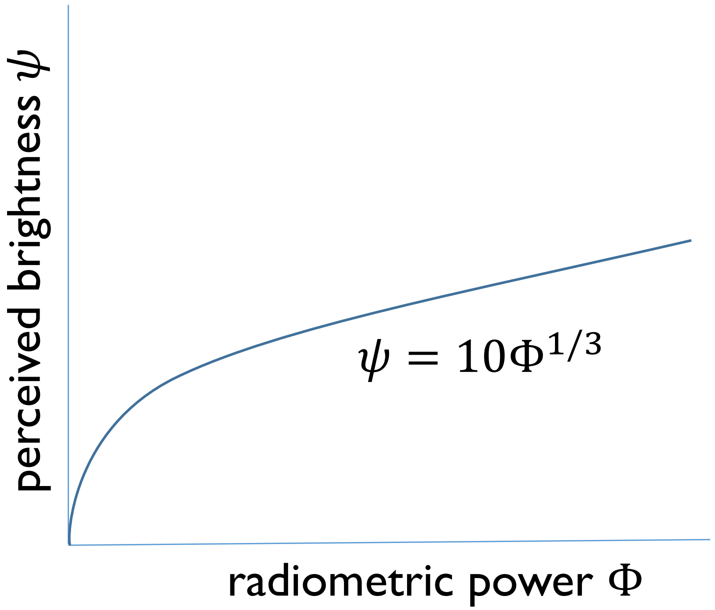
Human sensation

Constant

Stimulus intensity

power exponent

Dr. Stanley Stevens showed that most human sensations follow a power-law relationship between stimuli and sensation.

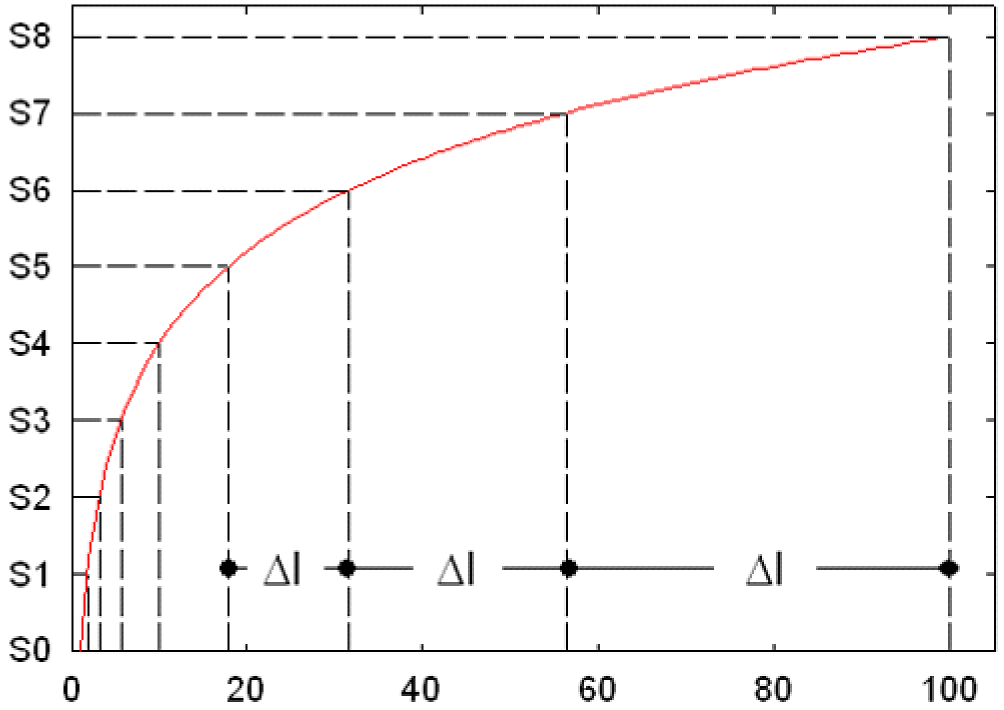


Stevens' model stated that human perception to brightness followed a power law.

Stevens' power law

Interpreting the power law.

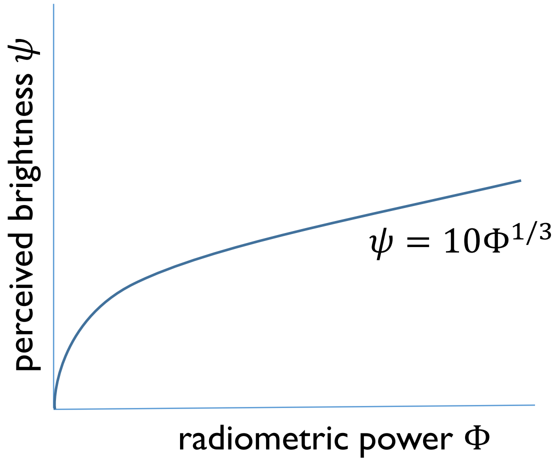
A constant (linear) increase in perceived brightness.

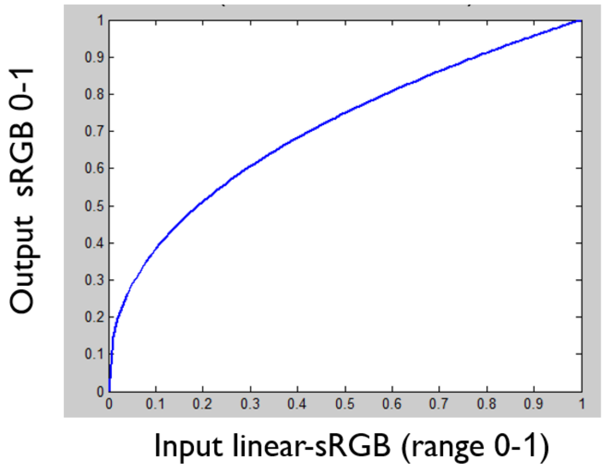


The radiant power needs to change exponentially.

sRGB gamma

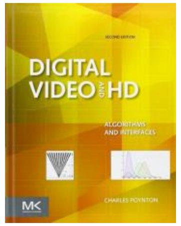
Steven's $\sqrt{3}$ power-law





sRGB gamma

- The sRGB gamma approximates Steven's $\frac{1}{3}$ power-law.
- The reason we apply gamma is that it remaps the linear color to fit better our visual system's nonlinear response to radiant power.
- There is a misconception in many graphics and image processing textbooks that gamma is applied to compensate for displays (CRTs). See a nice writeup about this by Poynton.[1]

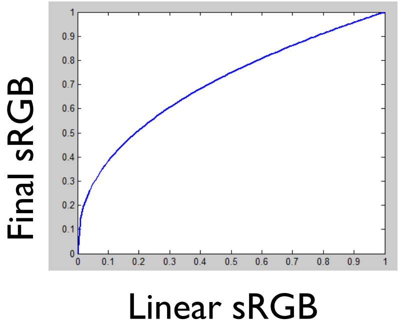


Poynton

[1] https://poynton.ca/PDFs/Rehabilitation_of_gamma.pdf

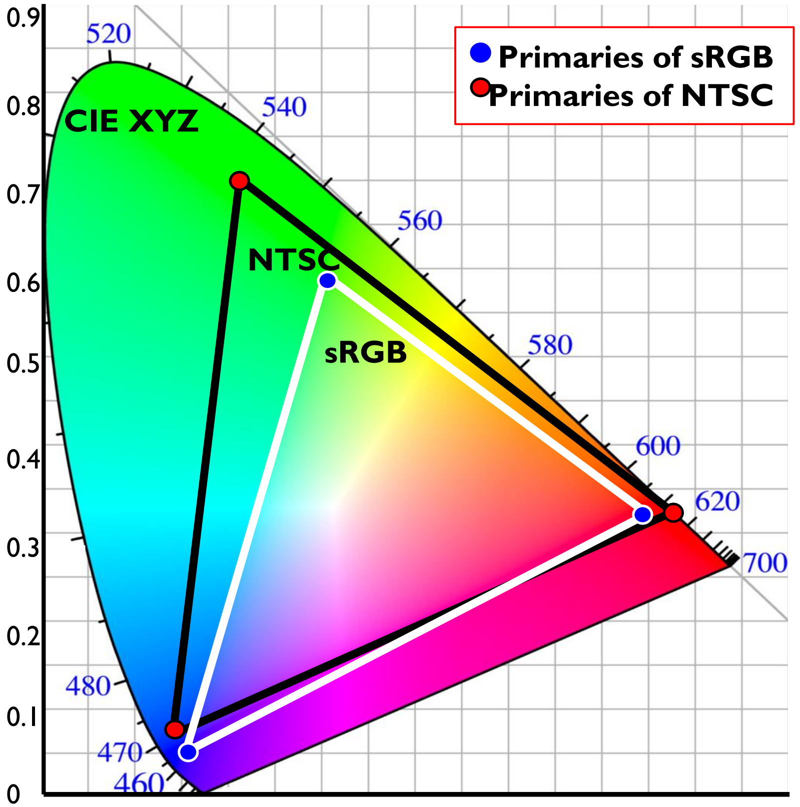
Before (linear sRGB) & after (sRGB)

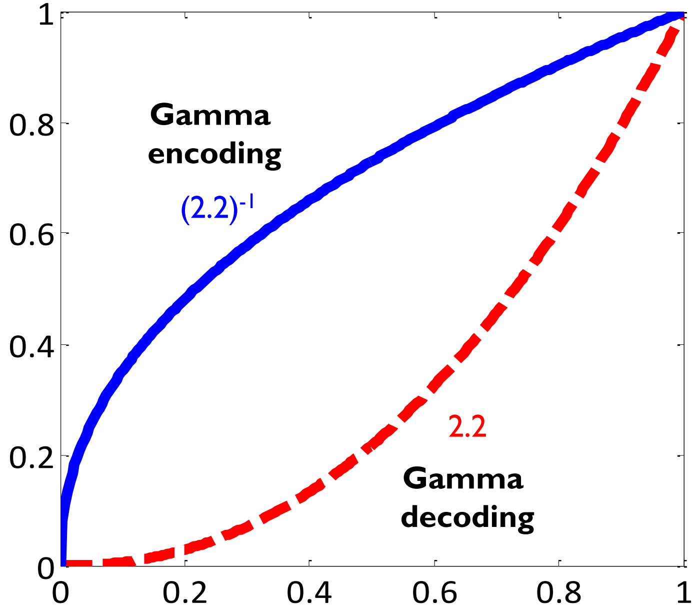
Linear sRGB



Final sRGB

Standardization is not new - NTSC/PAL





Both NTSC and sRGB used gamma encodings. Most color spaces use some type of perceptual encoding.

NTSC/sRGB
(know your color space!)

It is important to known which color space your image is in.

Linear-sRGB back to XYZ

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9505 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Linear-NTSC back to XYZ

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.6071 & 0.1736 & 0.1995 \\ 0.2990 & 0.5870 & 0.1140 \\ 0.0000 & 0.0661 & 1.1115 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Many color APIs (e.g., matlab, python) assume the default color space is NTSC. Many research papers use the wrong equations!

An additional fun fact

- Physical stimulus vs. human sensations
- Stevens' Power Law

$$S = kI^{a}$$

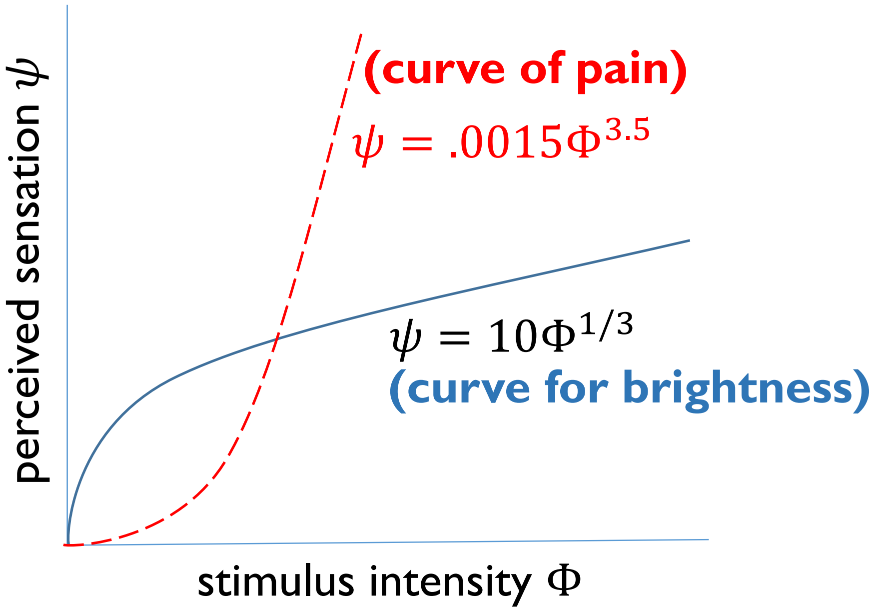
Human sensation

Constant

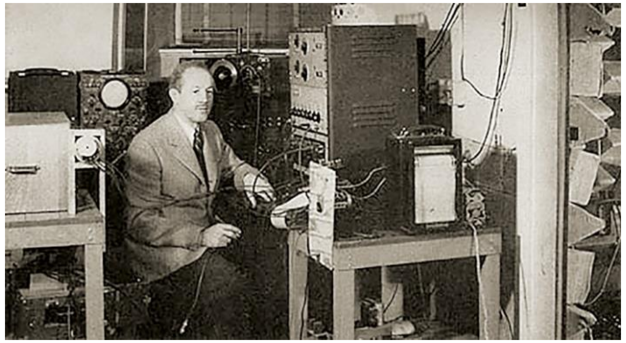
Stimulus intensity

power exponent

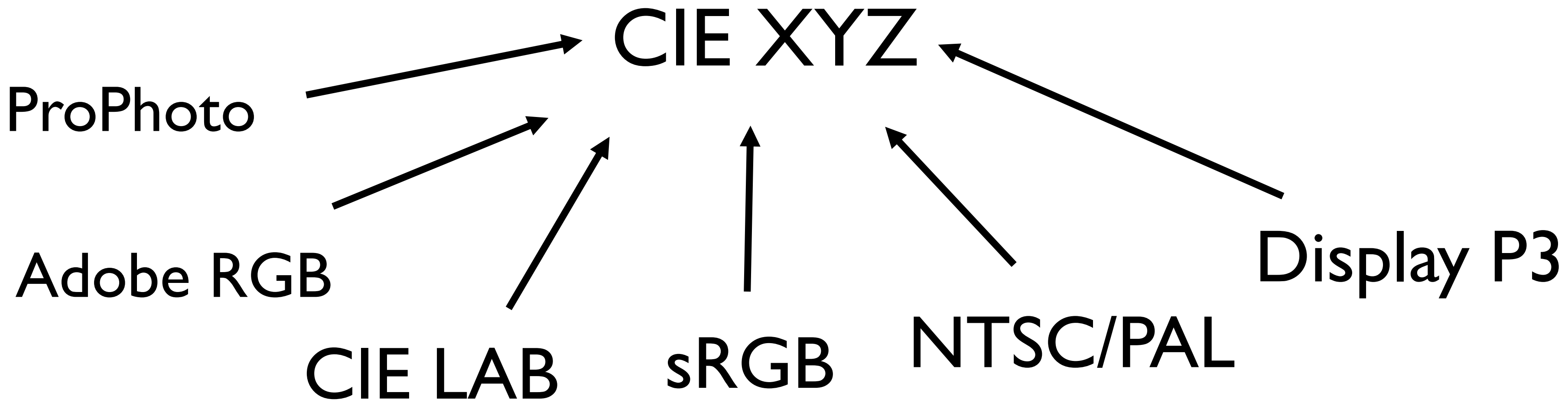
Dr. Stanley Stevens introduced showed that most human sensations follow a power-law relationship between stimuli and sensation.



Stevens also did experiment on the **pain sensation** of electrical shock! Turns out our sensitivity is the opposite than with radiometric power to brightness.



CIE XYZ: The mother color space



Other common color spaces

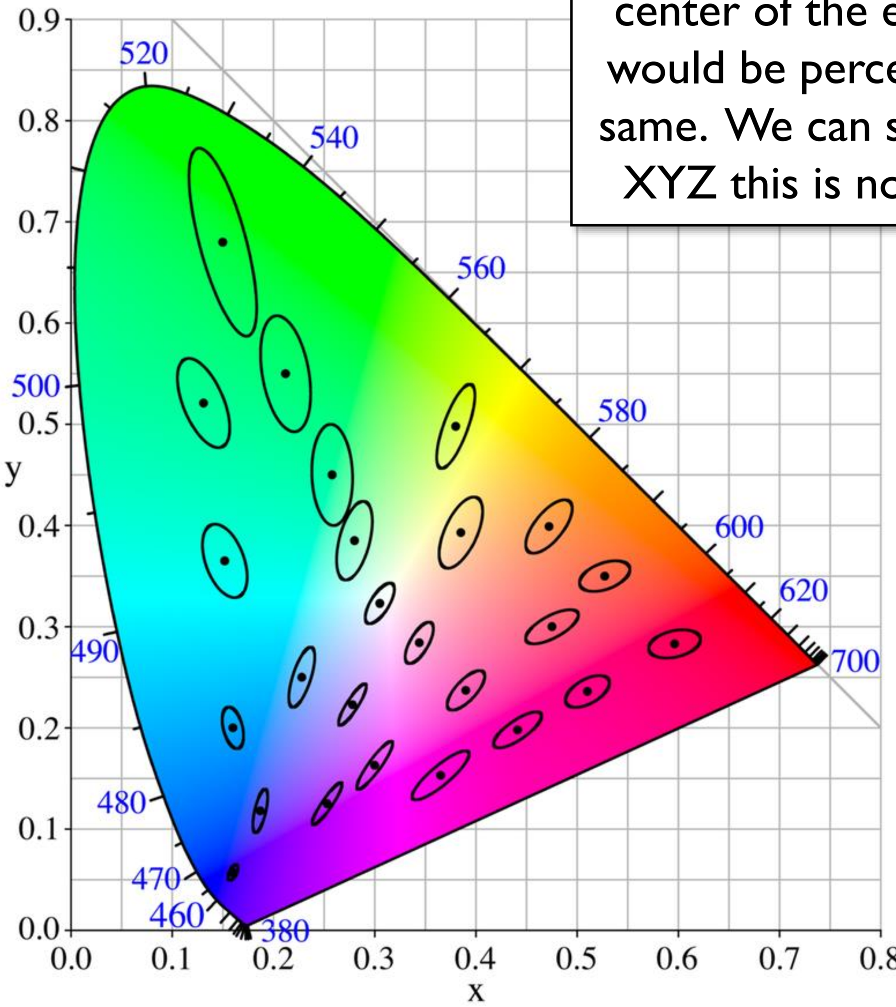
This tutorial does not go into the details of the mathematical transformations to other color spaces (we'd need another tutorial for that). You can find the transforms online.

The goal here is to explain the rationale behind each transform so you understand why the other color spaces are introduced.

CIE LAB space

- CIE LAB space (also written as CIE L*a*b*) was introduced as a perceptually uniform color space
- **Why?**
  - CIE XYZ provides a means to map between a physical SPD (radiometric measurement) to a colorimetric measurement (perceptual)
  - However, a uniform change in CIE XYZ space does result in an uniform change in perceived color difference (see diagram)
- CIE Lab transforms CIE to a new space where color (and brightness) differences are more uniform.

The ellipses shows the range of colors (around the center of the ellipse) that would be perceived as the same. We can see that CIE XYZ this is not uniform.



David MacAdam performed experiments on color perception. This plot is known as the MacAdam ellipses.

CIE 1976 LAB

- Considering the MacAdam experiments and the Steven's power-law, CIE LAB was derived in 1976 by applying various transformations to the CIE XYZ values that result in the following:
- L* represents a **perceptual brightness** measure between 0-100
  - L* is a non-linear (gamma) transformation of the Y component of CIE XYZ.
  - L is approximately a cube root of Y (directly from Steven's power law)
- a* and b* (often range ±50)
  - Both have similar non-linear transformations applied, and represent approximately:
  - a* values lying along colors related to red and green
  - b* values lying along colors related to yellow and blue
  - a*=b*=0 represents neutral grey colors

NOTE: CIE LAB requires the white point to be specified for the transformation. The default white point is D65.

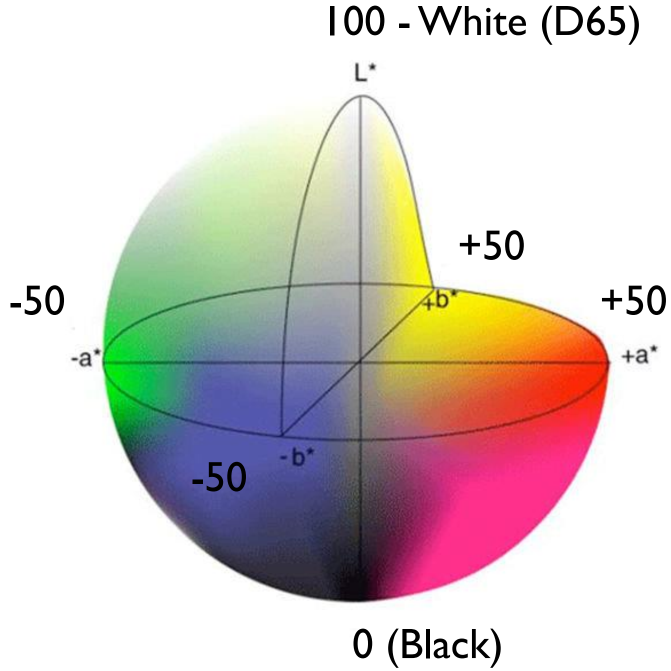
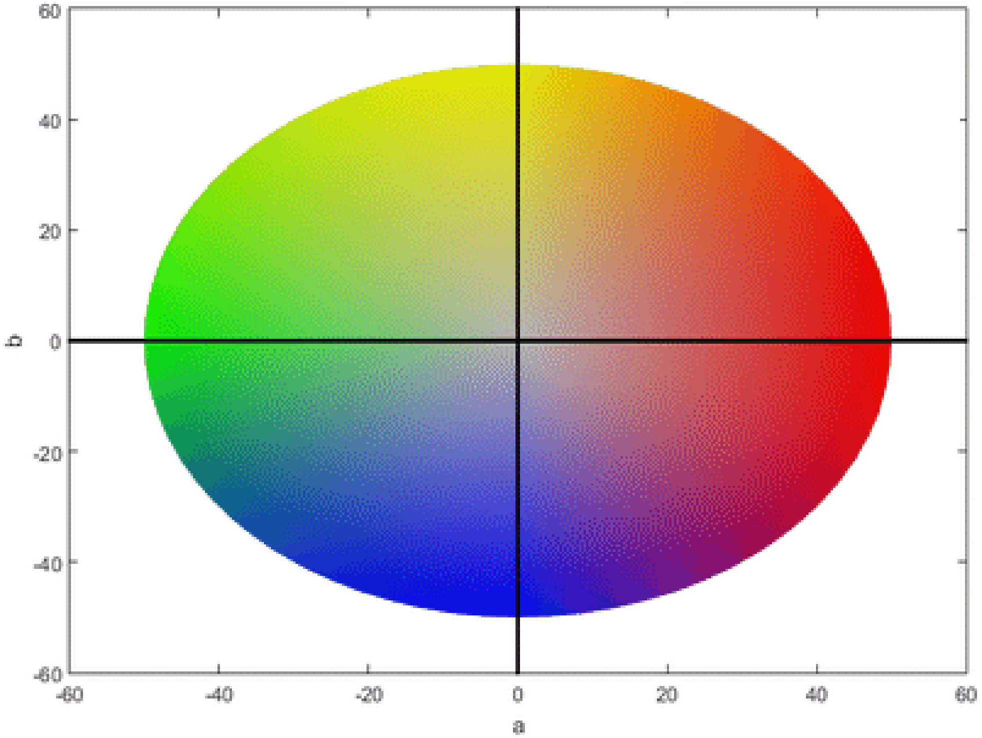


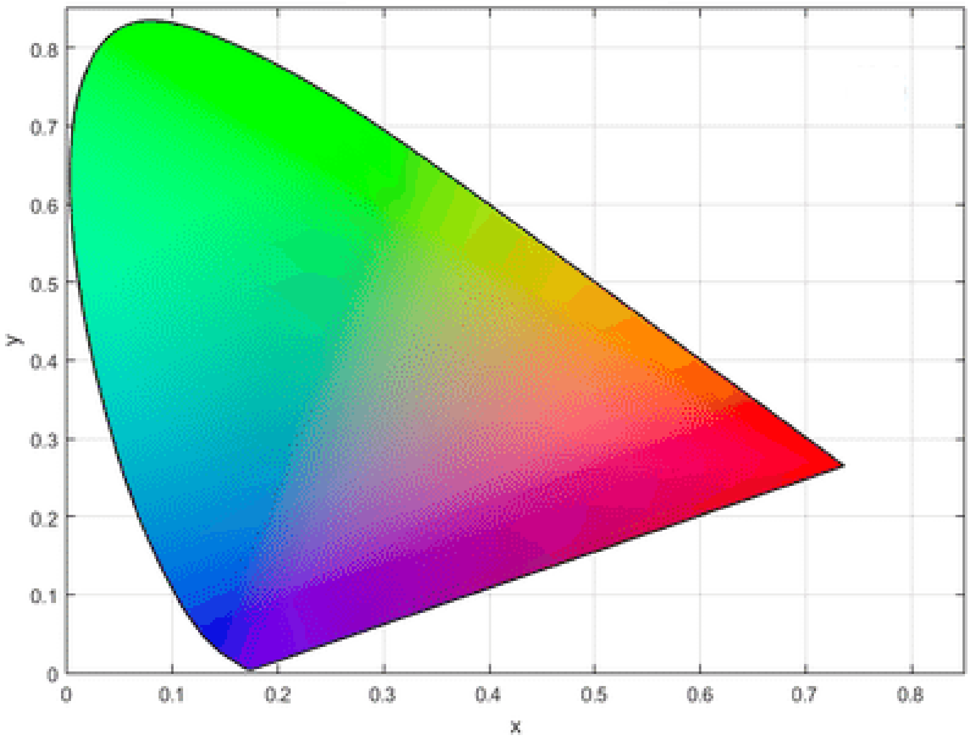
Image from Mohamed Cheriet

CIE LAB

CIE-L*ab space



CIE-xyY space



Chromaticity comparison's between CIE LAB and CIE XYX

Image from Bagdasar et al ICSTCC'17

Color error metric – CIE 2000 Delta E (ΔE)

- Delta E is a color metric based on L*ab space.
- Since L*ab is more uniformly perceptual, distances (e.g., Euclidean distance) in L*ab have more meaning than in CIE XYZ.
- Delta E values have an interpretation as follows.

Delta E	Perception
<= 1.0	Not perceptible by human eyes.
1 - 2	Perceptible through close observation.
2 - 10	Perceptible at a glance.
11 - 49	Colors are more similar than opposite
100	Colors are exact opposite

In general, a ΔE of 2 or less is considered to be very good. It means a standard observer could not tell that two colors are different unless they observe them very closely.

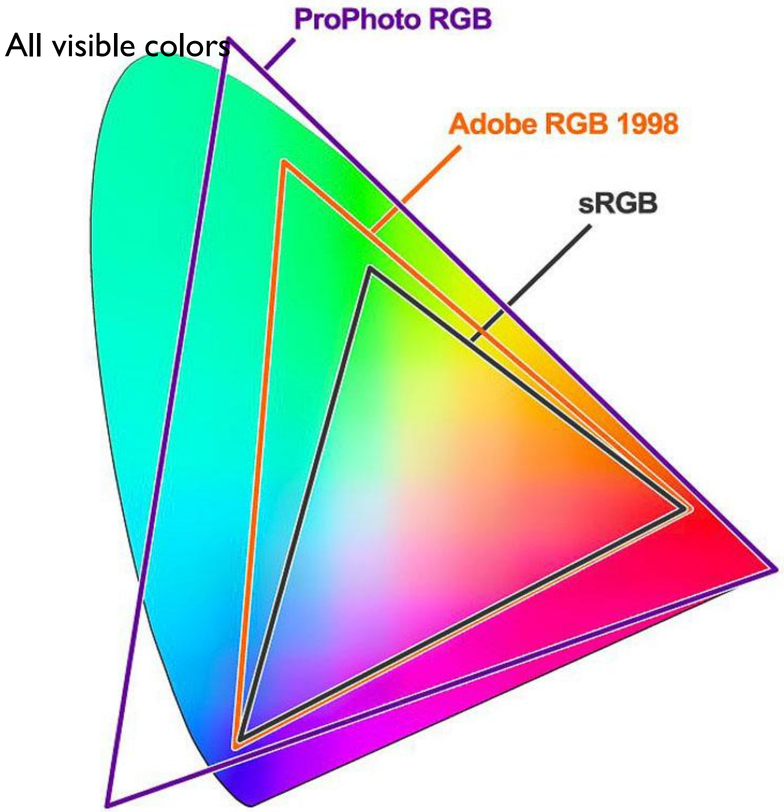
Table from
https://zschuessler.github.io/DeltaE/learn/

Other color spaces to be aware of

- Adobe RGB
  - Medium gamut color space
  - Used for photo-editing
- Display P3
  - Medium gamut color space
  - Used by Apple devices to accommodate better display technology
  - Similar to Adobe RGB
- ProPhoto (ROMM)
  - Developed by Kodak
  - Intended to encode a wide range of colors and dynamic range

These are known as "output referred" color spaces because they are defined for encoding images for display or output devices. The definition of color spaces also states the space's preferred dynamic range and viewing environment (although we rarely view in such conditions).

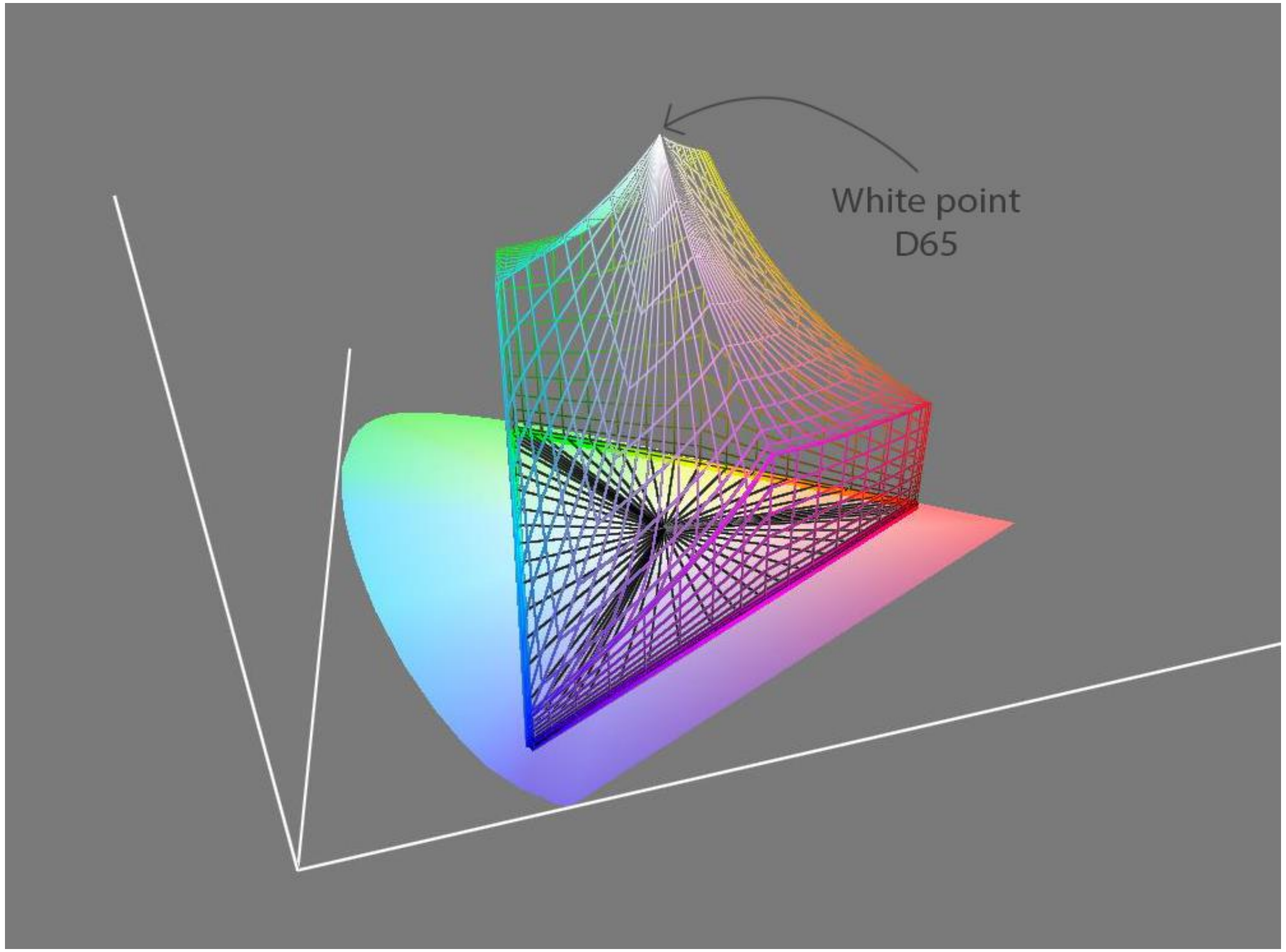
Color space’s gamut



CIE Yxy chromaticity

A color space’s gamut is the span of colors that can be represented. The 3D gamuts are plotted in CIE L*ab.

Gamuts expressed in chromaticity are misleading

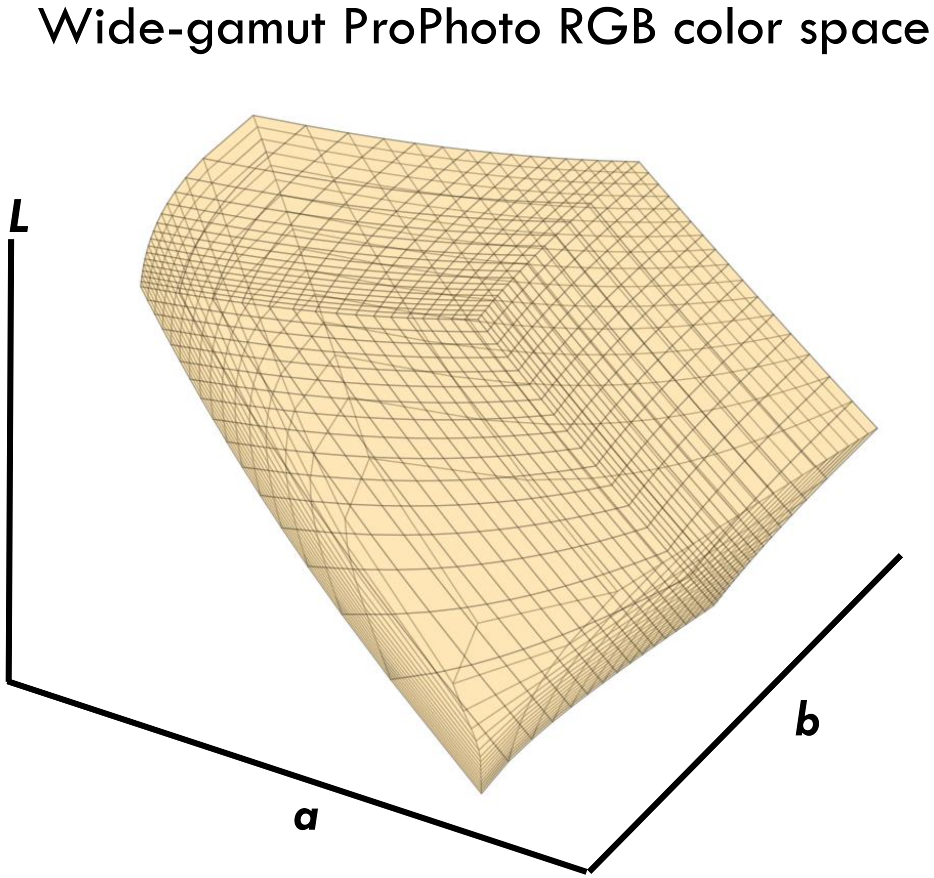


AdobeRGB plotted
in CIE XYZ and
then projected to
2D CIE Yxy chromaticity.

Image: Alp_Er_Tunga

*See slide 44 (Mark Fairchild's comments)

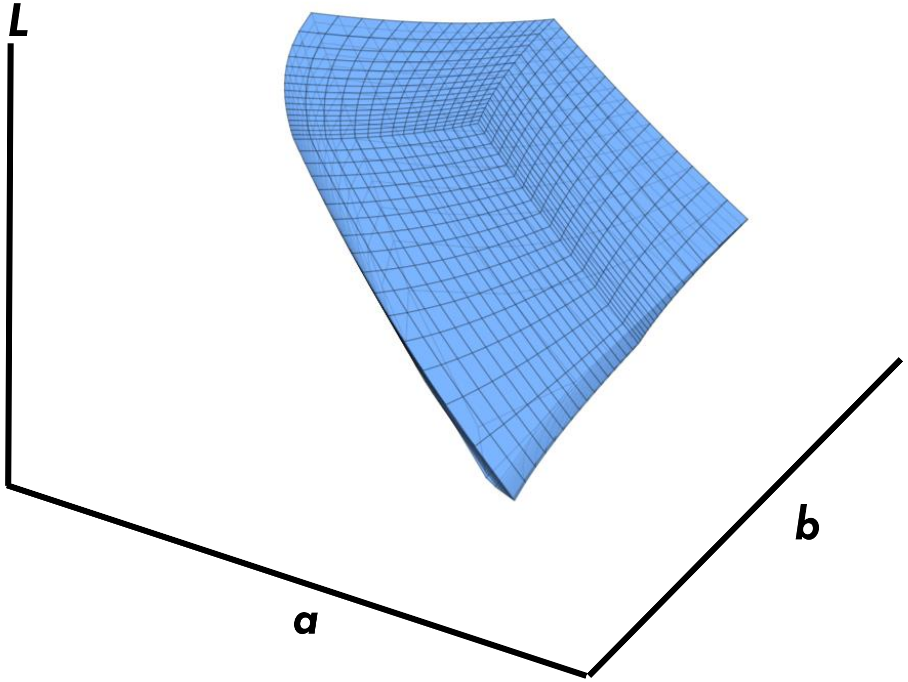
ProPhoto color space



ProPhoto encodes over 90% of surface colors (color from reflected light of a surface, i.e., not emitted light).
It is recommended to use 16-bit values per channel since the gamut is so large.
The white point is D50.

Adobe RGB/Display P3 color space

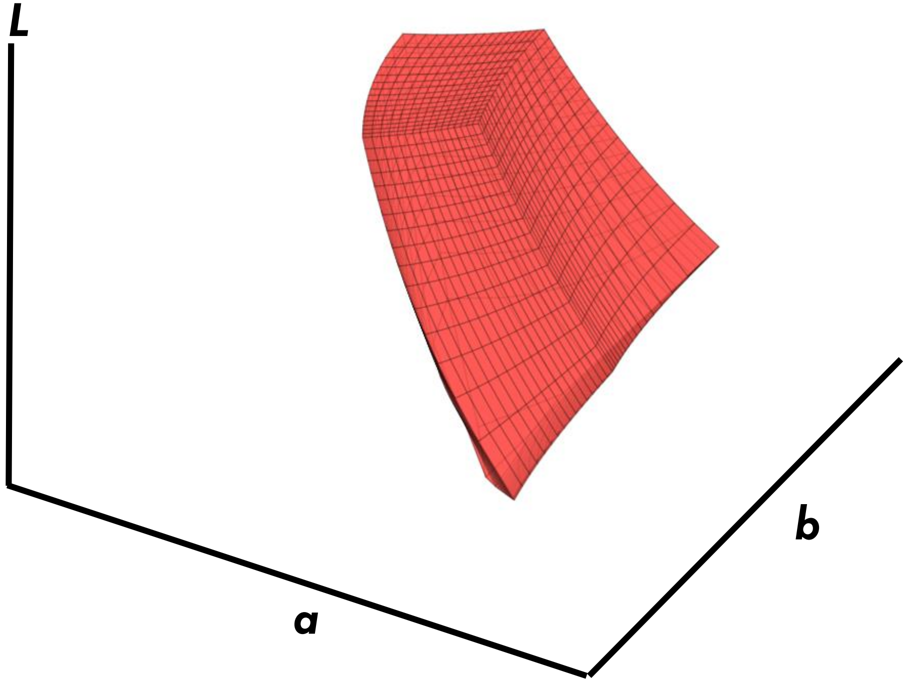
Medium-gamut AdobeRGB color space.
(Apple's Display-P3 is very similar).



AdobeRGB/Display P3 encodes over ~50% of surface colors. Display P3 is Apple's encoding color space.
It is recommended to use 10-bit per channel.
The white point is D65.

sRGB color space*

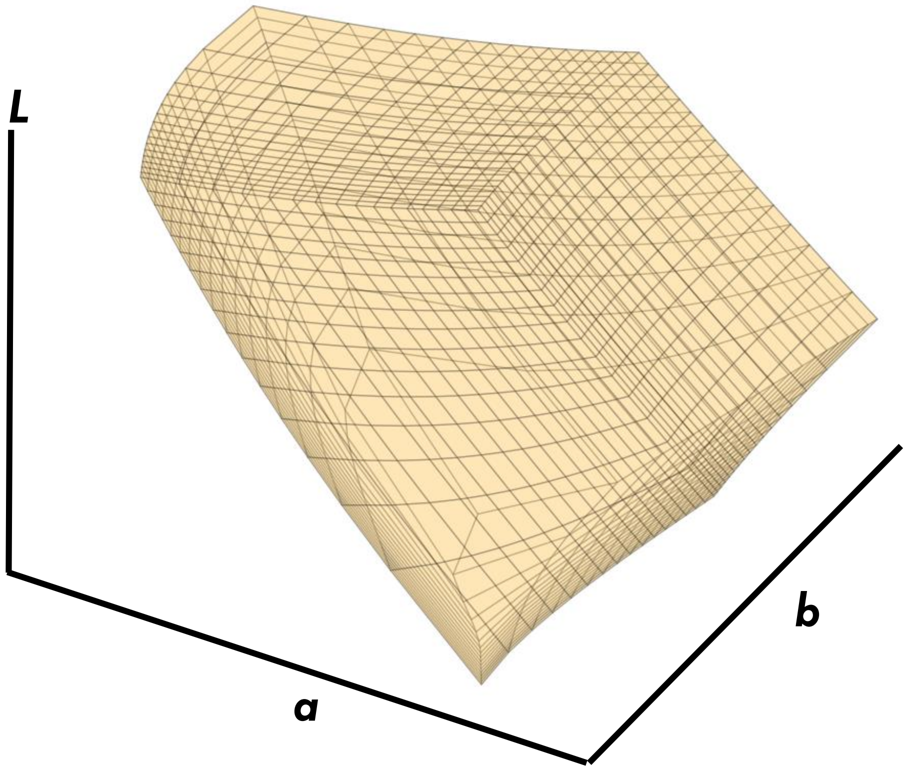
Small-gamut standard RGB (sRGB) color space



*Currently, sRGB is the most common color space (designed for 1990s display technology).
sRGB encodes ~30% of surface colors.
Developed for 8-bit encoding per channel.
The white point is D65.

Displaying different encodings

Wide-gamut ProPhoto RGB color space



PowerPoint expects images to be in sRGB. When encoded in a wider gamut color space, the image may appear dull. This may seem counter-intuitive because the wider gamut should encode more colors, but this is only possible when the software and display hardware are aware (and capable) of interpreting the color space correctly.

Now, are we really finally done with color?

Yes . . .

But remember that color appearance, measurement, and encoding is its own research field. My slides provide only a basic introduction. The CV community is bad at abusing color terminology or not putting in enough effort to understand color fully.

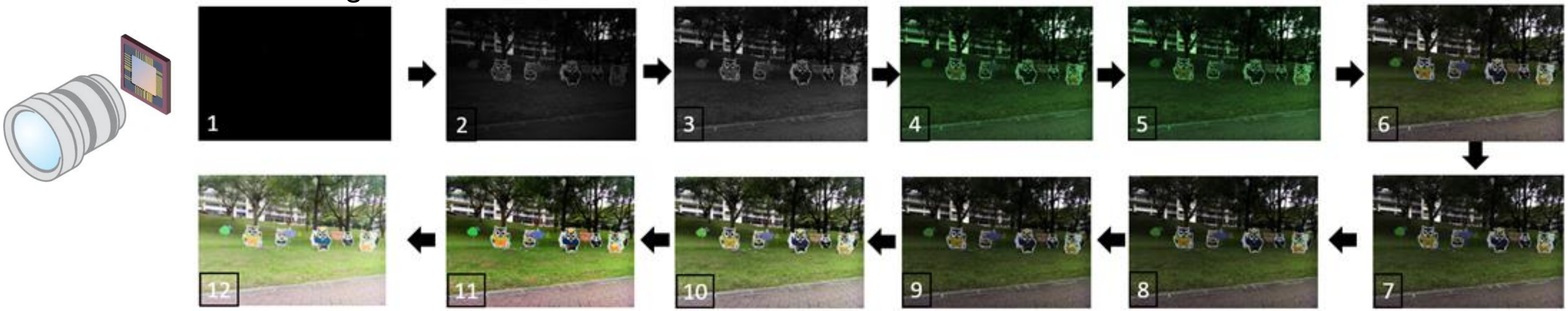
Congratulations!

Part 2: Overview of the in-camera rending pipeline

In-camera rendering

- The image directly captured from the camera's sensor needs to be processed.
- We can call this process "rendering," as the goal is to render a digital image suitable for viewing.

Sensor image



Final rendered image in sRGB

Intermediate states of the image being "rendered" to sRGB.

Image signal processor (ISP)

- An ISP is dedicated hardware that renders the sensor image to produce the final output.
- Companies such as Qualcomm, HiSilicon, Intel (and more) sell ISP chips (often as part of a System on a Chip – SoC).
  - Companies can customize the ISP.
- Many ISPs now have neural processing units (NPUs).

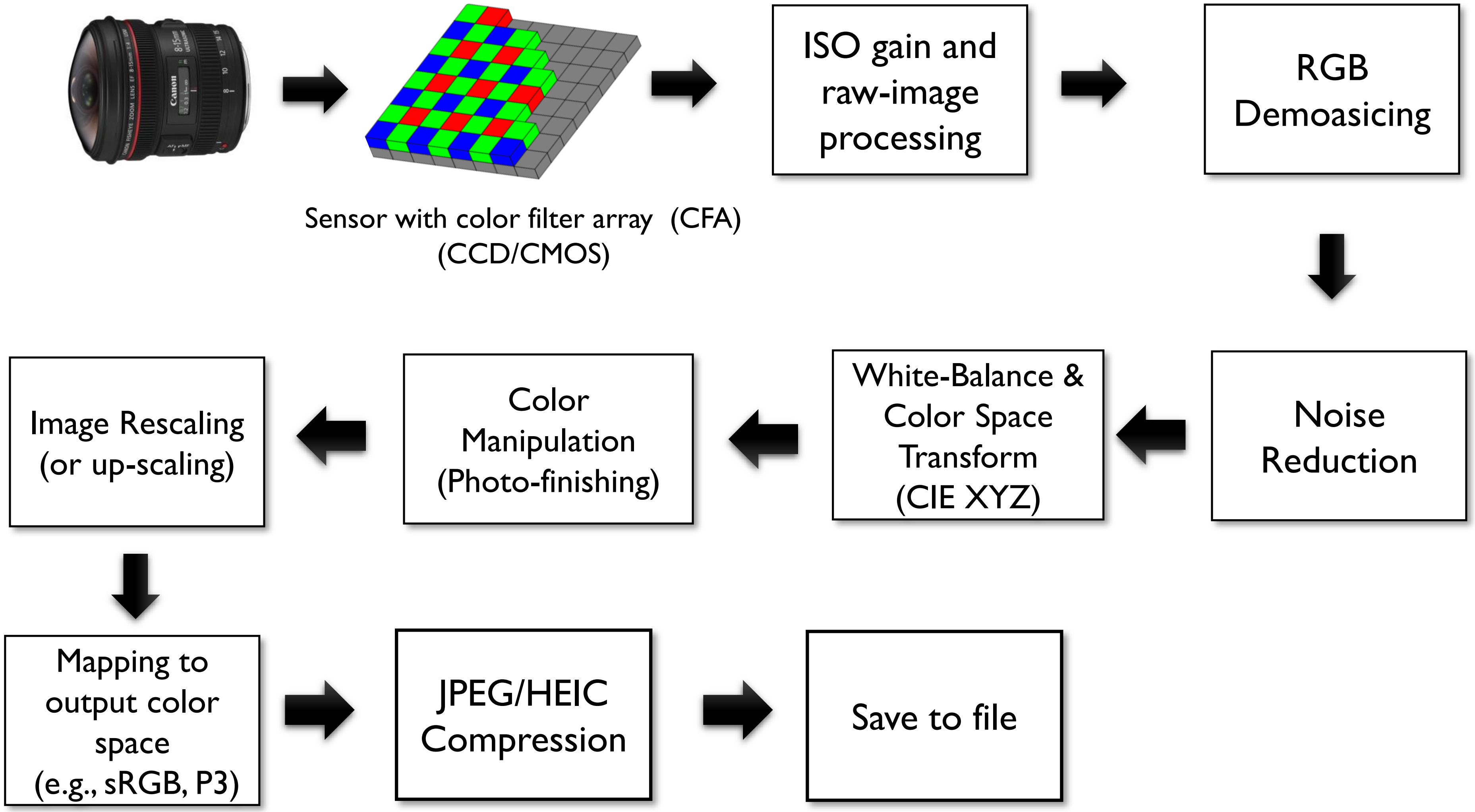
Samsung

Huawei

Apple

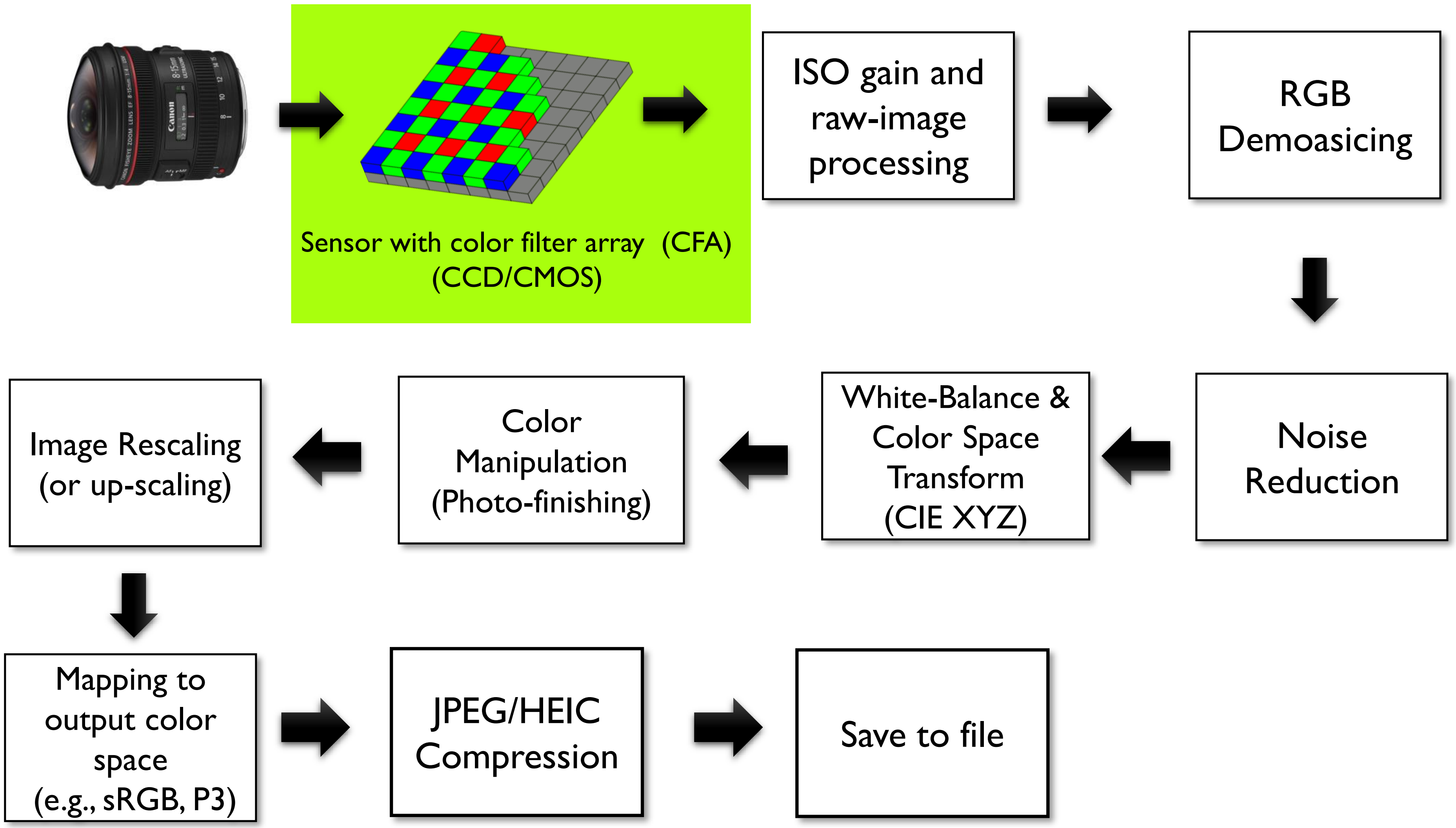
Samsung/Pixel/OnePlus/Xiaomi/….

A typical color imaging pipeline



NOTE: This diagram represents the steps applied on a typical consumer camera pipeline. ISPs may apply these steps in a different order or combine them in various ways. A modern camera ISP will undoubtedly be more complex but will almost certainly implement these steps in some manner.

A typical color imaging pipeline



Camera sensor

CMOS sensor

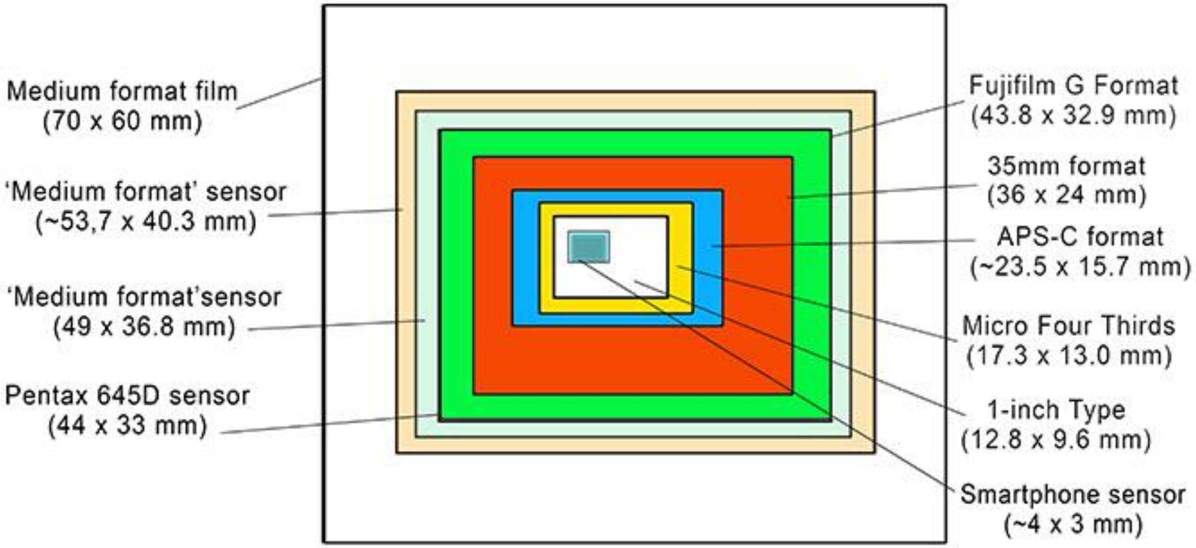
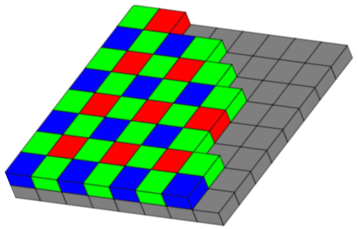


Figure from Photo Review website.

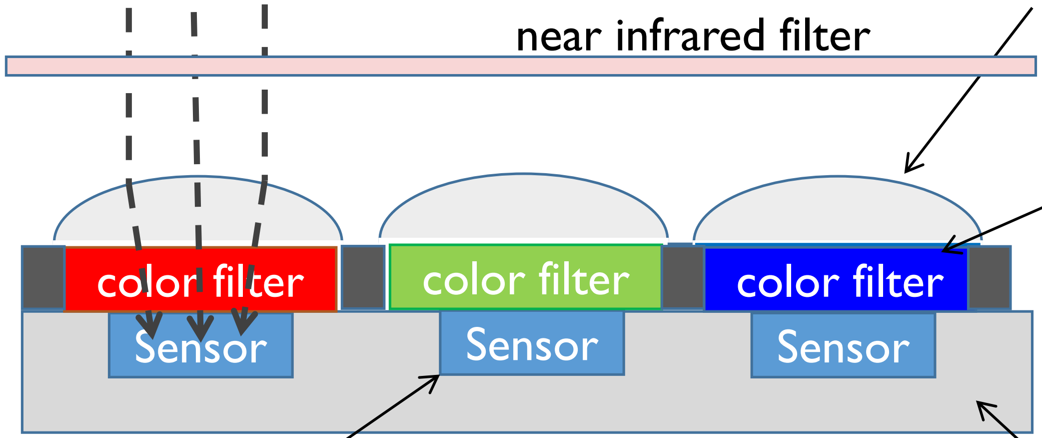
Almost all consumer camera sensors are based on complementary metal-oxide-semiconductor (CMOS) technology.
We generally describe sensors in terms of number of pixels and size. The larger the sensor, the better the noise performance as more light can fall on each pixel.   Smart phones have small sensors!

Camera sensor RGB values

A Near Infrared (NIR) filter is often placed before the sensor. (This is sometimes called a "hot mirror"). This is because red filters often respond to NIR light.



Color filter array or "Bayer" pattern.



Micro-lenses are placed over the diode to help increase light collection on the sensor

Color filters place over the sensor. This forms a Color Filter Array (CFA) also called a "Bayer Pattern" after inventor Bryce Bayer.

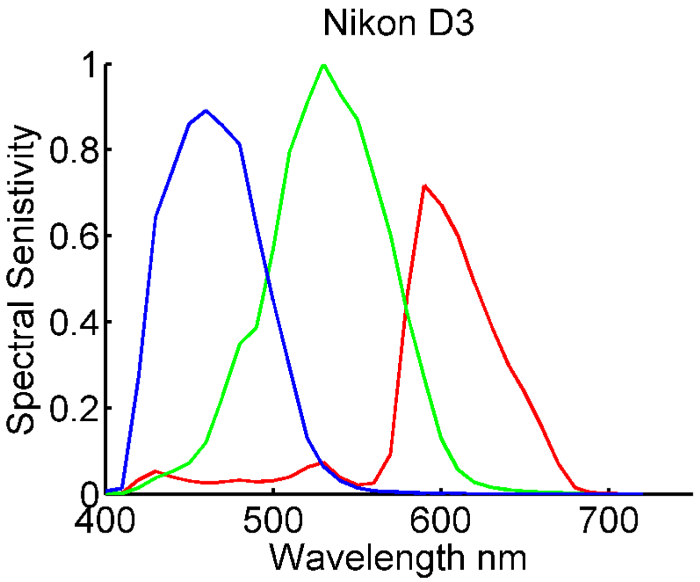
Photodiode
Photons hit the diode and force out electrons. This design is similar to a solar cell!

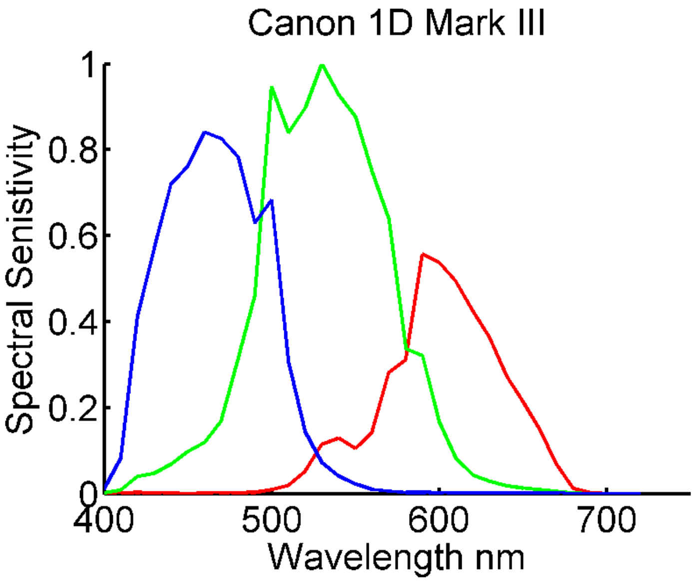
Silicon/Circuitry

Bryce Bayer (Kodak)

Camera RGB sensitivity

- The color filter array (CFA) on the camera filters the light into three *sensor-specific* RGB primaries





Plotted from camera sensitivity database by Dr. Jinwei Gu from Rochester Institute of Technology (RIT). Dr. Gu is now at SenseTime (USA). http://www.cis.rit.edu/jwgu/research/camspec/

Measuring camera sensitivities

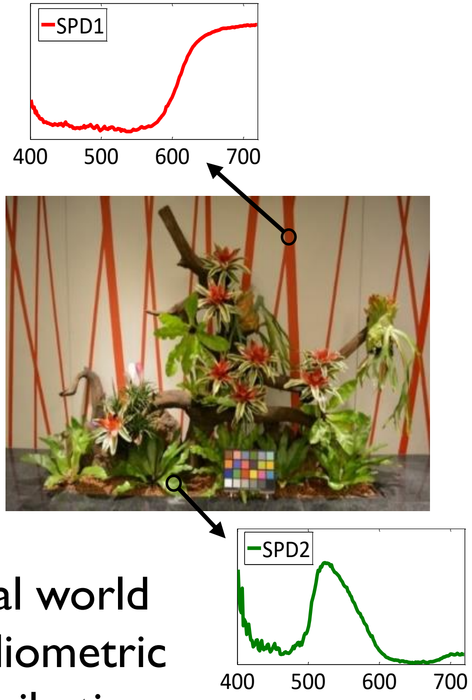
- It is not easy to get information on a camera's spectral sensitivities.
  - This process is called camera or sensor characterization.
  - The sensitivity needs to factor in the entire camera form factor: lens, NIR filter, and CFA.
- You need specialized equipment to measure camera spectral sensitivities.
  - But of course, reviewer 2 will say obtaining sensitivities curves is easy. . . .

Image Engineering GmbH & Co. KG

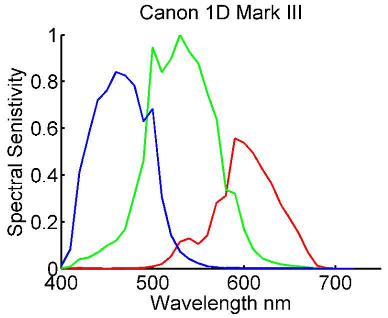
camSPEC device for measuring camera spectral sensitivity.

https://www.image-engineering.de/products/equipment/measurement-devices/588-camspecs-express

Sensor raw-RGB image



Remember: physical world is measured by radiometric spectral power distributions.



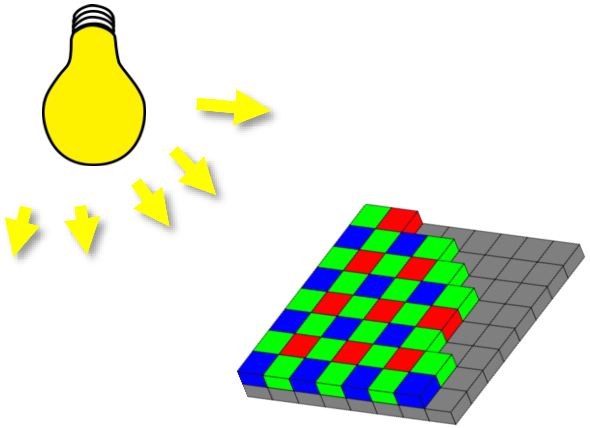
Camera spectral sensitivities.

Your camera sensor RGB filter is sensitive to different regions of the incoming SPD.

raw-RGB represents the physical world's SPD "projected" onto the sensor's spectral filters.

Sensors are linear to irradiance

- Camera sensors are decent light measuring devices.
- If you double the amount of light hitting a sensor's pixel, the digital value output of that pixel will double.

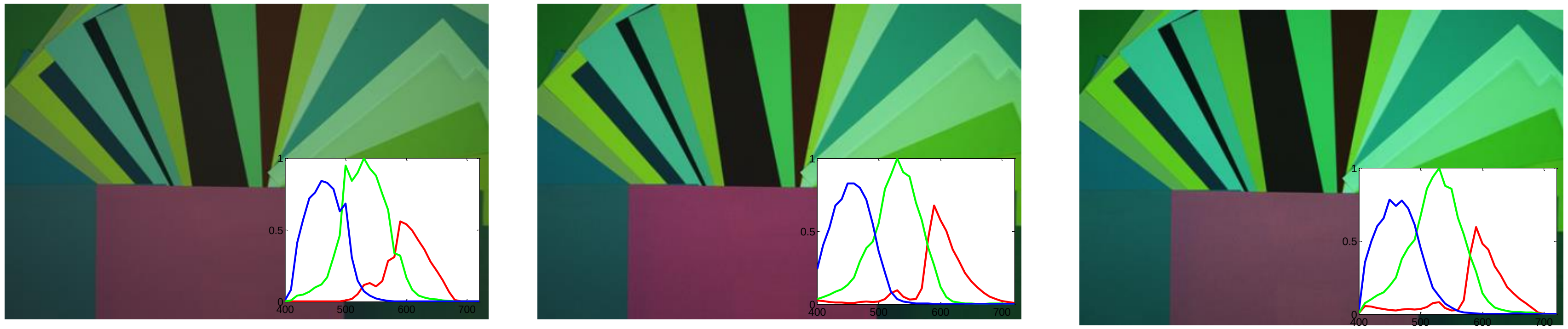


Sensor output is linear with respect to irradiance falling over the sensor over a certain amount of time.

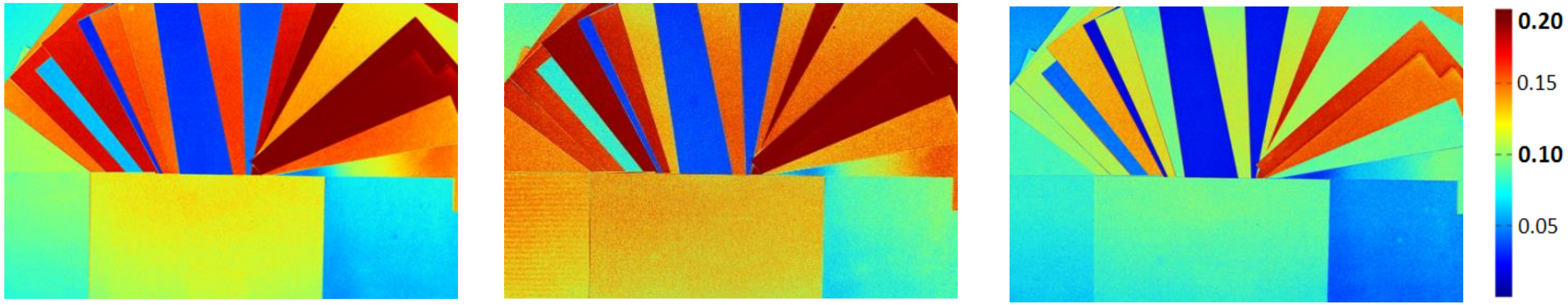
$$I = i * t$$

Digital value I is a linear function of irradiate i and exposure t.

IMPORTANT: raw-RGB sensor images are <u>not</u> in a standard color space.



Canon 1D **Nikon D40** **Sony α57**



||Canon 1D – Nikon D40||$_2$ **||Canon 1D - Sony α57||$_2$** **||Nikon D40 - Sony α57||$_2$**

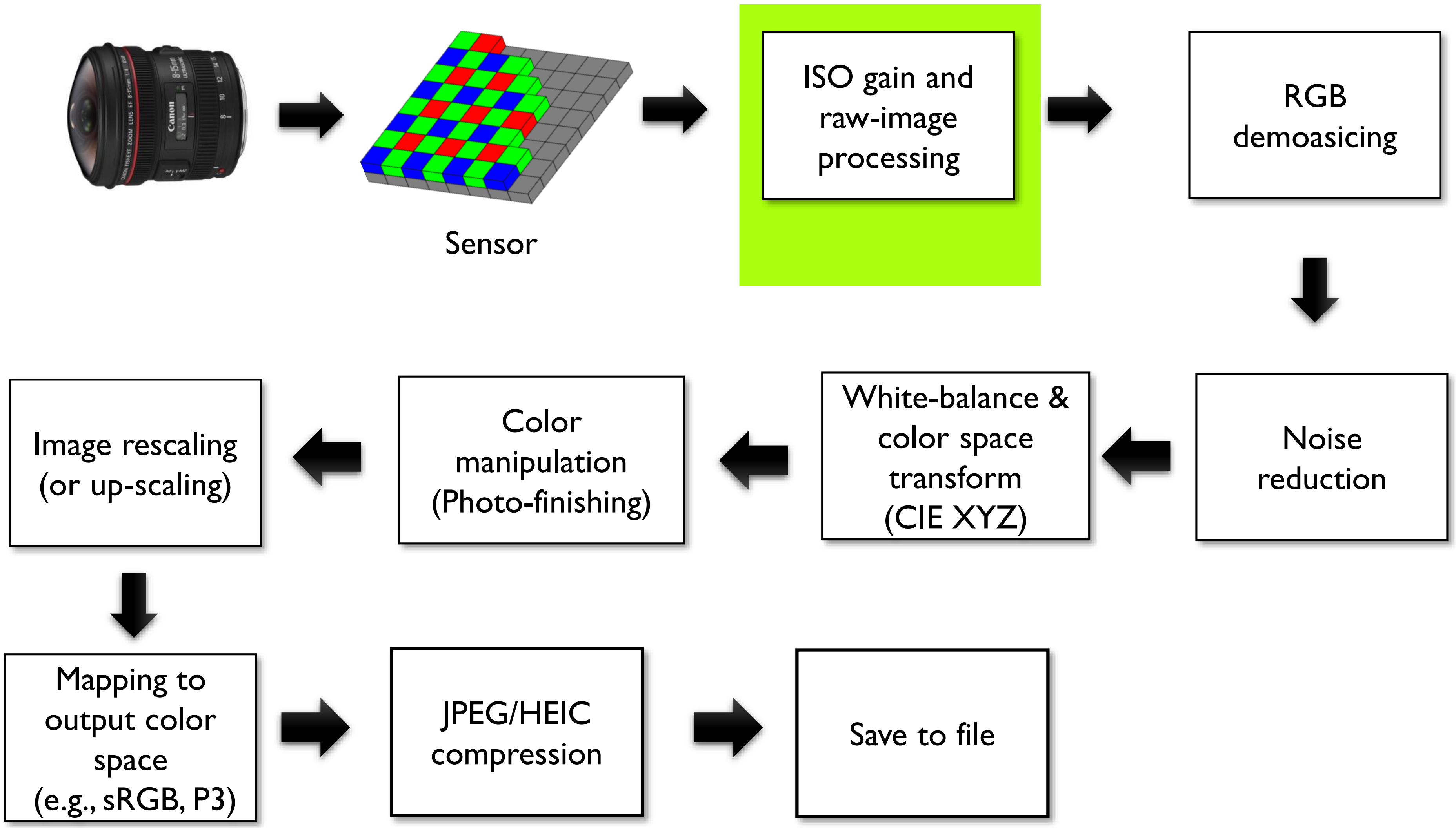
Color plots show L2 distance between the raw-RGB values with different cameras.

Displaying raw-RGB images

- Inserting a raw-RGB image in your slides, research paper, etc will result in strange colors.
- Why? Our devices (computers, printers, etc) expect the image to be in a standard color space like sRGB.

This is a raw-RGB image. Why does it look bad?
Because the raw-RGB values are not sRGB values.

A typical color imaging pipeline



ISO signal amplification (gain)

- Imaging sensor signal is amplified and digitized.
- Amplification to assist A/D conversion.
  - Need to get the voltage to the range required for the desired digital output.
- This gain is used to accommodate camera ISO settings.
  - Gain to signal applied on the sensor.
  - Note – gaining the signal also gains image noise.

Different ISO settings (note: the exposure will be shorter for higher ISO)

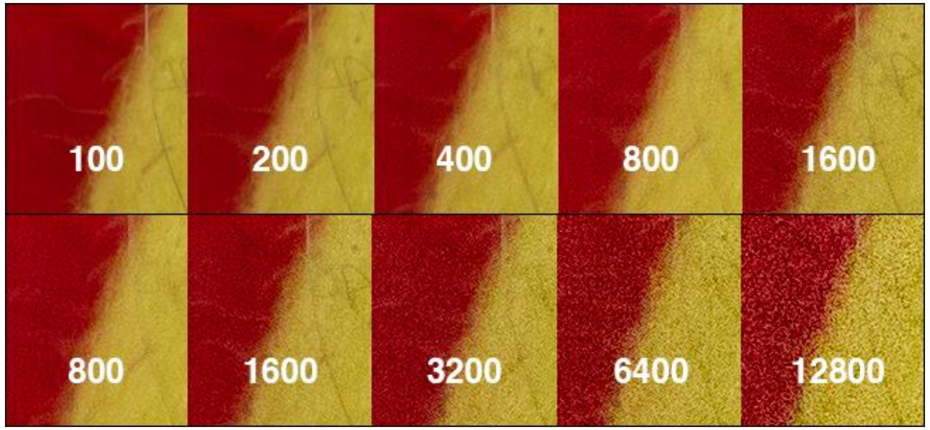


Image: Harry Guinness

ISO gain and raw-image processing

Pixel "intensity"

- We often talk about a pixel's intensity, however, a pixel's numerical value has no *unit*.

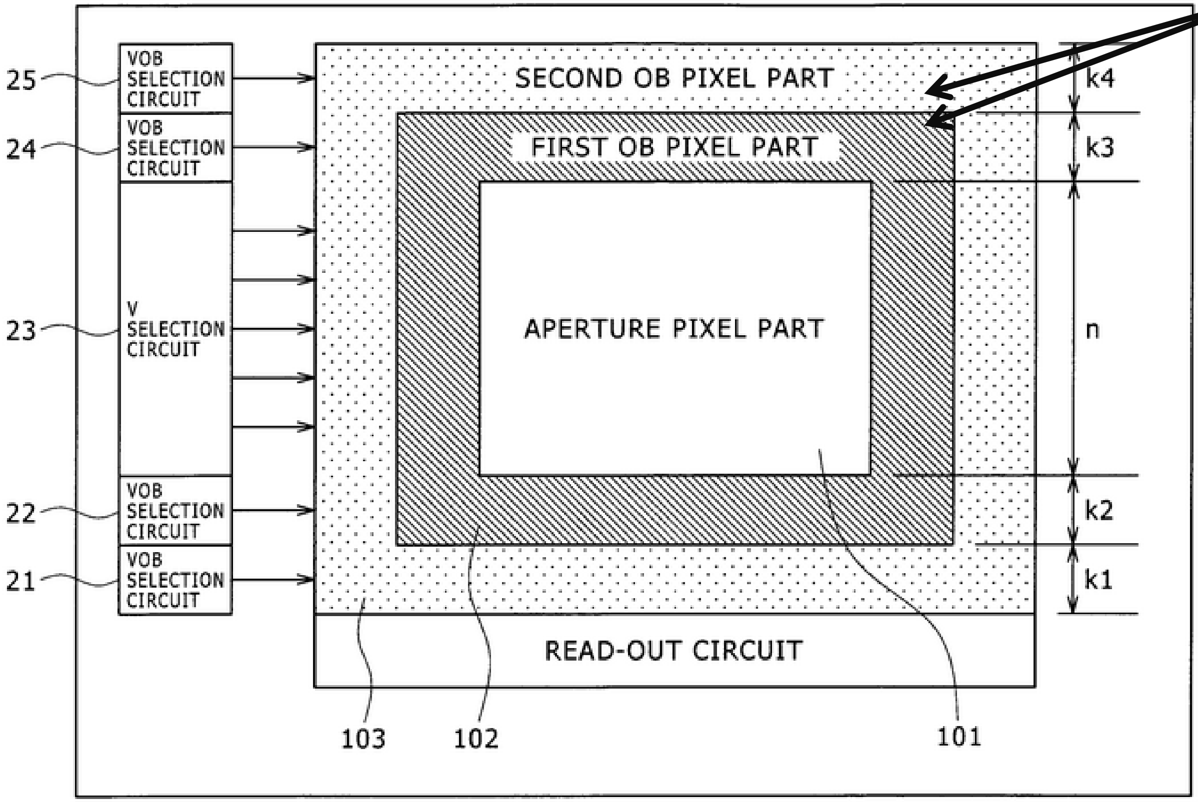
- The digital value of a pixel is based on several factors.
  - Exposure (which is a function of both shutter speed and exposure)
  - Gain (ISO setting on the camera)
  - Camera hardware that digitizes the signal.

- We typically rely on the ***relative*** digital values in the image and not the absolute digital values

Black light subtraction

- Sensor values for pixels with "no light" should be zero.
- However, this is not the case due to sensor noise.
  - The black level often changes as the sensor heats up.
- This can be corrected by capturing a set of pixels that do not see light
- Place a dark shield around the sensor.
- Subtract the level from the "black" pixels.

Optical black (OB)



Sensor area(s) capturing "optical black"

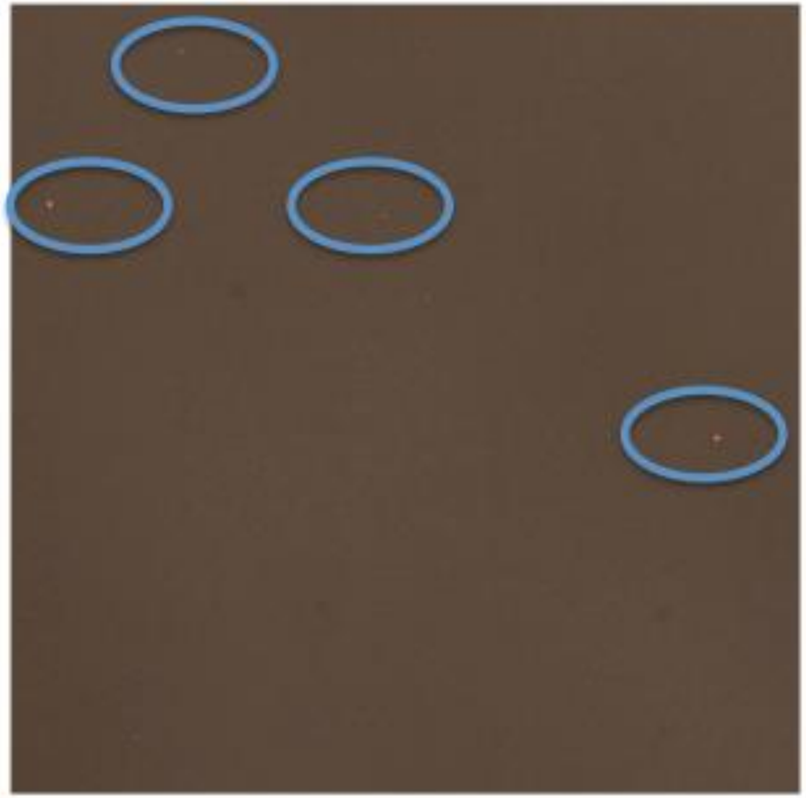
Black light capturing areas (likely exaggerated) from Sony US Patent US8227734B2 (Filed 2008) .

ISO gain and raw-image processing

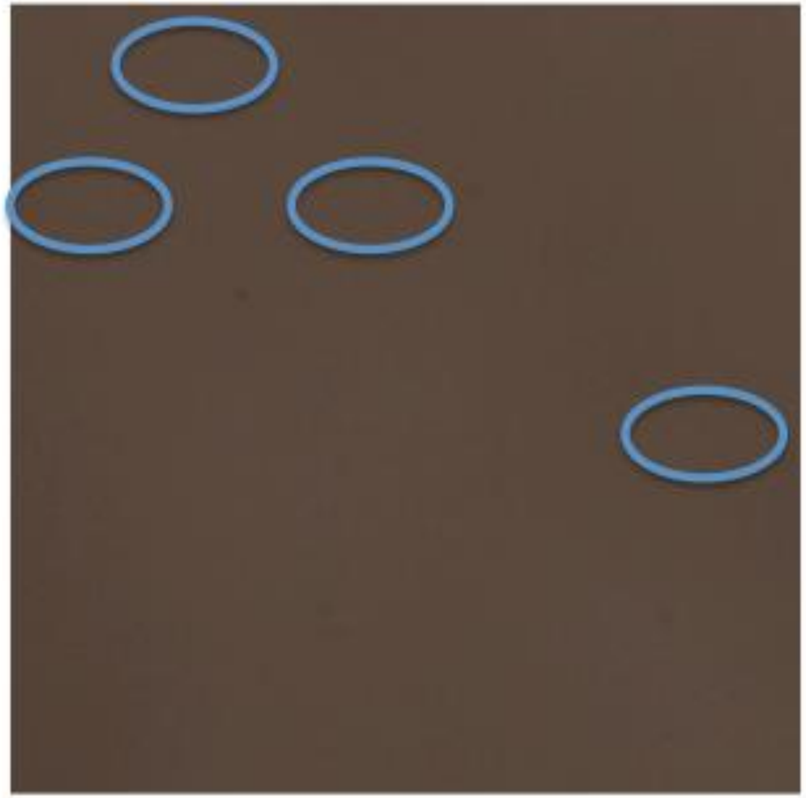
Defective pixel mask

- CMOS have pixels that are defective.
- Dead pixel masks are pre-calibrated at the factory
  - Using "dark current" calibration
  - Take an image with no light
  - Record locations reporting values to make "mask."
- Bad pixels in the mask are interpolated.

Defective pixel mask example



Identifying "dead pixels"

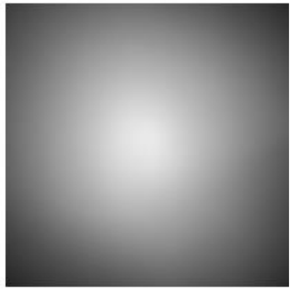


After interpolation

ISO gain and raw-image processing

Image courtesy of Lu Zheng and Moshe Ben-Ezra

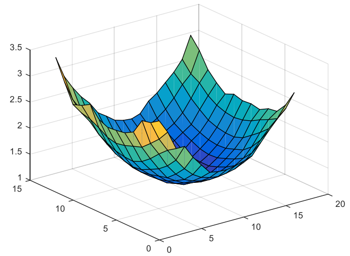
Flat-field correction



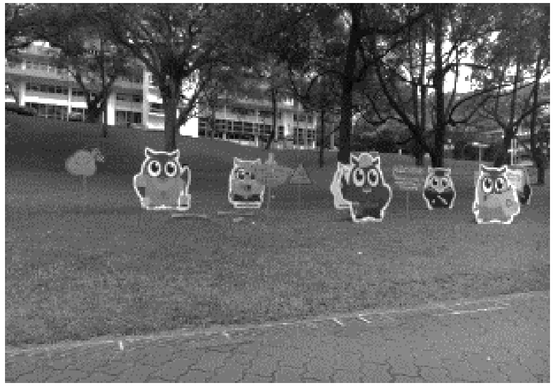
Uniform light falling on the sensor may not appear uniform in the raw-RGB image. This can be caused by the lens, sensor position in the camera housing, etc.

We want to correct this problem such that we get a "flat" (or uniform) output.

Before correction



Apply a correction gain over the sensor values.

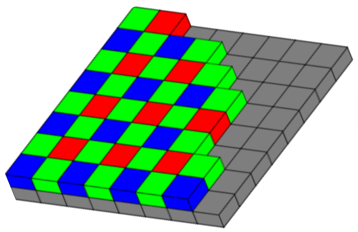


After correction

ISO gain and raw-image processing

This operation can also be called *lens shading correction*.

A typical color imaging pipeline



Sensor

ISO gain and raw-image processing

RGB demoasicing

Noise reduction

White-balance & color space transform (CIE XYZ)

Color manipulation (Photo-finishing)

Image rescaling (or up-scaling)

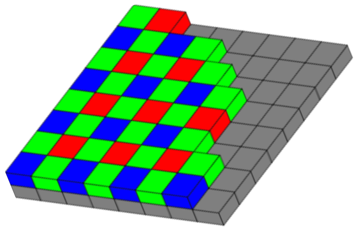
Mapping to output color space (e.g., sRGB, P3)

JPEG/HEIC compression

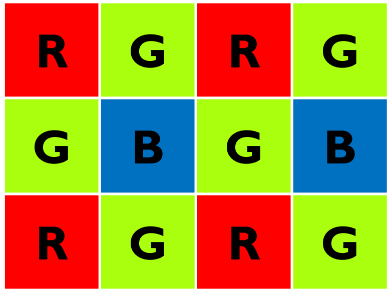
Save to file

CFA/Bayer pattern demosaicing

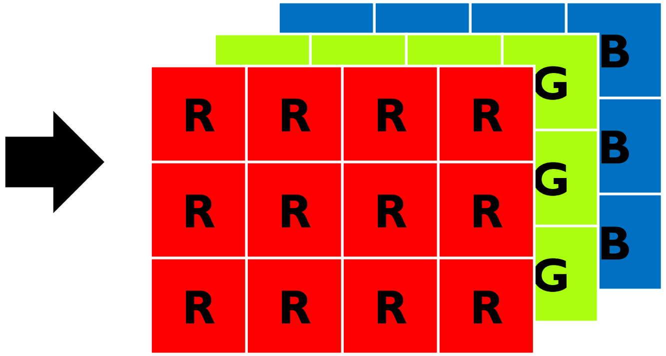
- Color filter array (CFA) pattern placed over pixel sensors.
- We want an RGB value at each pixel, so we need to perform interpolation.



Sensor with color filter array (CMOS)



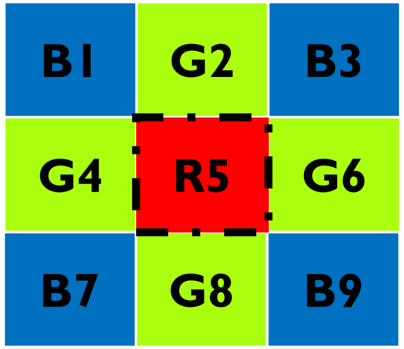
Sensor RGB layout



Desired output with RGB per pixel.

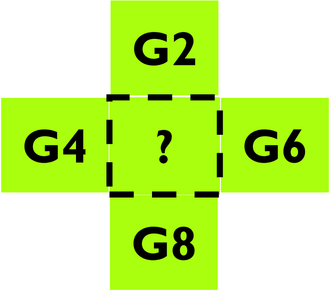
Simple interpolation

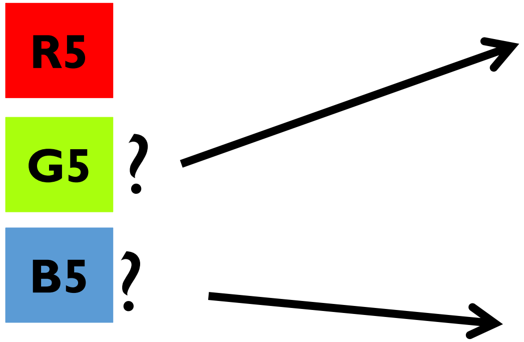
This is a zoomed up version of the Bayer pattern.



At location R5, we have a red pixel value, but no Green or Blue pixel.

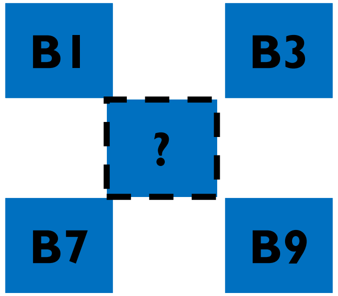
We need to estimate the G5 & B5 values at location R5.





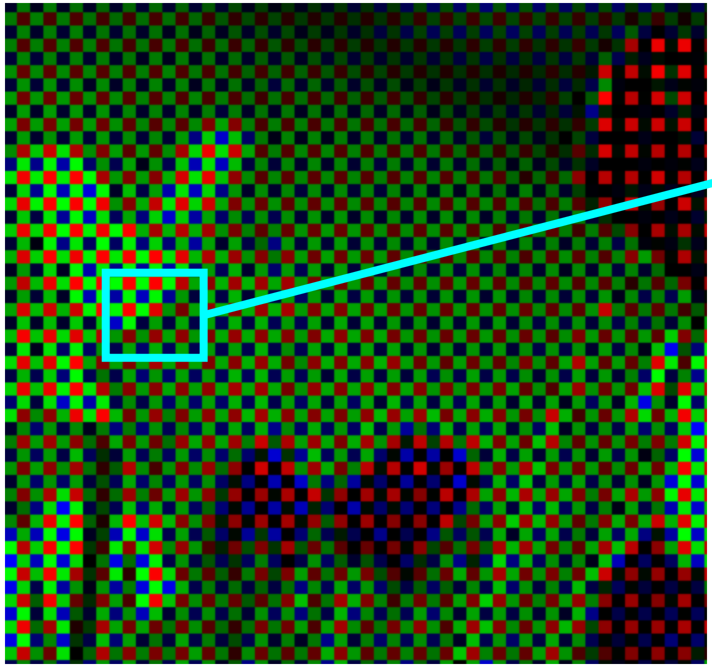
$$G5 = \frac{G2 + G4 + G6 + G8}{4}$$

$$B5 = \frac{B1 + B3 + B7 + B9}{4}$$

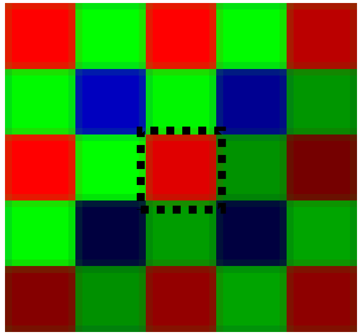


RGB Demoasicing

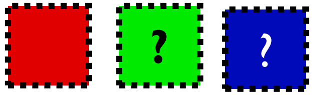
Simple "edge aware" interpolation



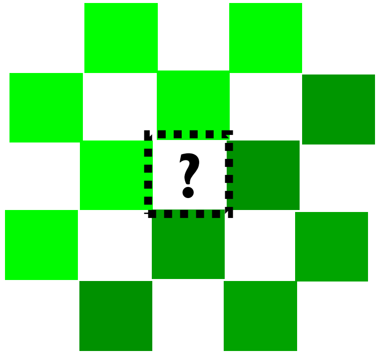
Captured raw-Bayer image



Neighborhood
about red pixel



R G B

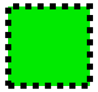


Neighboring
green values

0.8	0.8	0.8	0.4	0.2
0.8	0.9	0.9	0.3	0.2
0.7	0.9	1.0	0.3	0.2
0.5	0.3	0.3	0.2	0.2
0.1	0.2	0.2	0.2	0.2

Weight mask based on red pixel's
similarity to neighboring red values.

Missing green pixel value is computed
as a weighted-interpolation of the
neighboring green values.

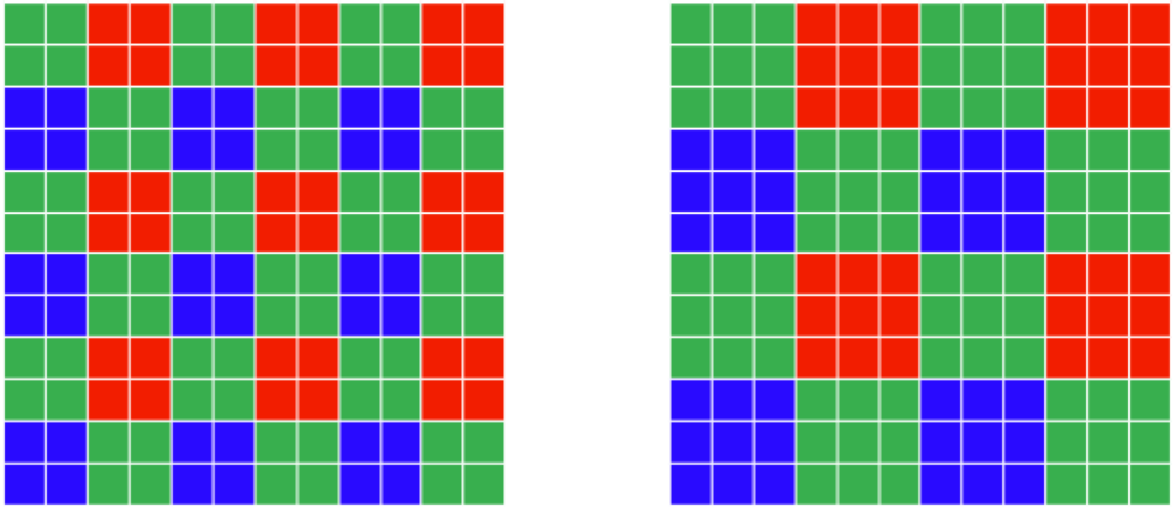


G

Do this procedure also for the blue pixel, B5.

Newer CFA/Bayer patterns

- Newer sensors are starting to use different patterns.
- Quad/Tetra (2x2) and Nona (3x3) are now common on smartphones.
- In low-light situations, the 2x2 or 3x3 layouts are "binned" into a single pixel (a process called binning).



Tetra CFA　　Nona CFA

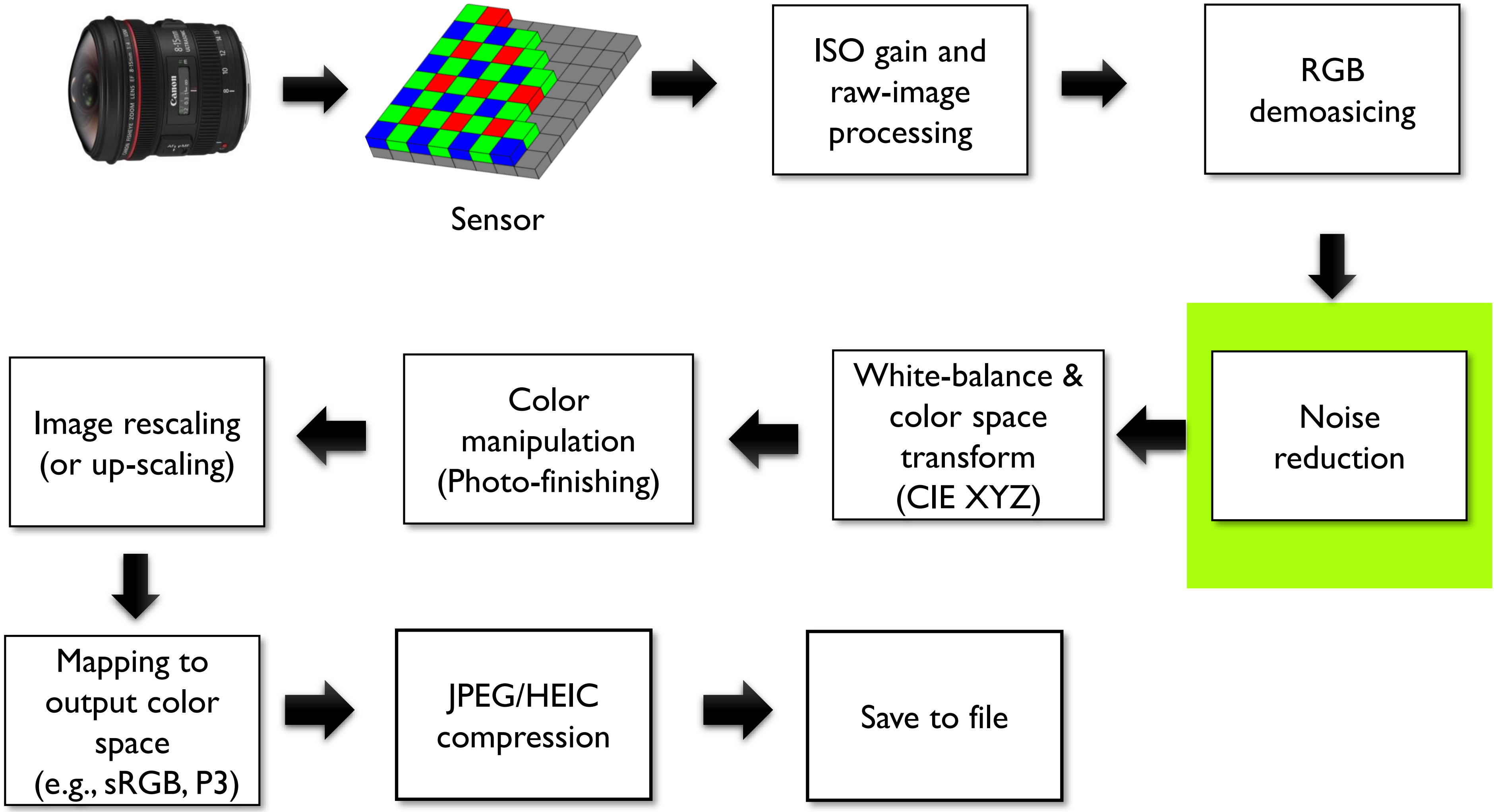
Image: Sang-wook Park and Jong-hyun Kim

RGB Demoasicing

Demosaicing in practice

- The prior examples are illustrative algorithms only
- **Camera IPSs use more complex and proprietary algorithms.**
- Demosaicing can be combined with additional processing
  - Highlight clipping
  - Sharpening
  - Noise reduction
- Demosaicing is an active research area!

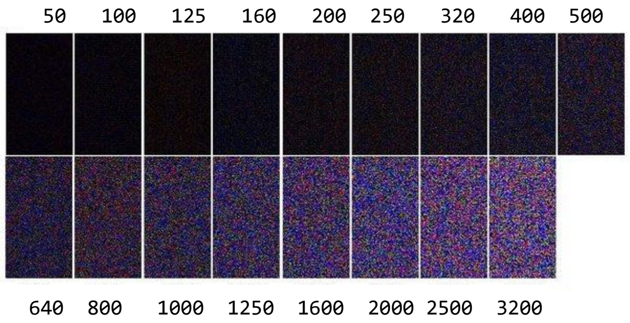
A typical color imaging pipeline



Noise reduction (NR)

- All sensors inherently have noise
- Most cameras apply additional NR after A/D conversion
- A simple method is described in the next slide
- For high-end cameras, it is likely that cameras apply different strategies depending on the ISO settings, e.g. high ISO will result in more noise, so a more aggressive NR could be used
- Smartphone cameras, because the sensor is small, apply aggressive noise reduction.

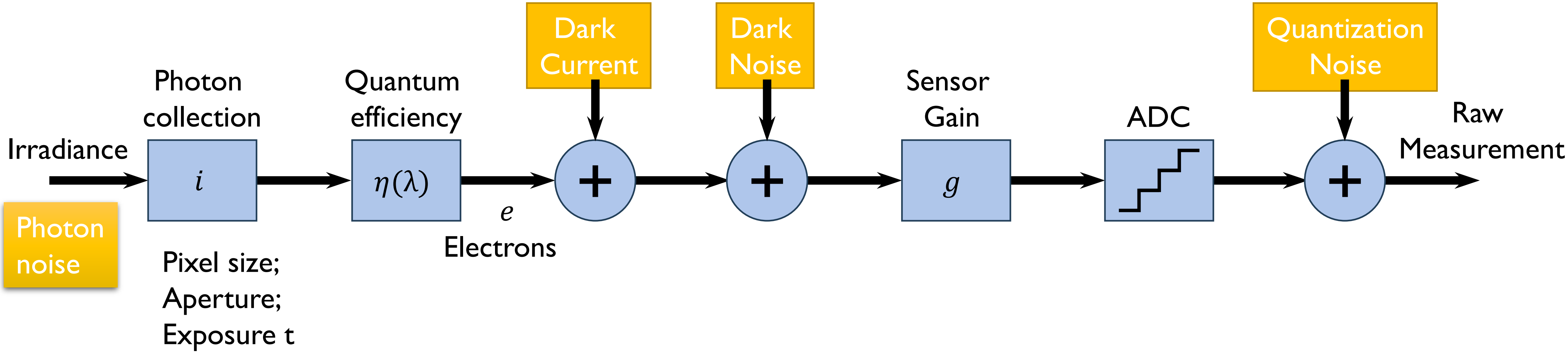
Camera ISO setting and noise



Noise Reduction

Sensor noise model

EMVA 1288 Standard



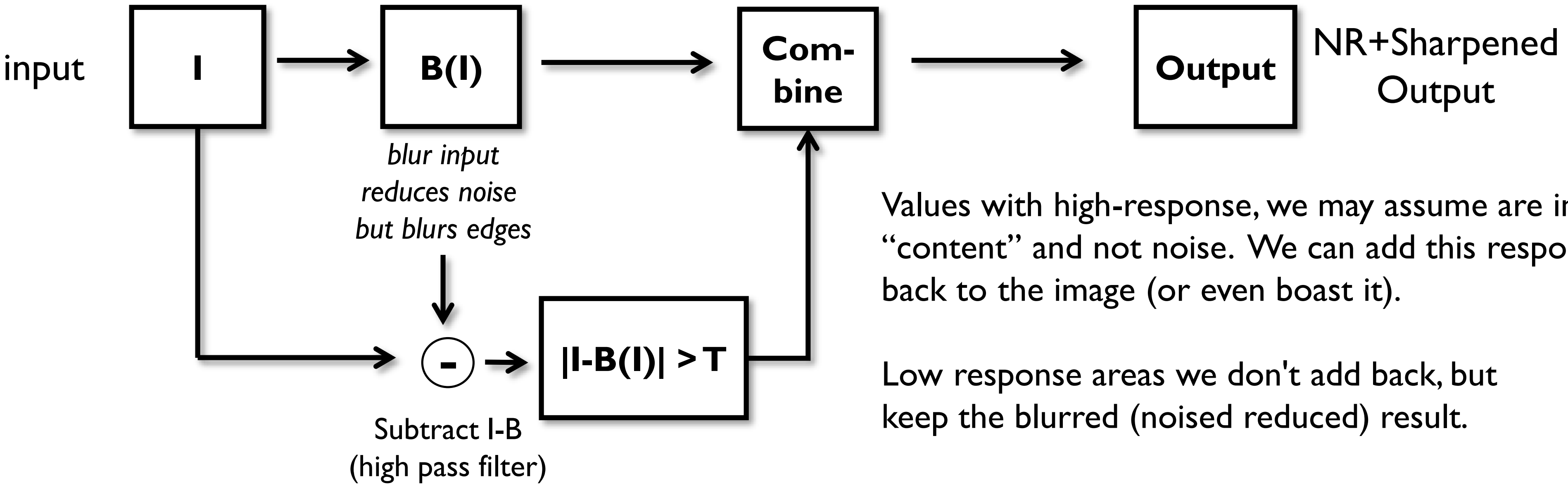
- Two main **sources** of image noise:
  1. The quantum nature of light (photon noise/shot noise); unrelated to the imaging sensor. Follows a Poisson distribution.
  2. Electronic sources associated with the imaging sensor circuitry (dark current and dark noise). Often follows a Normal distribution.
- Gain factor g amplifies noise.

Slide credit: Ali Mosleh

A simple noise reduction approach

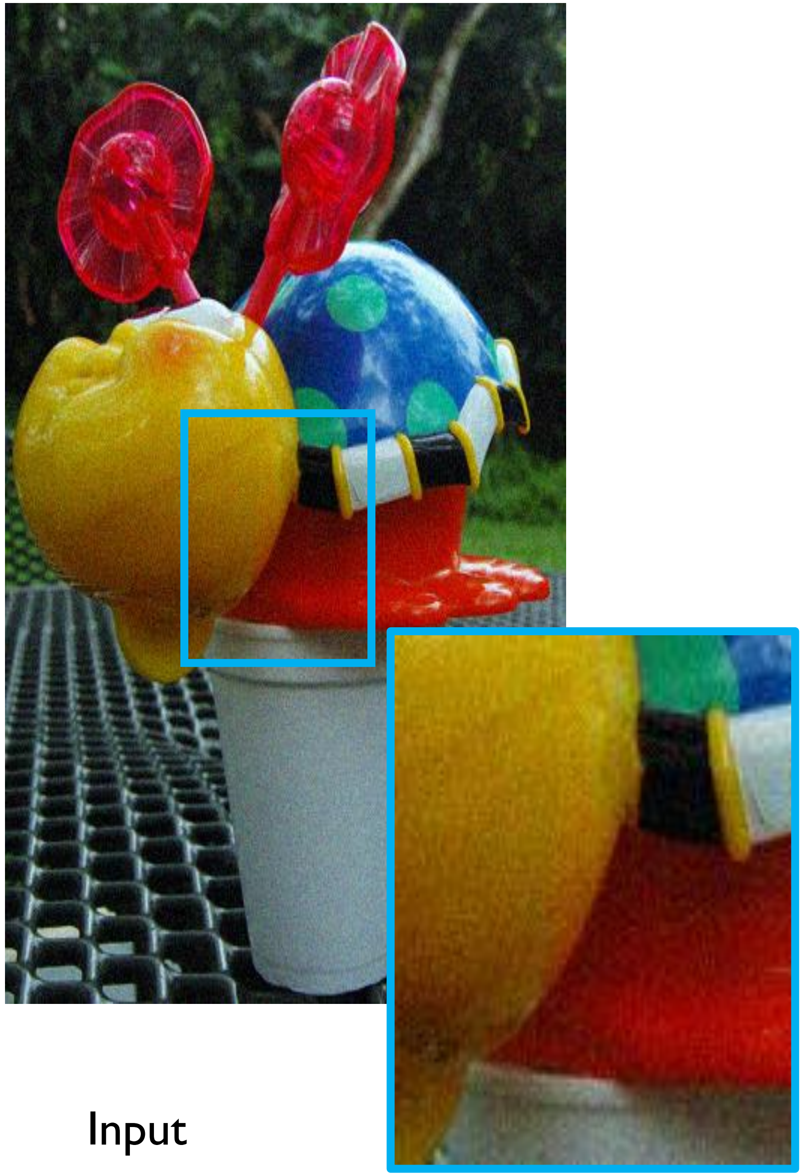
- Blur the image based on the ISO setting (higher ISO = more blur)
- Blurring will reduce noise, but also remove detail.
- Add image detail back for regions that have a high signal. We can even boost some parts of the signal to enhance detail (i.e. "sharpening")

Sketch of the procedure here

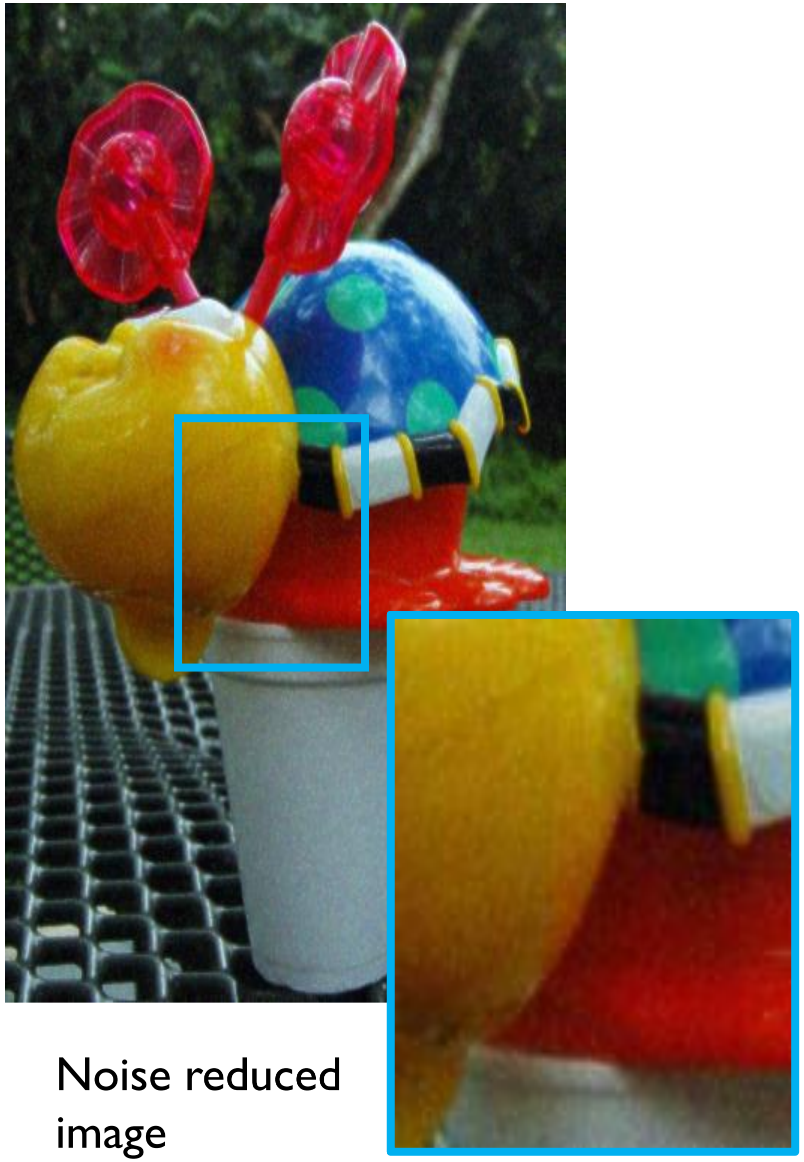


Values with high-response, we may assume are image "content" and not noise. We can add this response back to the image (or even boast it).

Low response areas we don't add back, but keep the blurred (noised reduced) result.

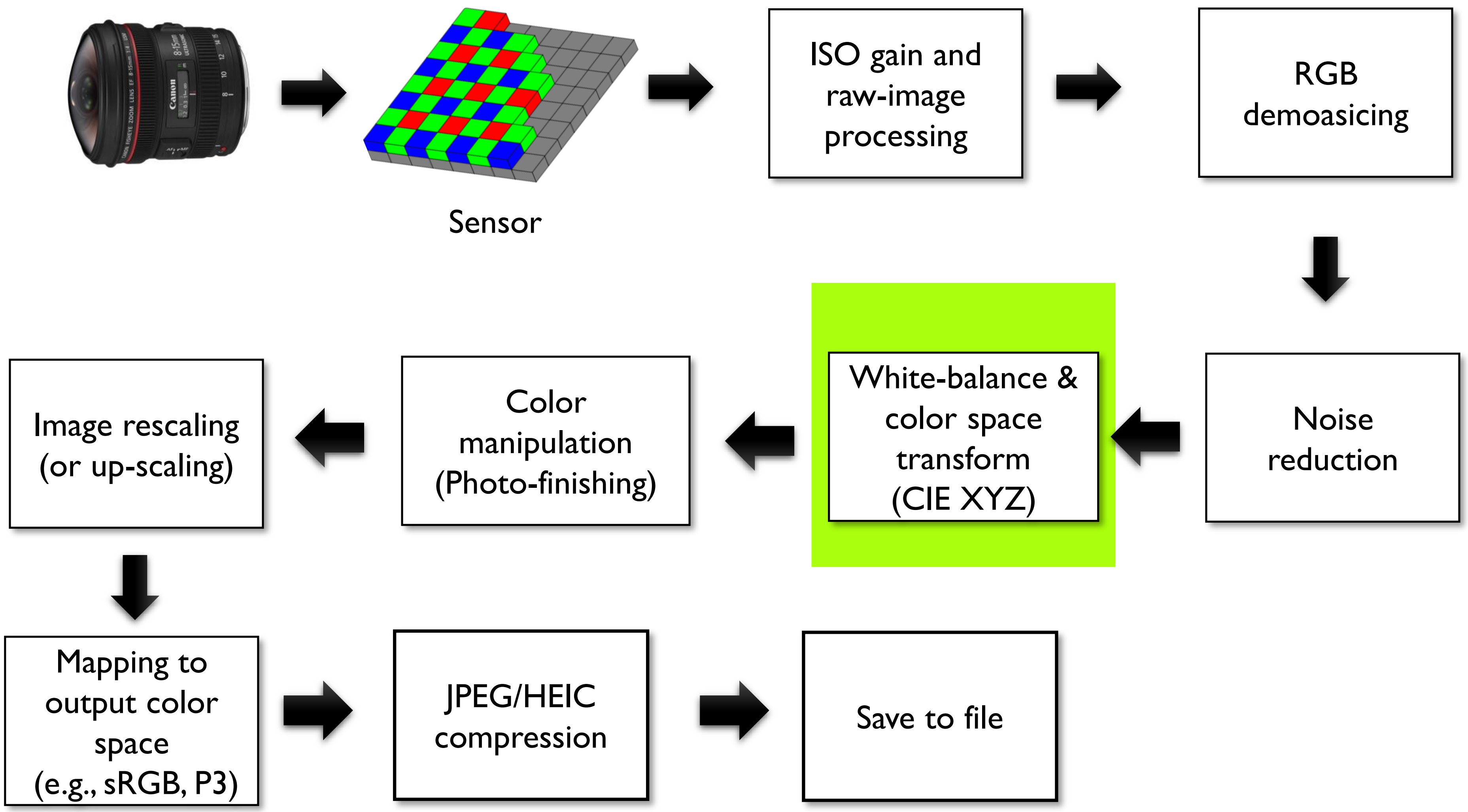


Input



Noise reduced image

A typical color imaging pipeline

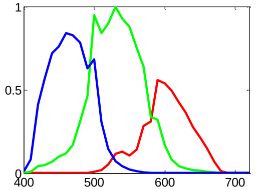


Color mapping/colorimetric stage

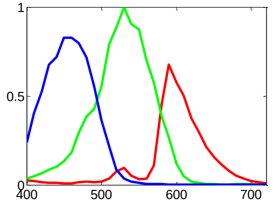
- This step in the IPS converts the sensor raw-RGB values to a device independent color space

Camera sensors have their own spectral response.

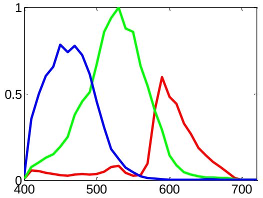
We need to map it into a standard response (CIE XYZ).



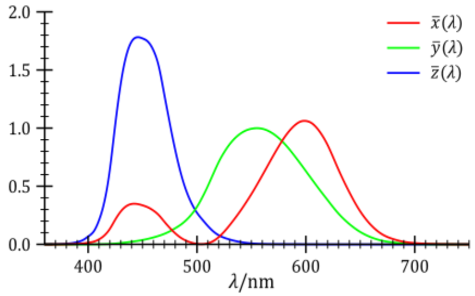
Canon 1D



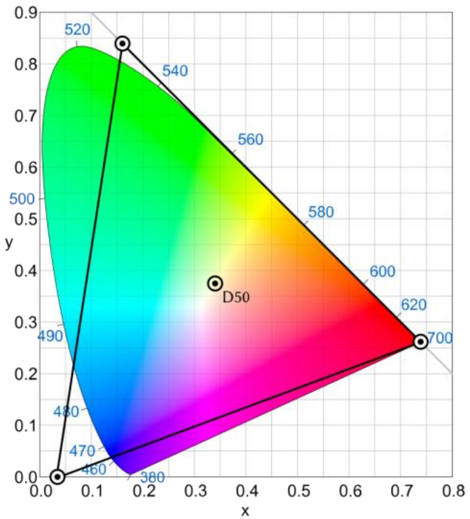
Nikon D40



Sony α57



CIE XYZ



ProPhoto RGB

We will use CIE XYZ in this tutorial, but most cameras use a related space called ProPhoto.

White-Balance & Color Space Transform (CIE XYZ)

Two step procedure

(1) apply a white-balance correction to the raw-RGB values

(2) map the white-balanced raw-RGB values to CIE XYZ

White balance

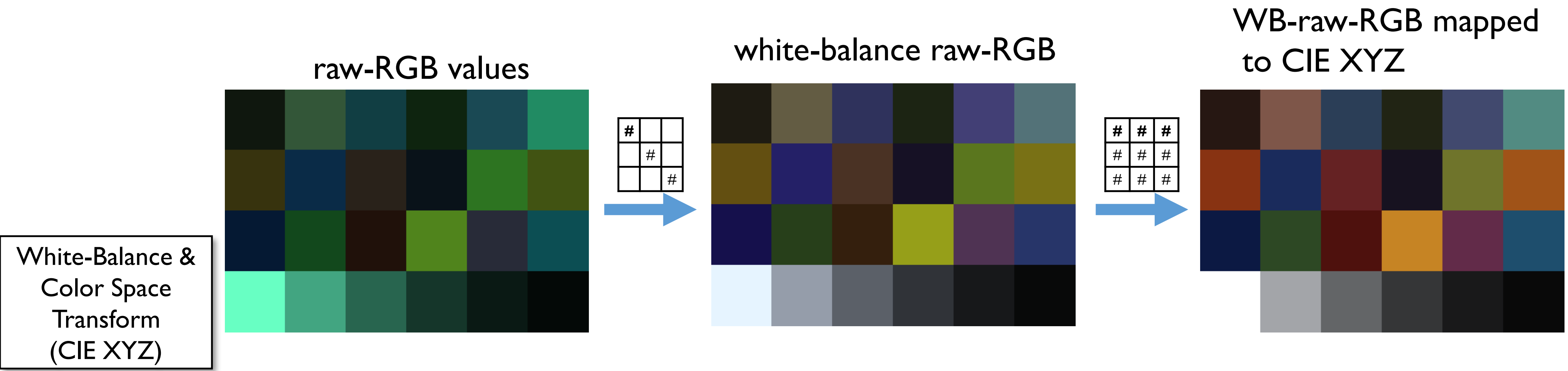
$$\begin{bmatrix} \# & & \\ & \# & \\ & & \# \end{bmatrix}$$

3x3 diagonal matrix

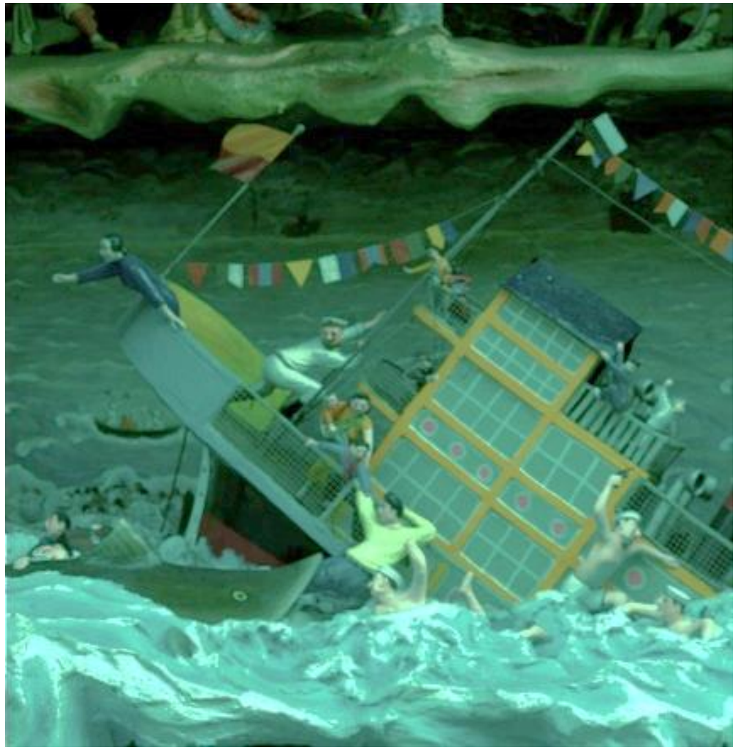
Color space transform (CST)

$$\begin{bmatrix} \# & \# & \# \\ \# & \# & \# \\ \# & \# & \# \end{bmatrix}$$

3x3 full matrix (or polynomial function)

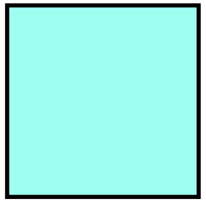


How does white balance (WB) work?

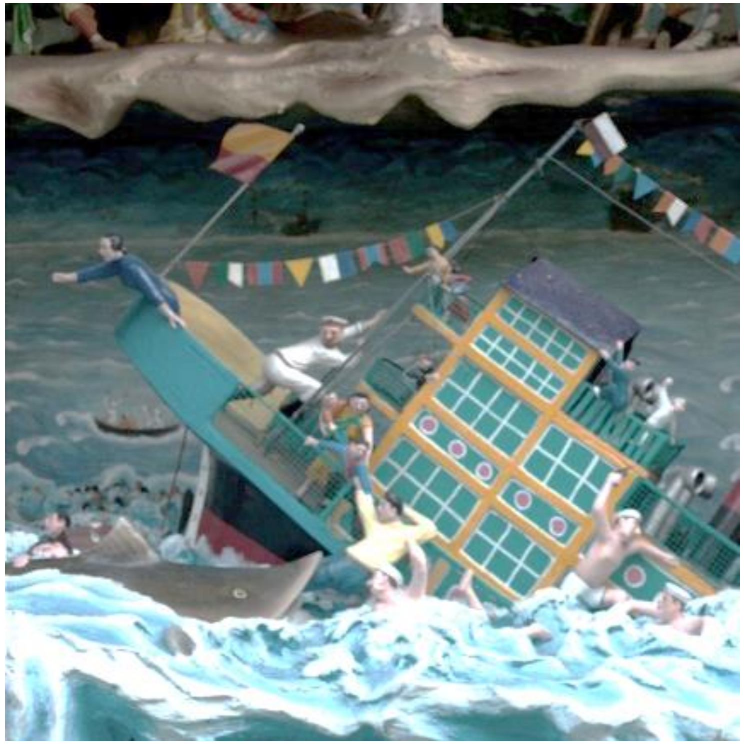


raw-RGB sensor image
(pre-white-balance correction)

Sensor's response to illumination (ℓ)



$$\begin{bmatrix} \ell_r \\ \ell_g \\ \ell_b \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.8 \\ 0.8 \end{bmatrix}$$



"White-balanced" raw-RGB image

White-balance diagonal matrix

$$\begin{bmatrix} r_{wb} \\ g_{wb} \\ b_{wb} \end{bmatrix} = \begin{bmatrix} 1/\ell_r & 0 & 0 \\ 0 & 1/\ell_g & 0 \\ 0 & 0 & 1/\ell_b \end{bmatrix} \begin{bmatrix} r \\ g \\ b \end{bmatrix}$$

White balance
(computational color constancy)

- **The challenging part for white-balance is determining the proper white-balance setting!**
- Users can manually set the white balance
    - Camera specific white-balance matrices for common illuminations
    - These can be manually selected by the user
- Otherwise auto white balance (AWB) is performed
    - In computer vision, we often refer to AWB as "illumination estimation"
    - Since the hard part is trying to determine what the illumination in the scene is.

White-Balance &
Color Space
Transform
(CIE XYZ)

WB manual settings

WB SETTINGS	COLOR TEMPERATURE	LIGHT SOURCES
	10000 - 15000 K	Clear Blue Sky
	6500 - 8000 K	Cloudy Sky / Shade
	6000 - 7000 K	Noon Sunlight
	5500 - 6500 K	Average Daylight
	5000 - 5500 K	Electronic Flash
	4000 - 5000 K	Fluorescent Light
	3000 - 4000 K	Early AM / Late PM
	2500 - 3000 K	Domestic Lightning
	1000 - 2000 K	Candle Flame

Cameras can pre-calibrate their sensor's response for common illuminations.
Typical mapping of WB icons to related color temperature.

Image from ExposureGuide.com

Examples of manual WB matrices

Nikon D7000

Sunny

$$\begin{bmatrix} 2.0273 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & 1.3906 \end{bmatrix}$$

Incandescent

$$\begin{bmatrix} 1.3047 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & 2.2148 \end{bmatrix}$$

Shade

$$\begin{bmatrix} 2.4922 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & 1.1367 \end{bmatrix}$$

Canon 1D

Daylight

$$\begin{bmatrix} 2.0938 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & 1.5020 \end{bmatrix}$$

Tungsten

$$\begin{bmatrix} 1.4511 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & 2.3487 \end{bmatrix}$$

Shade

$$\begin{bmatrix} 2.4628 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & 1.2275 \end{bmatrix}$$

Sony A57K

Daylight

$$\begin{bmatrix} 2.6836 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & 1.5586 \end{bmatrix}$$

Tungsten

$$\begin{bmatrix} 1.6523 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & 2.7422 \end{bmatrix}$$

Shade

$$\begin{bmatrix} 3.1953 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & 1.2891 \end{bmatrix}$$

White-Balance & Color Space Transform (CIE XYZ)

Pre-calibrated white-balance matrices for different brands of cameras.

Auto white balance (AWB)

- If manual white balance is not used, then an AWB algorithm is performed.
- AWB must determine the sensor's raw-RGB response to the scene illumination from an arbitrary image.
- AWB is not easy and this remains an open research problem.

White-Balance & Color Space Transform (CIE XYZ)

AWB is not easy

raw-RGB input image before white-balance

Given an arbitrary raw-RGB image, determine what is the camera's response to the illumination.

The idea is that something that is *white** is a natural reflector of the scene's illuminations SPD.

So, if we can identify what is "white" in the raw-RGB image, we are observing the sensor's RGB response to the illumination.

* It doesn’t have to be "white", but grey – sometimes we call these scene points "achromatic" or "neutral" regions.

AWB: "Gray world" algorithm

- This methods assumes that the average reflectance of a scene is achromatic (i.e. gray)
  - Gray is just the white point not at its brightest, so it serves as an estimate of the illuminant
  - This means that image average should have equal energy, i.e. R=G=B
- Based on this assumption, the algorithm adjusts the input average to be gray as follows:

First, estimate the average response:

$$R_{avg} = \frac{1}{Nr}\sum Rsensor(\mathrm{r}) \quad G_{avg} = \frac{1}{Ng}\sum Gsensor(\mathrm{g}) \quad B_{avg} = \frac{1}{Nb}\sum Bsensor(\mathrm{b})$$

r = red pixels values, g=green pixels values, b =blue pixels values
Nr = # of red pixels, Ng = # of green pixels, Nb = # blue pixels

Note: # of pixel per channel may be different if white balance is applied to the RAW image before demosaicing. Some pipelines may also transform into another colorspace, e.g. LMS, to perform the white-balance procedure.

AWB: "Gray world" algorithm

- Based on the image average R/G/B value, white balance can be expressed as a matrix as:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} G_{avg}/R_{avg} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & G_{avg}/B_{avg} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

White-balanced sensor RGB

Sensor RGB

Matrix scales each channel by its average and then normalizes to the green channel average.

White-Balance & Color Space Transform (CIE XYZ)

AWB: "White patch" algorithm

- This methods assumes that "highlights" (bright spots) represent specular reflections of the illuminant
  - This means that maximum R, G, B values are a good estimate of the white point
- Based on this assumption, the algorithm works as follows:

$$R_{max} = max(Rsensor(\mathrm{r})) \qquad G_{max} = max(\ Gsensor(\mathrm{g})) \quad B_{max} = max(\ Bsensor(\mathrm{b}))$$

r = red pixels values, g=green pixels values, b =blue pixels values

White-Balance & Color Space Transform (CIE XYZ)

AWB: "White patch" algorithm

- Based on RGB max, white balance can be expressed as a matrix as:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} G_{max}/R_{max} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & G_{max}/B_{max} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

White-balanced sensor raw-RGB

Sensor raw-RGB

Matrix scales each channel by its maximum value and then normalizes to the green channel's maximum.

White-Balance & Color Space Transform (CIE XYZ)

AWB example

Input

Gray World

White Patch

White-Balance &
Color Space
Transform
(CIE XYZ)

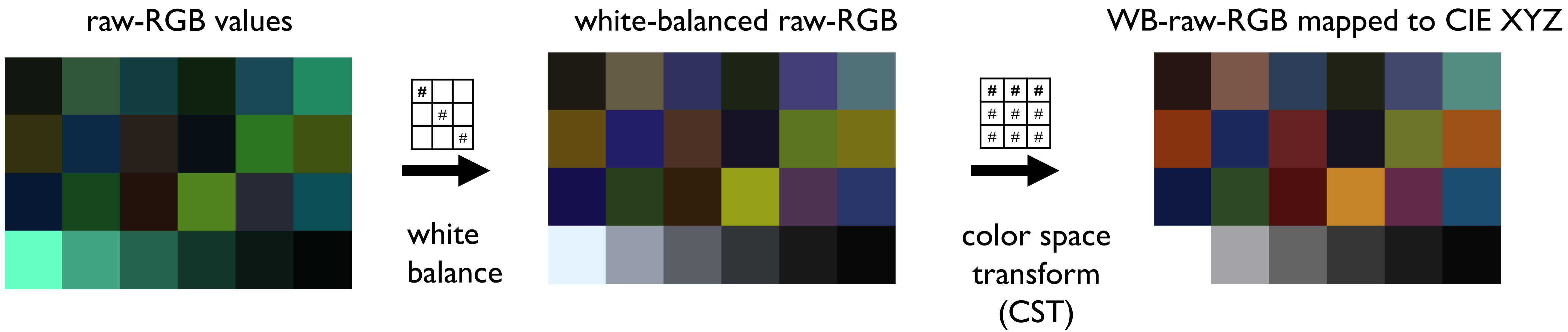
Better AWB methods

- Gray world and white patch are *very basic* algorithms.
  - These both tend to fail when the image is dominated by large regions of a single color (e.g. a sky image).
- **There are many AWB methods in the literature.**
- Camera ISPs often still use simple algorithms with lots of "tuning" . . .

Color space transform – part 2

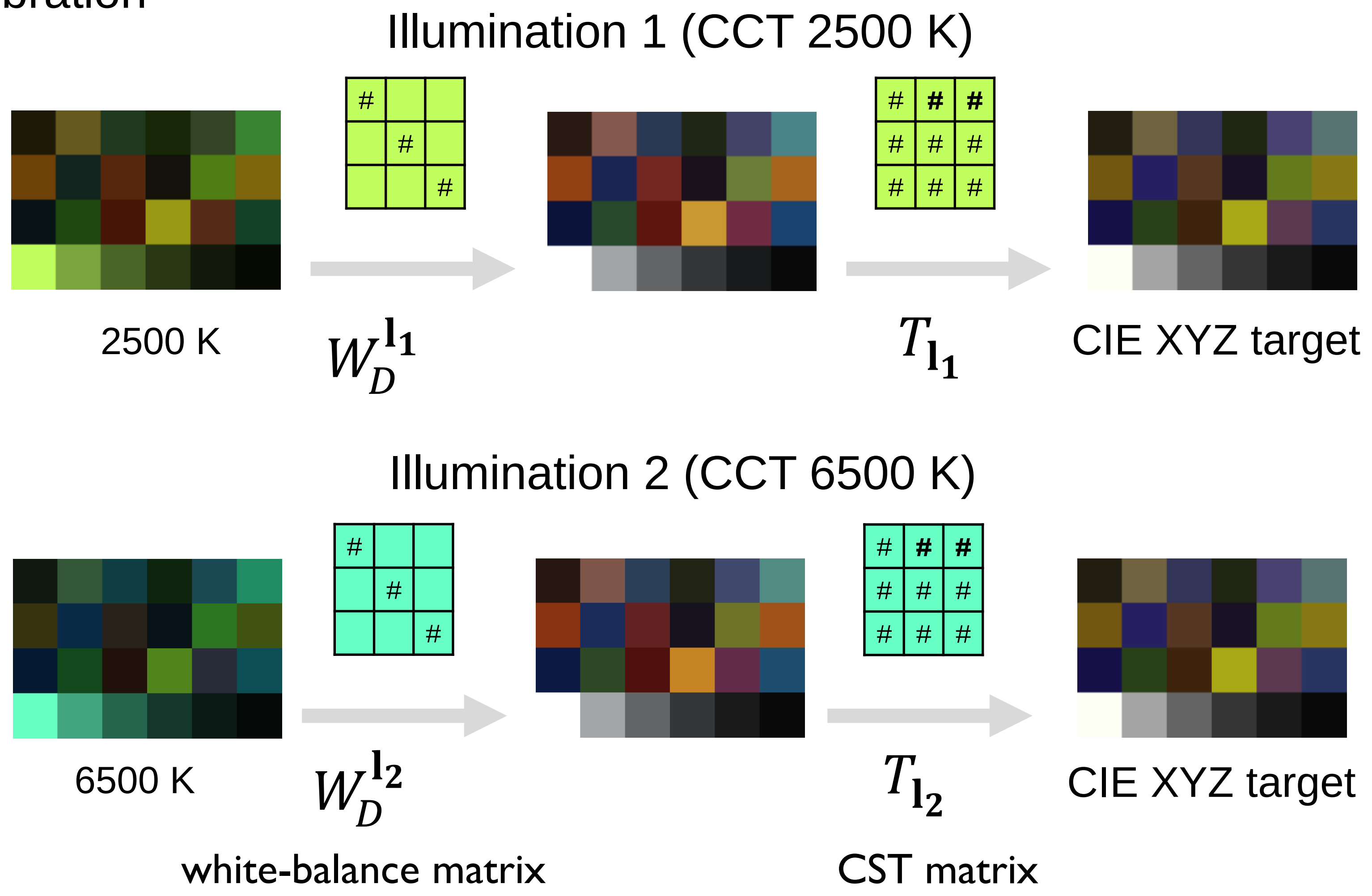
- Process used on cameras involves interpolation from factory presets.
- The need for interpolation is related to white-balance only approximating true color constancy.

Color space transform is applied after the white balance. In fact, the matrix we use to perform the CST is based on the white-balance CCT.



Color space transform (1/3)

Factory pre-calibration



White-Balance & Color Space Transform (CIE XYZ)

CST matrices ($T_{\mathbf{l}_1}$ and $T_{\mathbf{l}_2}$) are calibrated for two different illuminations (l1 and l2). Depending on the temperature of the white-balance, we use the corresponding CST.

Lightboxes for calibration

Lightboxes are used to perform this calibration under different illuminations.
Lightboxes are able to reproduce standard illuminants (e.g., D65, incandescent, fluorescent, etc)

GTI lightbox

X-rite lightbox

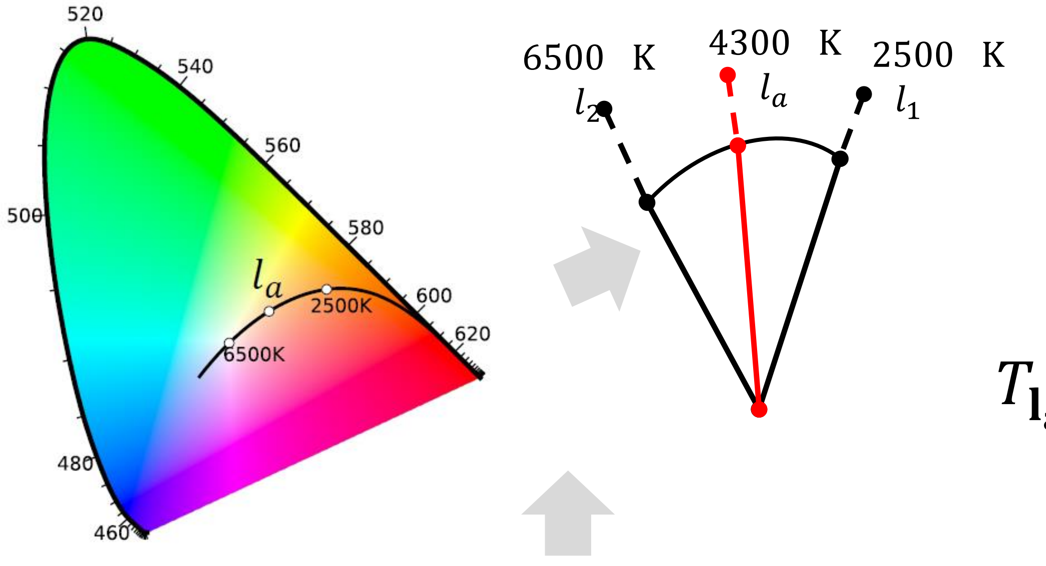
TELELUMEN

The Recording and Playback of Light

Telelumen can replace the
light source in a lightbox
to allow tunable SPDs.

Color space transform (2/3)

Interpolation process



$$g = \frac{\mathrm{CCT}_{\mathbf{l_a}}^{-1} - \mathrm{CCT}_{\mathbf{l_2}}^{-1}}{\mathrm{CCT}_{\mathbf{l_1}}^{-1} - \mathrm{CCT}_{\mathbf{l_2}}^{-1}}$$

$$T_{\mathbf{l_a}} = gT_{\mathbf{l_1}} + (1-g)T_{\mathbf{l_2}}$$

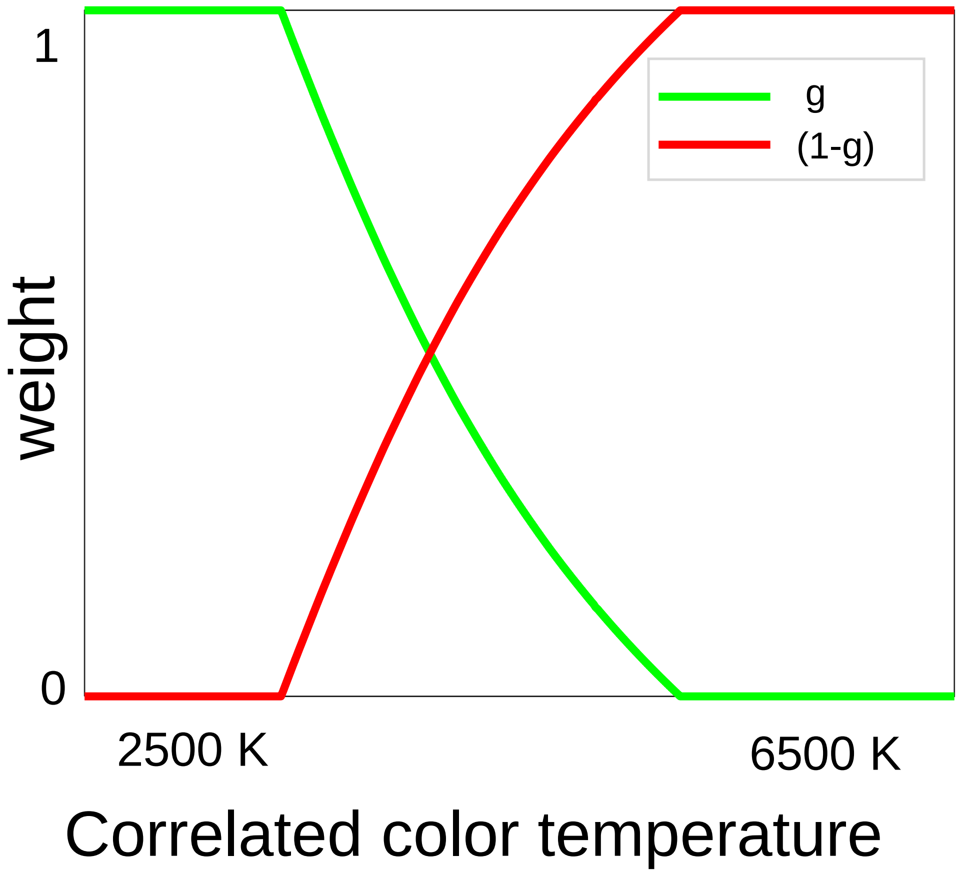
Given a new illumination (la) and its estimated correlated color temperature (CCT), we construct a CST matrix by blending the two factory pre-calibrated matrices.

Color space transform (3/3)

Weighting functions

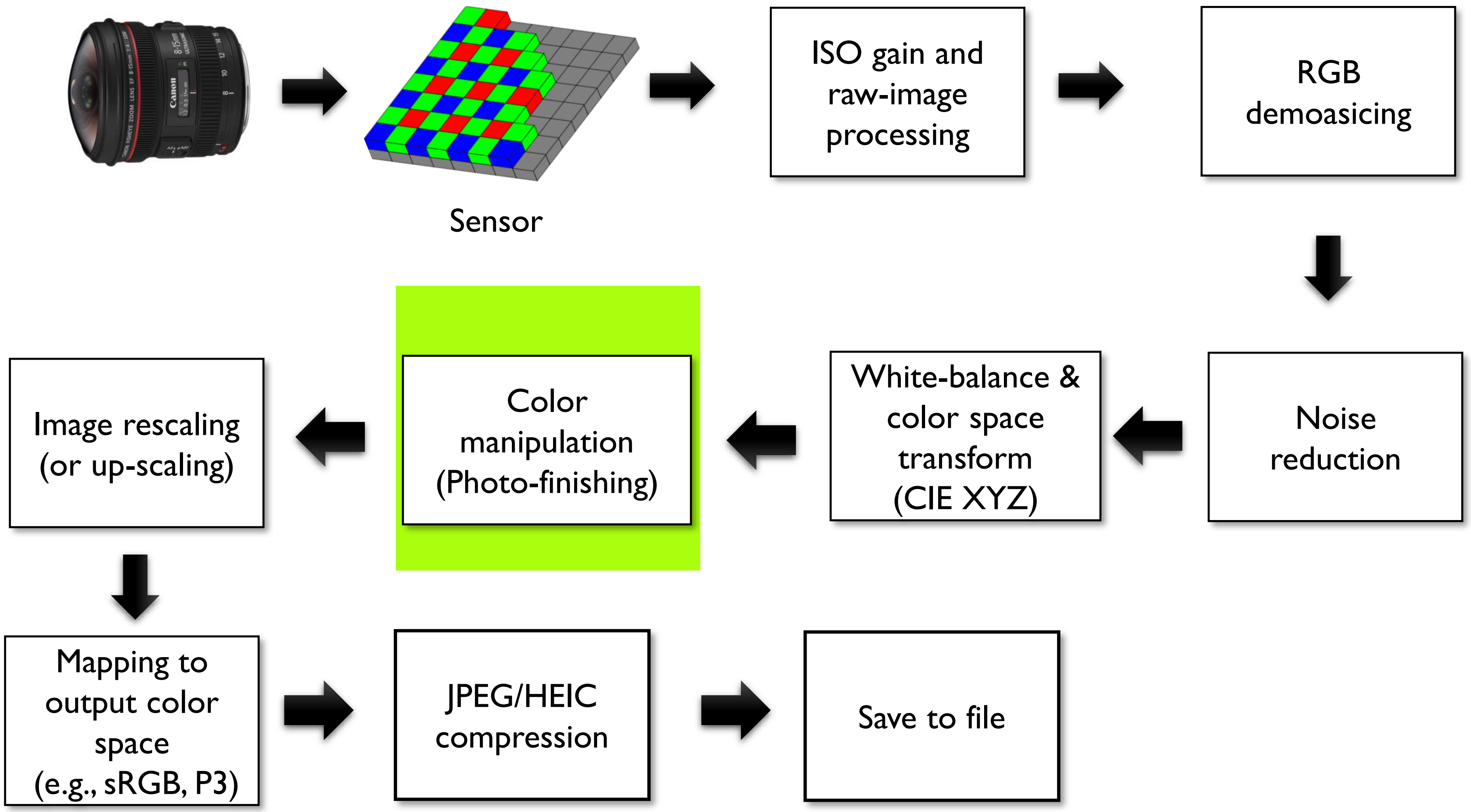
$$g = \frac{\mathrm{CCT}_{\mathbf{l_a}}^{-1} - \mathrm{CCT}_{\mathbf{l_2}}^{-1}}{\mathrm{CCT}_{\mathbf{l_1}}^{-1} - \mathrm{CCT}_{\mathbf{l_2}}^{-1}}$$

$$T_{\mathbf{l_a}} = gT_{\mathbf{l_1}} + (1-g)T_{\mathbf{l_2}}$$



White-Balance & Color Space Transform (CIE XYZ)

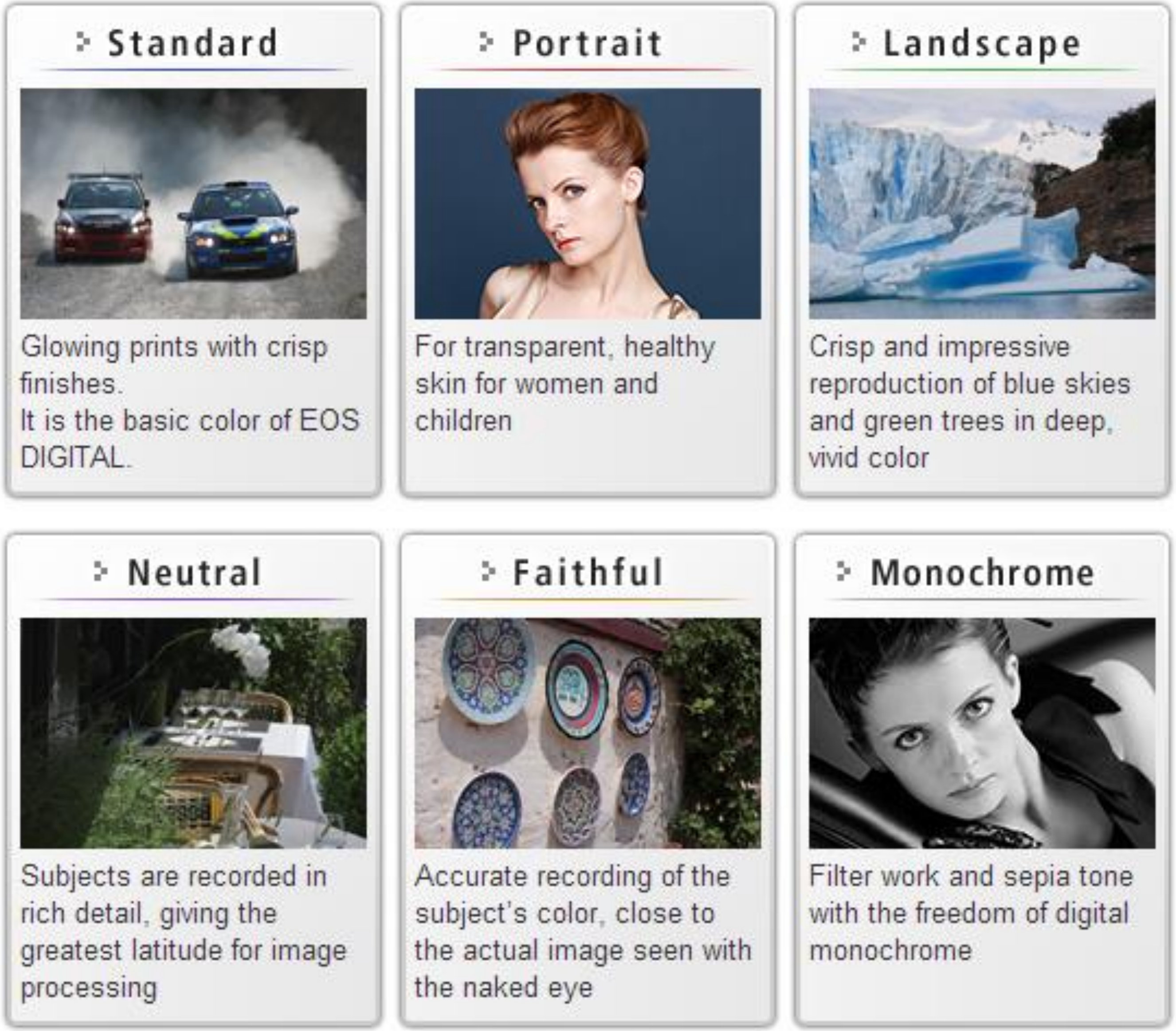
A typical color imaging pipeline



Color manipulation

- This is the stage where a camera applies its "secret sauce" to make the images look good.
- This procedure is called by many names:
  - Color manipulation
  - Photo-finishing
  - Color rendering or selective color rendering
  - Yuv processing engine
- DSLR will often allow the user to select various photo-finishing styles.
- Smartphones often compute this per-image.
- Photo-finishing may also tied to geographical regions!

DSLR "picture" styles



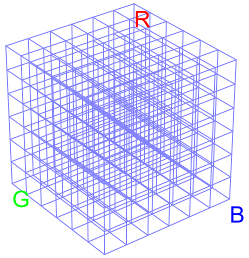
From Canon's user manual

Picture styles

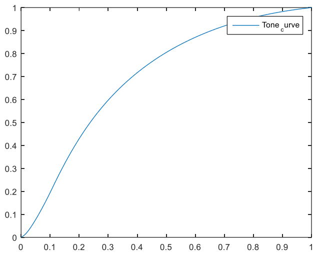
Example of four different picture styles from Nikon
This image is the **same** raw-RGB image processed in four different ways.

Color
Manipulation
(Photo-finishing)

Nonlinear color manipulation



3D Look up table (LUT)



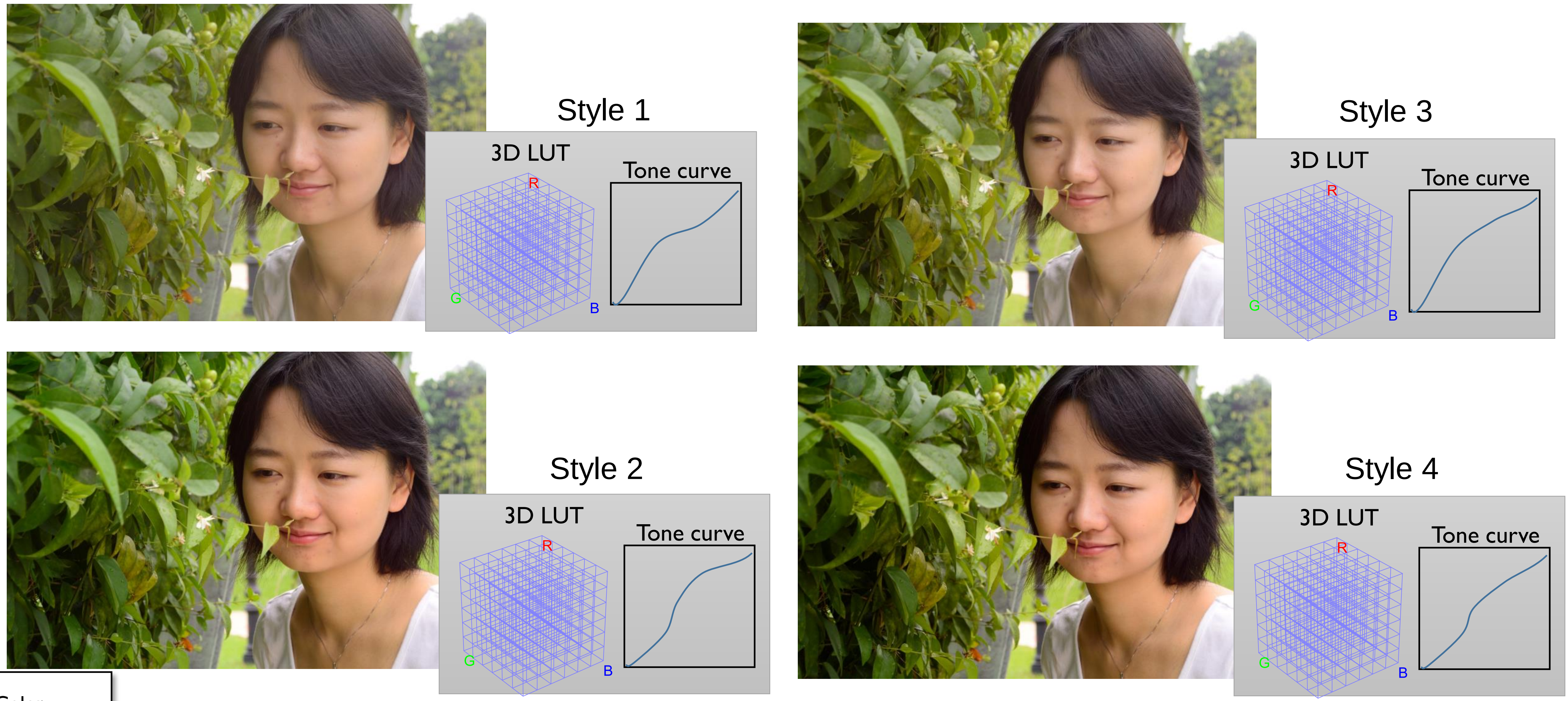
1D Tone Curve

Color manipulation can be implemented using a 3D look up table (LUT) and a 1D LUT tone-curve.
The 3D LUT table acts like a 3D function: $f(R,G,B) \rightarrow R',G',B'$
The 1D LUT table is applied per channel: $g(R) \rightarrow R', g(G) \rightarrow G', g(B) \rightarrow B'$

The 3D and 1D LUT can change based on picture style.

Color Manipulation (Photo-finishing)

Picture styles

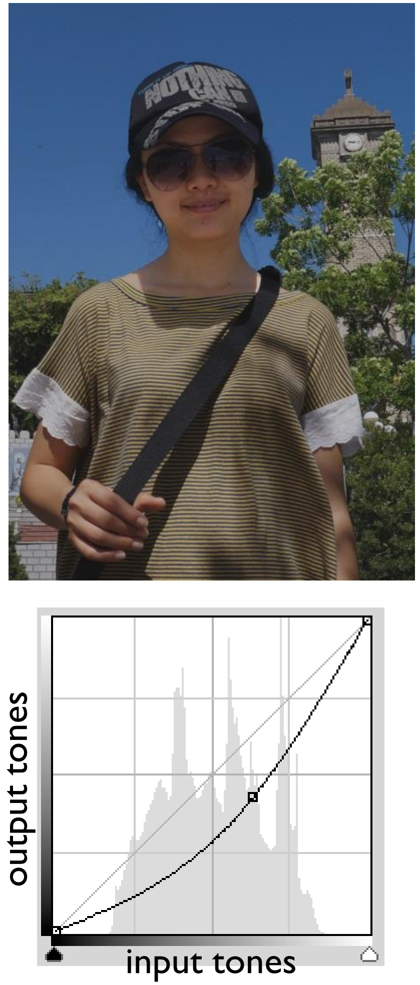


Color Manipulation (Photo-finishing)

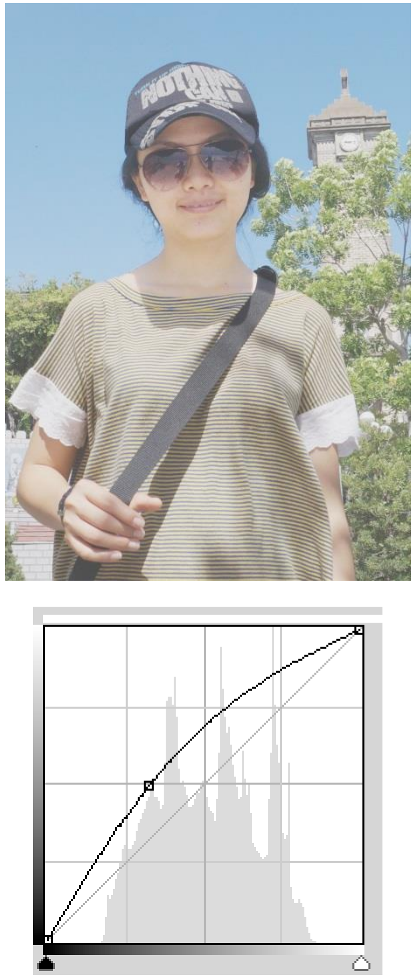
Each style has its own 3D LUT and 1D LUT.

Global tone map example (1D LUT)

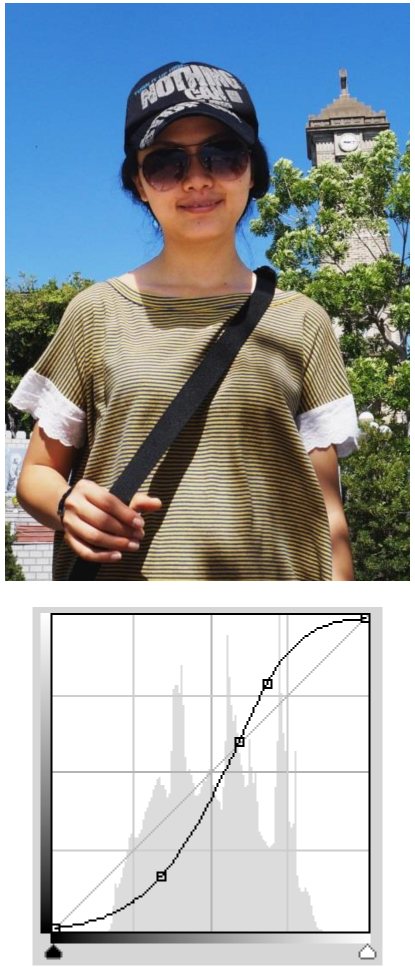
Input



Darkening the image

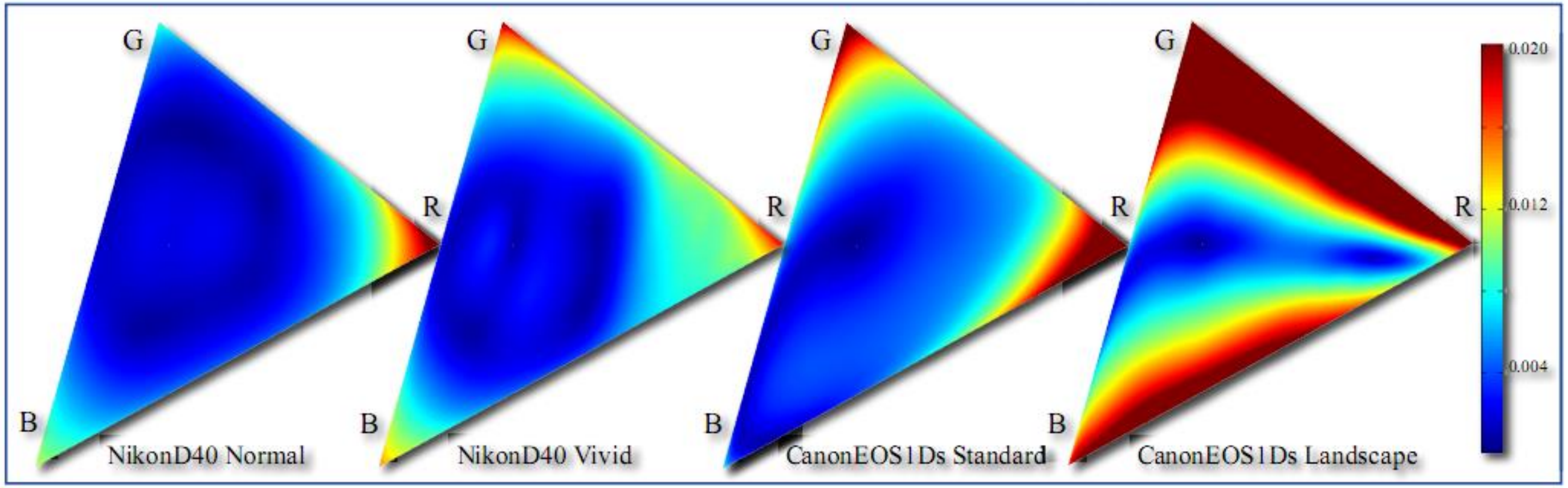


Brightening the image

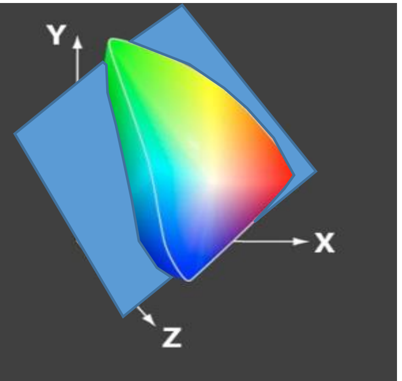


Enhancing contrast (called an S-curve)

3D LUT color manipulation visualization



Visualization as a **displacement map** of a *slice*
of the 3D LUT mapping, warping an input and output value

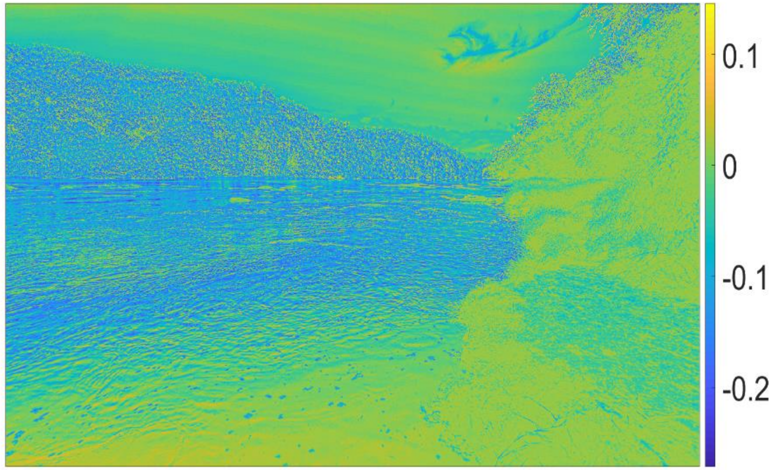


Local tone mapping (LTM)

Global tone-mapping
Camera mode - Manual

Local tone-mapping
Camera mode - Auto

NOTE: On many cameras, esp smartphones, a local tone map is applied as part of the photo-finishing. This helps bring out highlights in the image.



Difference map between image before and after LTM

Color
Manipulation
(Photo-finishing)

Selective color manipulation

- "Select" colors can be manipulated, especially skin tone.
- Sometimes called preferred color correction (PCC)

Selected color regions
can be manipulated.

Image: Gromovataya

Color and Imaging Conference (CIC) papers

Examples of papers addressing preferred skin color.

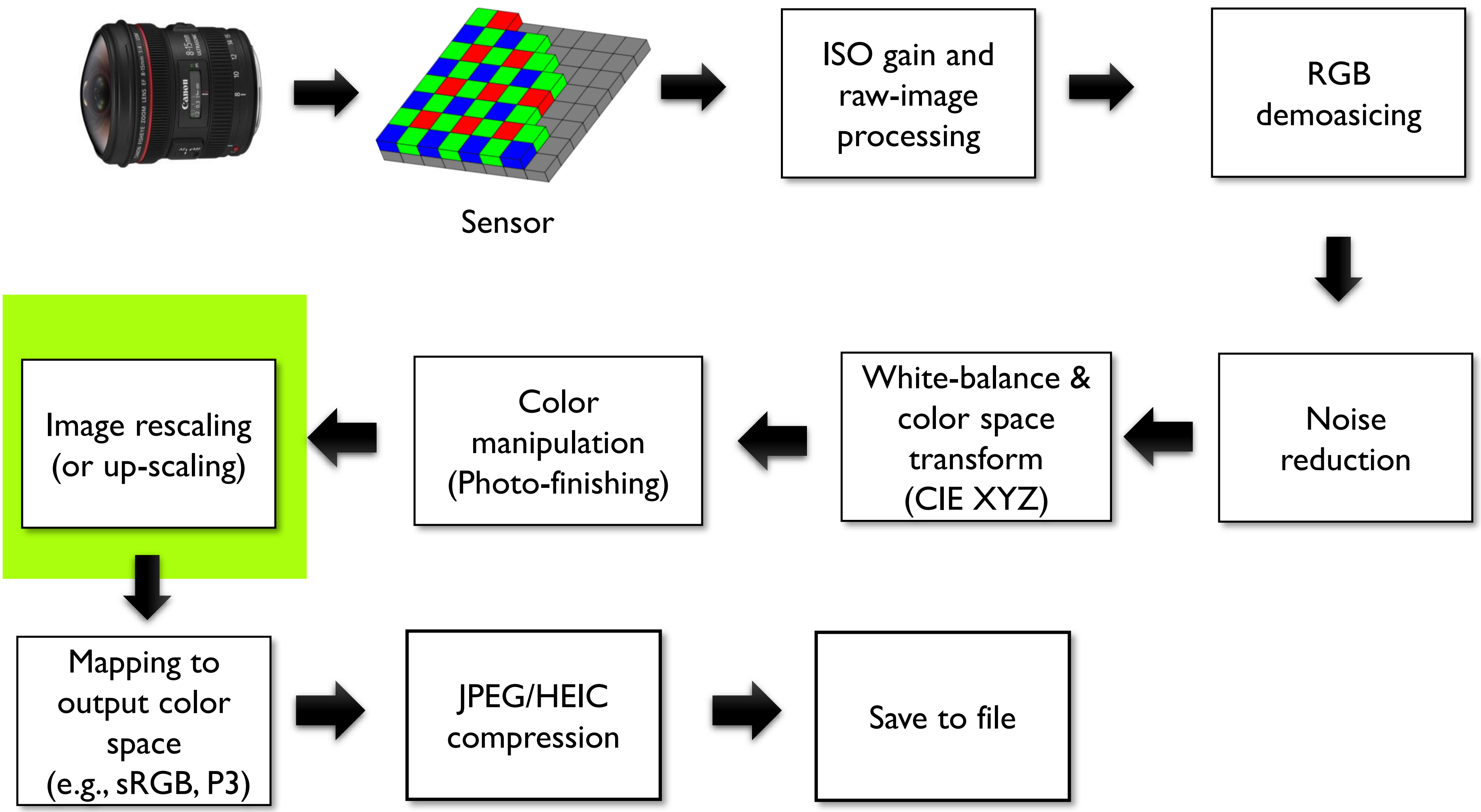
Investigation of Effect of Skin Tone to Facial Attractiveness, *Yan Lu[1], Jie Yang[1], Kaida Xiao[1], Michael Pointer[1], Changjun Li[2], and Sophie Wuerger[3]; [1]University of Leeds (UK), [2]University of Science and Technology Liaoning (China) and [3]University of Liverpool (UK)*

Preferred Skin Colours Observed by Three Ethnic Groups under different Ambient Lighting Conditions, *Mingkai Cao, Ming Ronnier Luo, Rui Peng, Yuechen Zhu, and Xiaoxuan Liu, Zhejiang University, and Guoxiang Liu, Huawei Technologies Co, Ltd. (China)*

Preferred Skin Reproduction Centres for Different Skin Groups, *Rui Peng, Ming Ronnier Luo, Mingkai Cao, Yuechen Zhu, and Xiaoxuan Liu, State Key Laboratory of Modern Optical Instrumentation, and Guoxiang Liu, Hisilicon (China)*

Are We Alike? Skin Color Perception in Portrait Image and AR-based Humanoid Emoji, *Yuchun Yan and Hyeon-Jeong Suk, KAIST (South Korea)*

Typical color imaging pipeline



Re-scaling image and sRGB conversion

Often, the entire image is processed by the ISP and then rescaled for the view-finder or to fit the requested output resolution.

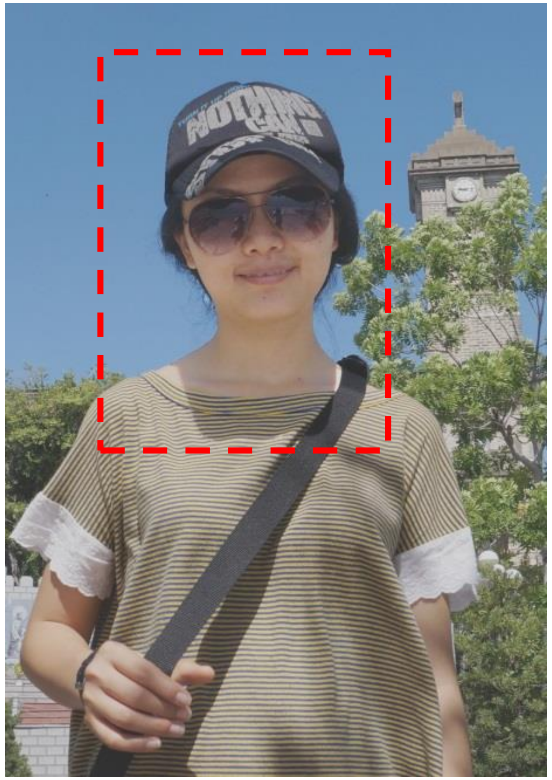
Full-size image

Rescaled for view-finder

Rescaled for preferred output size

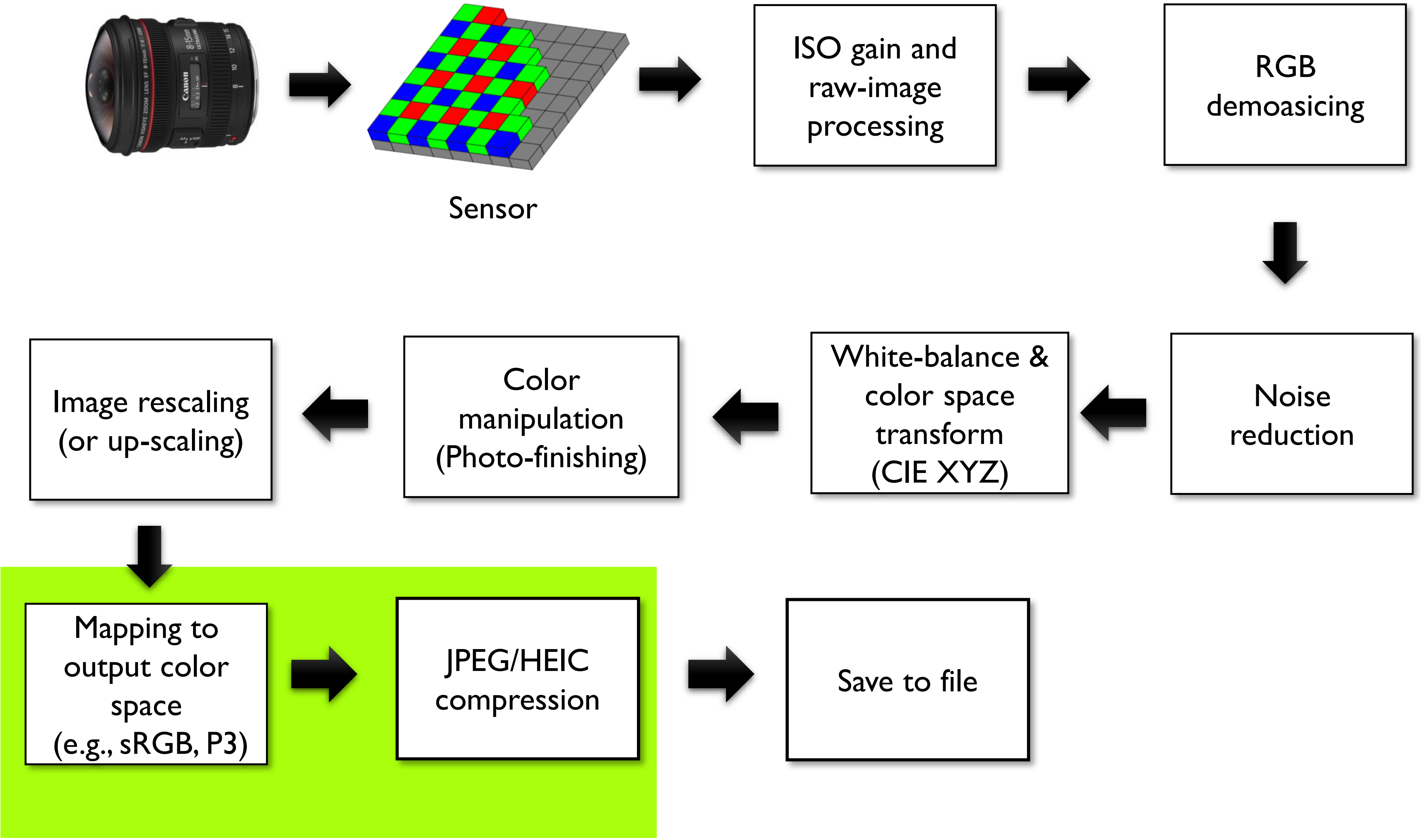
Rescale can also be "digital zoom"

Full frame



Digital zoom
(super res)

Typical color imaging pipeline



Final sRGB conversion (or other color space)

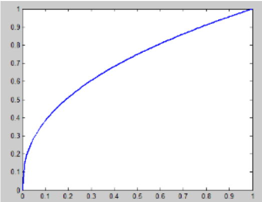
- Map from *photo-finished* CIE XYZ image to sRGB
- Apply the sRGB $(2.2)^{-1}$ gamma encoding

Photo-finished CIE XYZ

Covert to linear sRGB

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 3.2404542 & -1.5371385 & 0.4985314 \\ -0.9692660 & 1.8760108 & 0.0415560 \\ 0.0556434 & -0.2040259 & 1.0572252 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

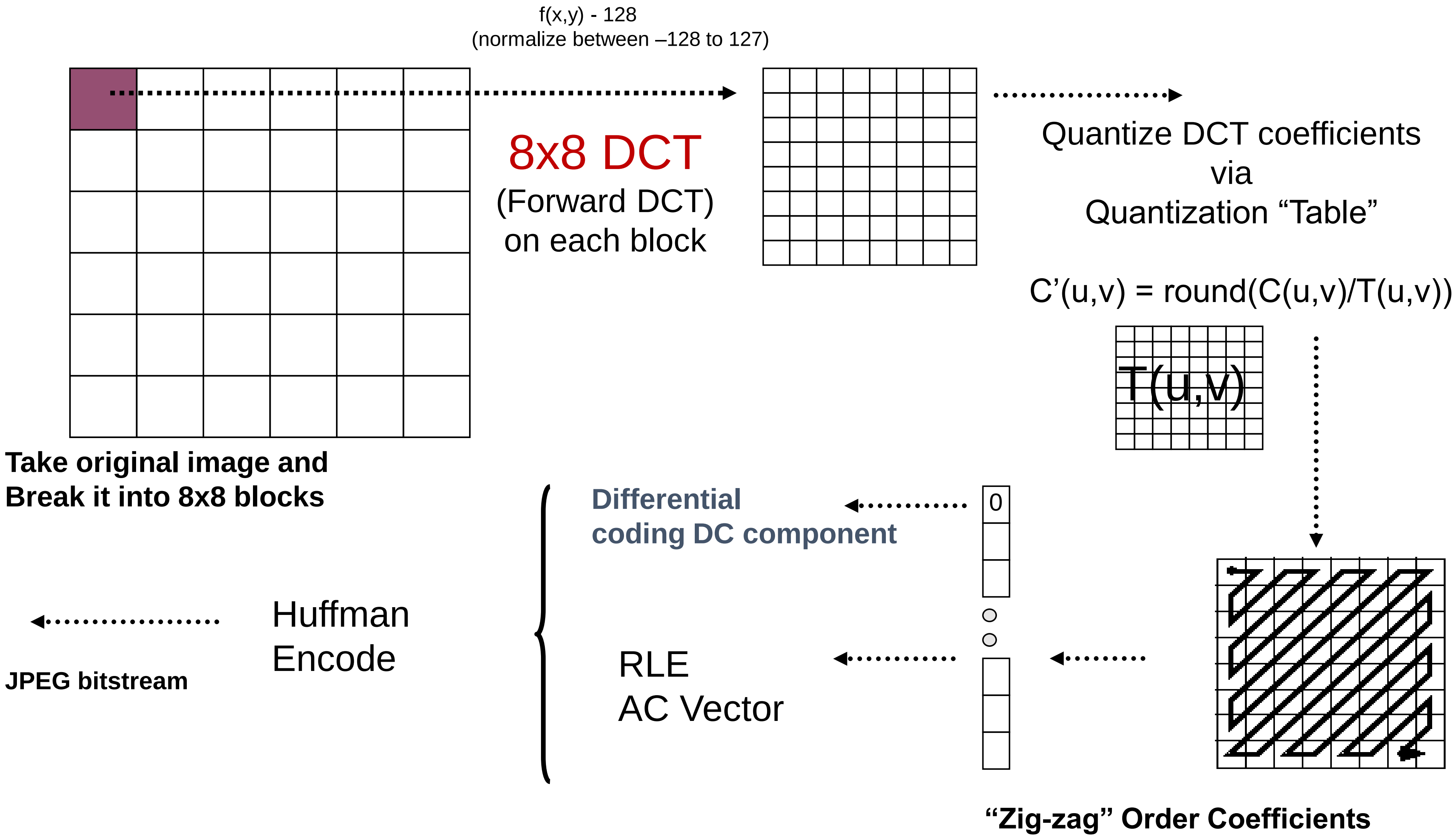
Apply sRGB gamma



Map to sRGB output

sRGB is known as an "output-referred" or "display-referred" color space. It is intended for use with display devices.

JPEG compression scheme



JPEG applies almost every compression trick known.
1) Transform coding, 2) psychovisual (loss), 3) Run-length-encoding (RLE), 4) Difference coding, and Huffman.

JPEG quality

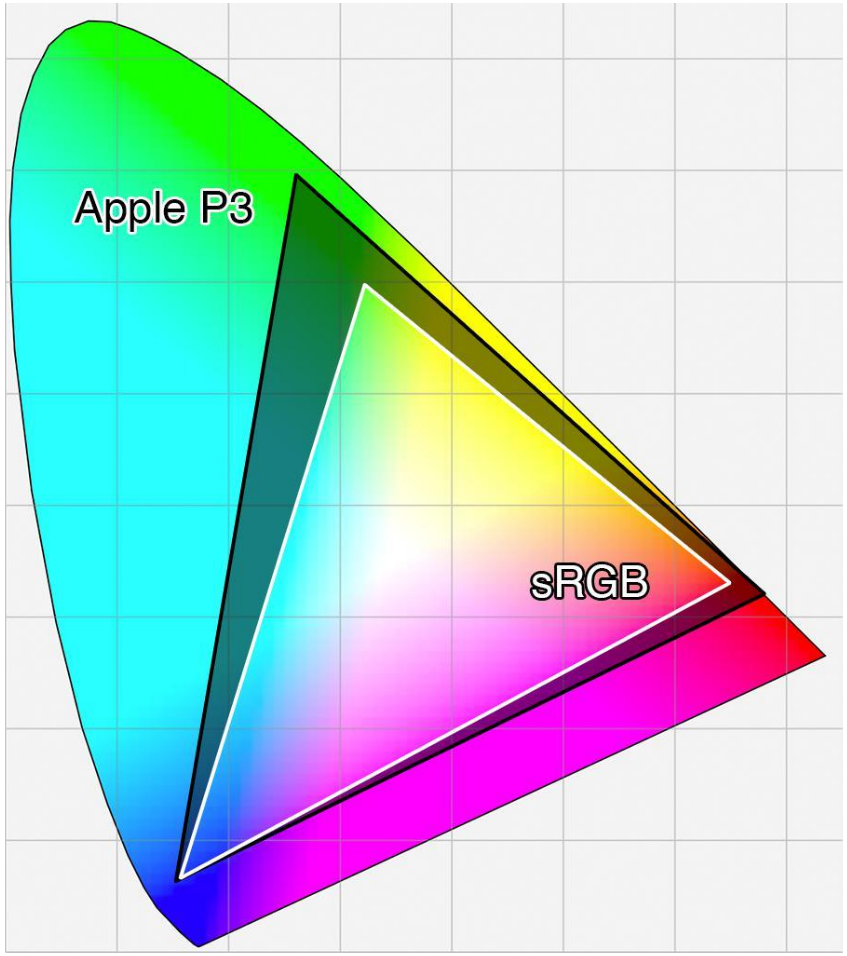
- The amount of quantization applied on the DCT coefficients amounts to a "quality" factor
  - More quantization = better compression (smaller file size)
  - More quantization = lower quality
- Cameras generally allow a range that you can select

Image from nphotomag.com

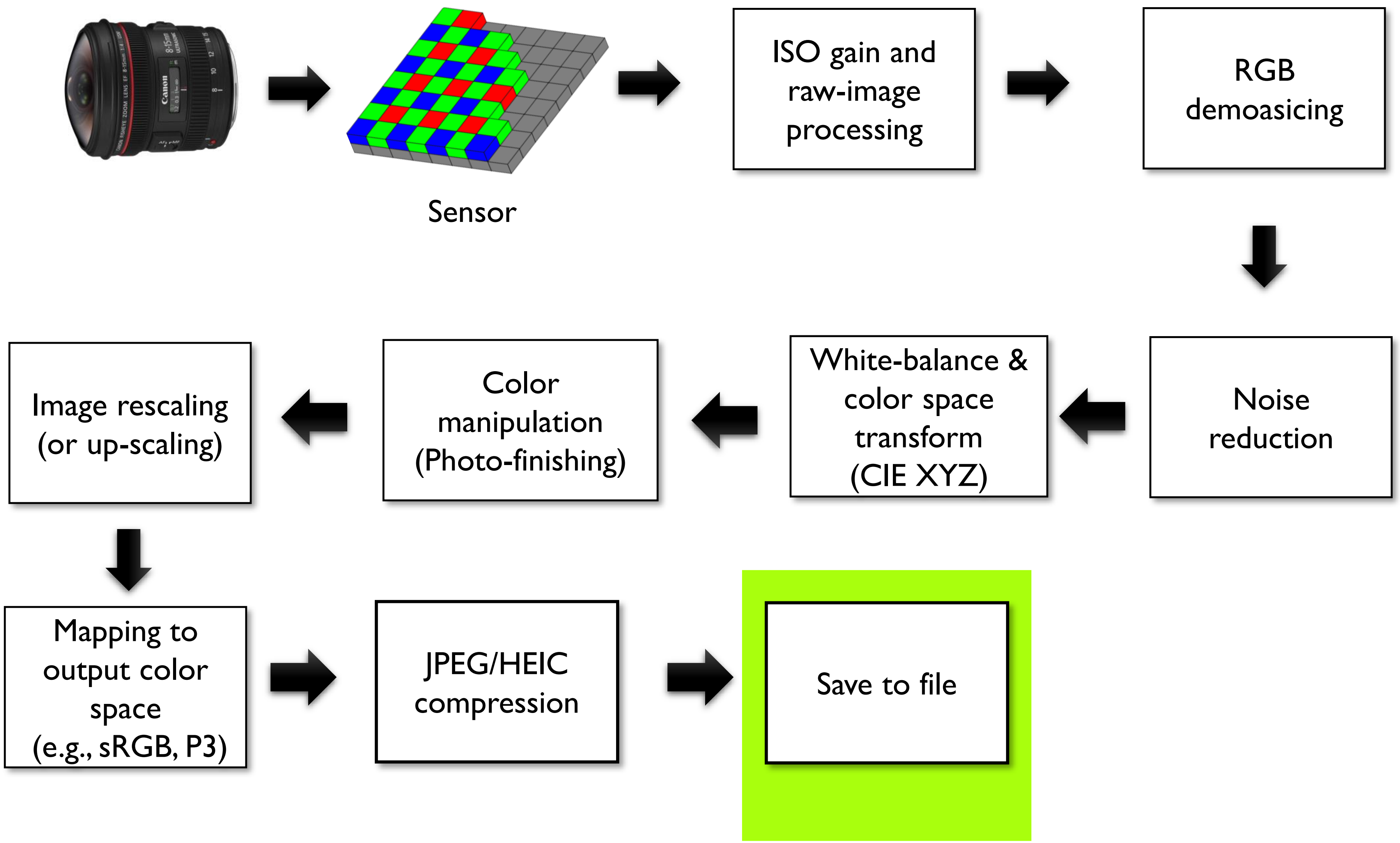
JPEG Compression

Note: sRGB/JPEG is slowly being replaced

- sRGB was developed for monitors in the 1990s – it is an old standard.
- High Efficient Image Encoding (HEIC)
  - Better compression than JPEG
- Apple iPhone has started to use HEIC to *replace JPEG*
- HEIC supports multiple color spaces. Apple uses Display P3 – a variation on a Digitial Cinema Initiative P3 space.
- The P3 gamut is 25% wider than sRGB
- There is also a gamma encoding similar to sRGB.

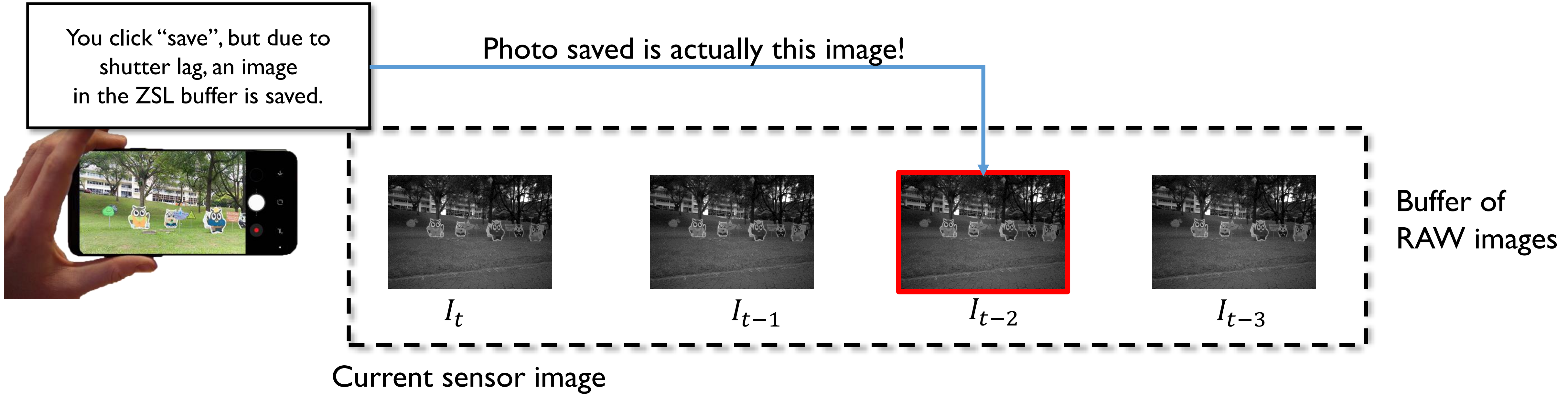


Typical color imaging pipeline

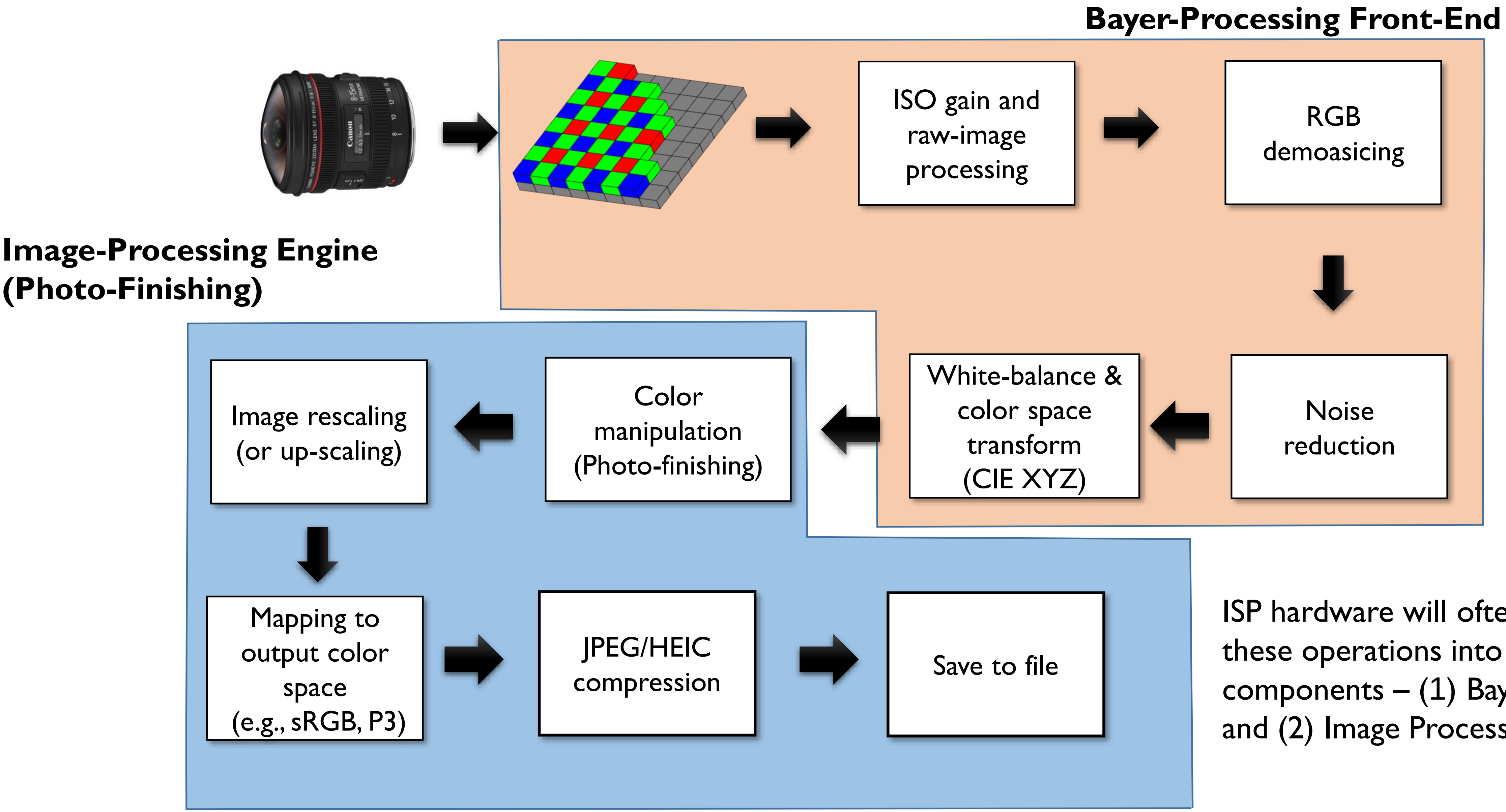


Circular buffer for ZSL and multi-frame processing

- Many ISPs store the last few RAW images in memory (i.e., a circular buffer).
- These images can be used for many purposes (image stabilization, temporal noise reduction, etc.).
- Another purpose is to ensure "Zero Shutter Lag" (ZSL).
  - There is often a delay between when a person presses the "capture" and the camera capturing the image. This is called "shutter lag".
  - So, a previous image in the buffer can saved instead.



ISP organization



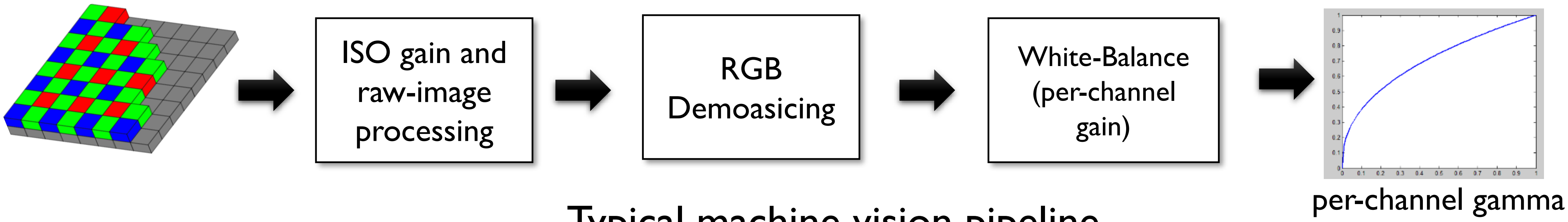
ISP hardware will often divide these operations into two components – (1) Bayer-Processing and (2) Image Processing.

Pipeline comments

- **Again, it is important to stress that the exact steps mentioned in these notes only serve as a guide to what takes place in a camera**
- Smartphone camera pipelines are more complex.
- Note: for the different camera makes/models, the operations could be performed in a different order and in different ways (e.g., combining sharpening with demosaicing).

What about machine vision cameras?

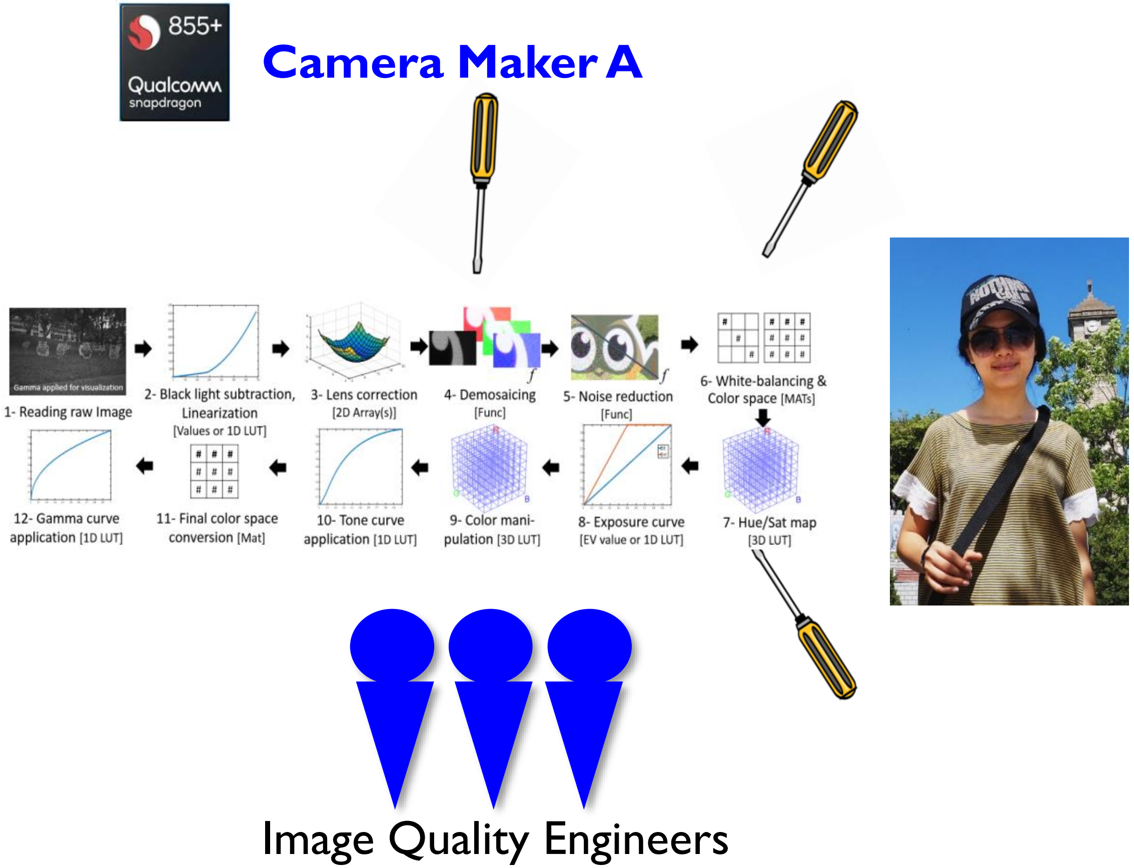
- Some industrial/machine vision cameras provide minimal ISP processing
- For example, some will only perform white-balance and apply a gamma to the raw-RGB values.
- This means the output is in a camera-specific color space.

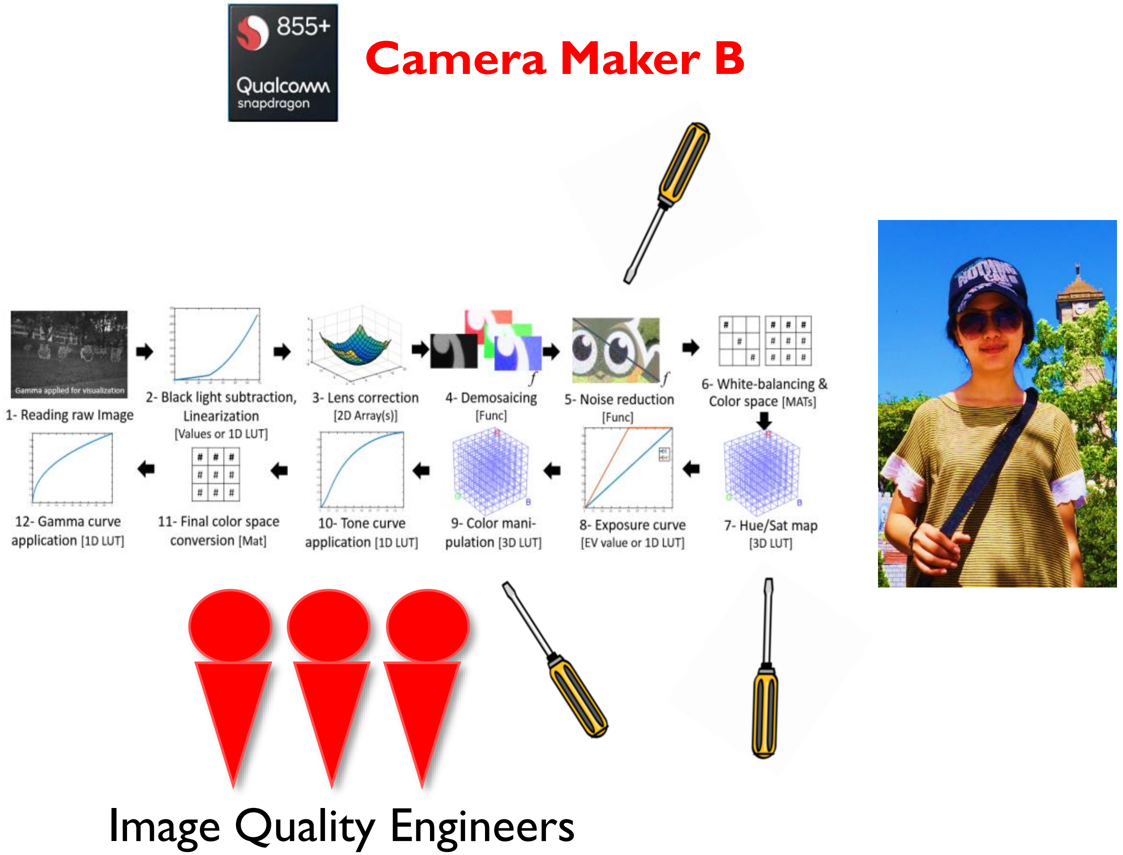


Typical machine vision pipeline

Point grey grasshopper camera.

IPS Tuning



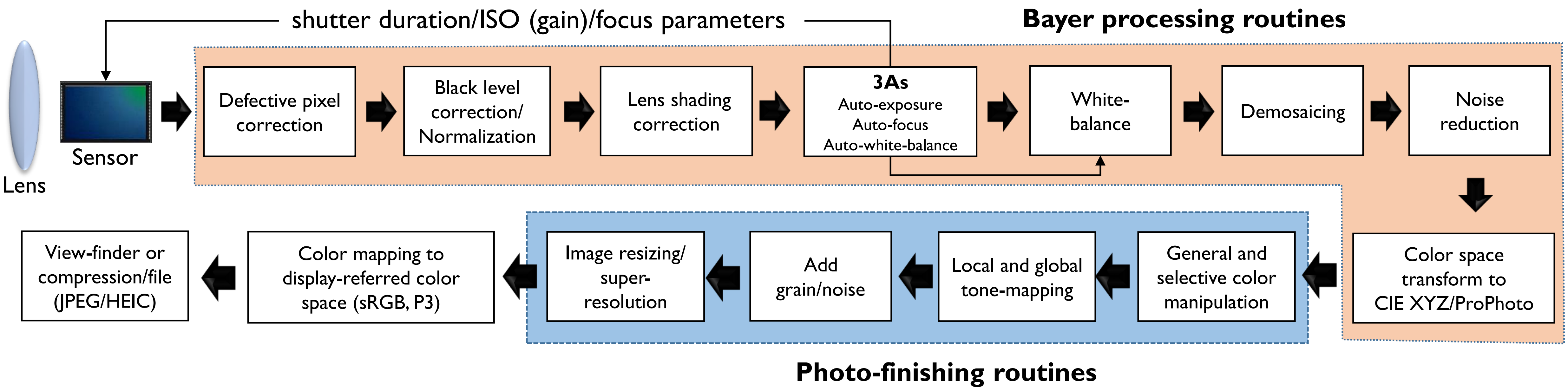


The algorithms on an ISP are often predefined. Camera engineers can "tune" the algorithm parameters to produce the output they want. The "tuning" of the ISP is a labor-intensive procedure.

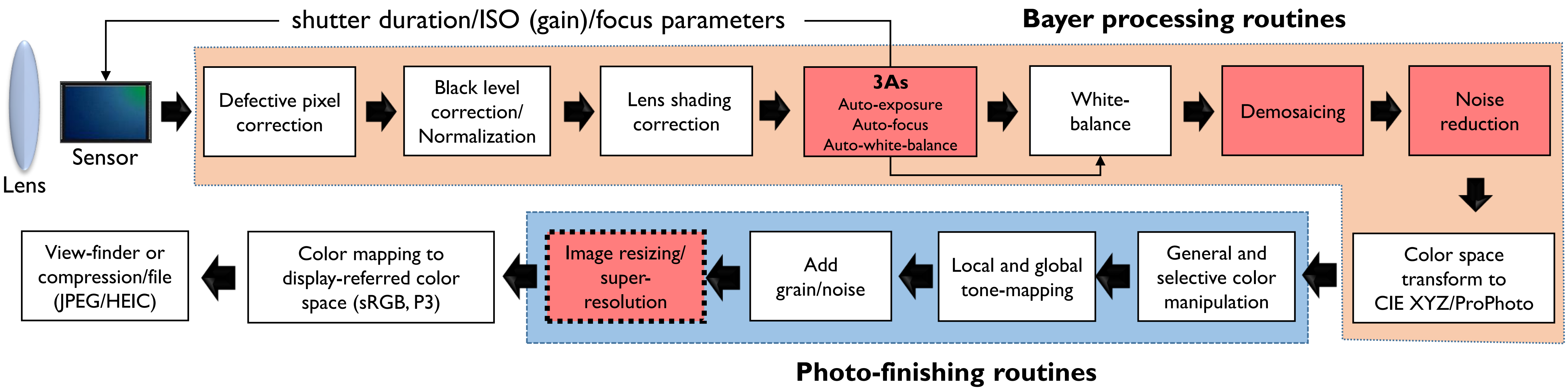
Congratulations!

Part 3:
AI targeting ISP components

A bit more complex ISP



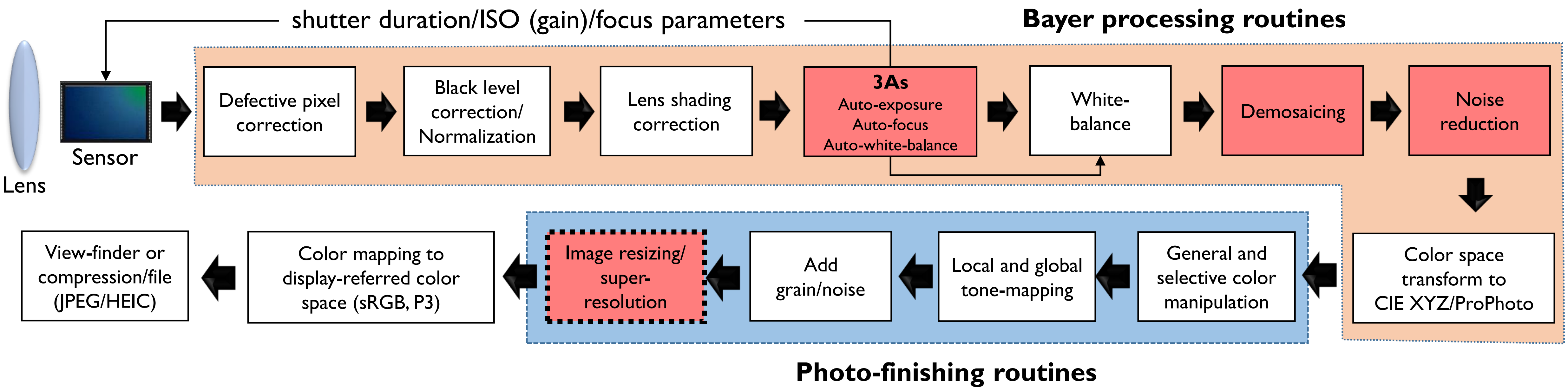
Use deep learning for hard problems



The highlighted components are camera pipeline steps that are challenging and areas AI can make notable gains:

- AWB (illumination estimation)
- Demosaicing
- Noise reduction
- Super-resolution

Use deep learning for hard problems



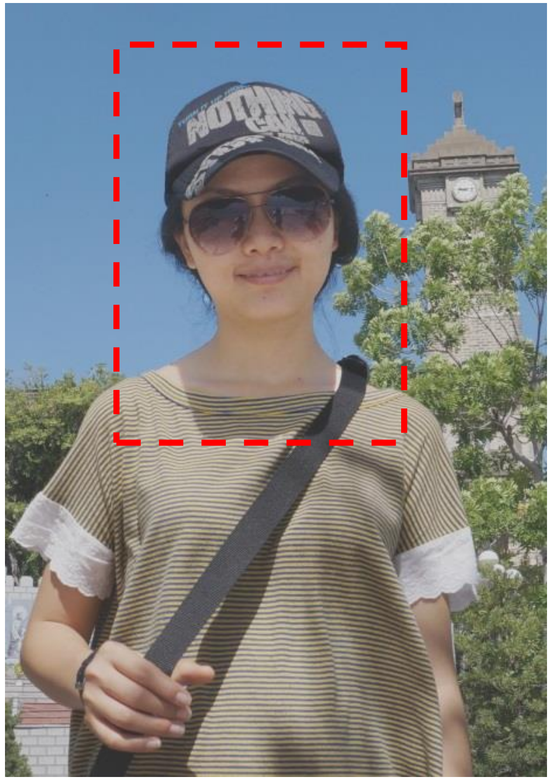
The highlighted components are camera pipeline steps that are challenging and areas AI can make notable gains:

AWB (illumination estimation)
Demosaicing
Noise reduction
Super-resolution

Digital zoom

A distinguishing feature in the smartphone camera market is zoom quality.

Full frame



Digital zoom
(super res)

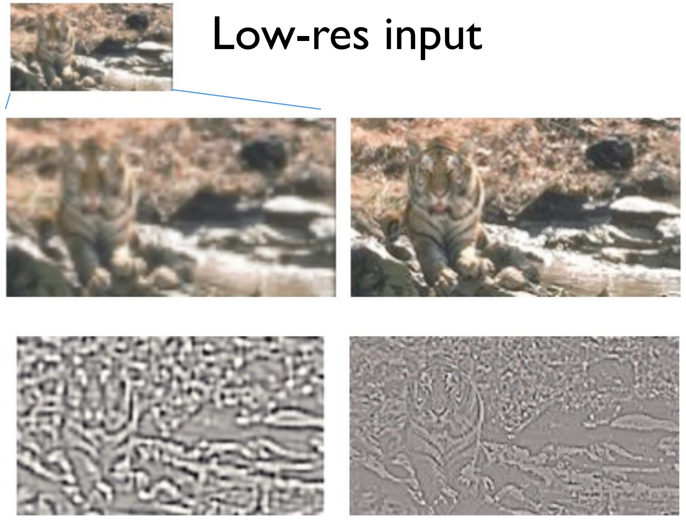
Machine learning (ML) for super-resolution

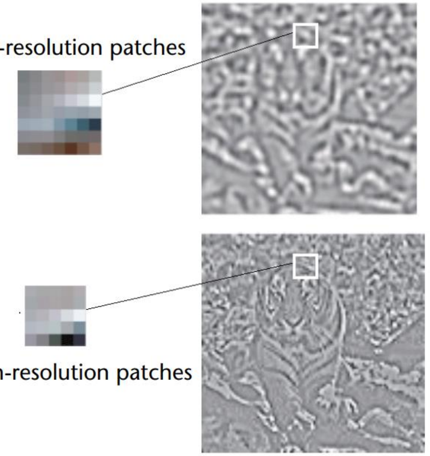
- SR has been addressed by machine learning methods for a long time.

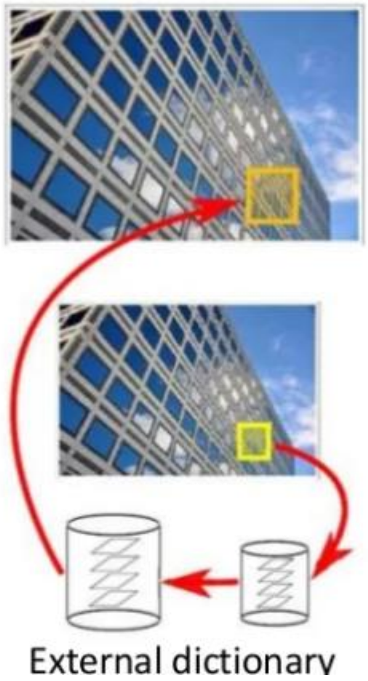
- Required "training data"
  - Quality of results are directly correlated to training data suitability.

- Before deep learning, used "non-learnable" machine learning.
  - Hand-crafted features
  - Conditional random fields
  - K-Nearest Neighbor
  - Support vector machines

Early example – Freeman 2002

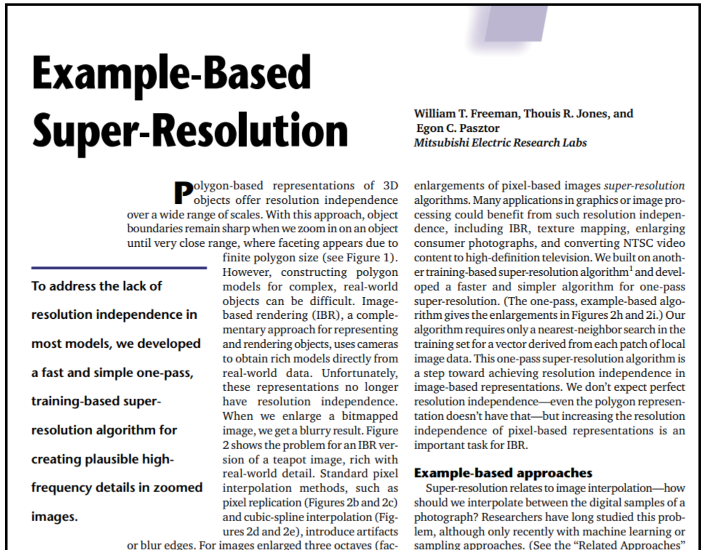






Training images were small photo collection

Search dictionary
for similar low-res patches,
replace with high-res patch.



Example-Based Super-Resolution

William T. Freeman, Thouis R. Jones, and Egon C. Pasztor
Mitsubishi Electric Research Labs

To address the lack of resolution independence in most models, we developed a fast and simple one-pass, training-based super-resolution algorithm for creating plausible high-frequency details in zoomed images.

Polygon-based representations of 3D objects offer resolution independence over a wide range of scales. With this approach, object boundaries remain sharp when we zoom in on an object until very close range, where faceting appears due to finite polygon size (see Figure 1). However, constructing polygon models for complex, real-world objects can be difficult. Image-based rendering (IBR), a complementary approach for representing and rendering objects, uses cameras to obtain rich models directly from real-world data. Unfortunately, these representations no longer have resolution independence. When we enlarge a bitmapped image, we get a blurry result. Figure 2 shows the problem for an IBR version of a teapot image, rich with real-world detail. Standard pixel interpolation methods, such as pixel replication (Figures 2b and 2c) and cubic-spline interpolation (Figures 2d and 2e), introduce artifacts or blur edges. For images enlarged three octaves (fac-

enlargements of pixel-based images *super-resolution* algorithms. Many applications in graphics or image processing could benefit from such resolution independence, including IBR, texture mapping, enlarging consumer photographs, and converting NTSC video content to high-definition television. We built on another training-based super-resolution algorithm[1] and developed a faster and simpler algorithm for one-pass super-resolution. (The one-pass, example-based algorithm gives the enlargements in Figures 2h and 2i.) Our algorithm requires only a nearest-neighbor search in the training set for a vector derived from each patch of local image data. This one-pass super-resolution algorithm is a step toward achieving resolution independence in image-based representations. We don't expect perfect resolution independence—even the polygon representation doesn't have that—but increasing the resolution independence of pixel-based representations is an important task for IBR.

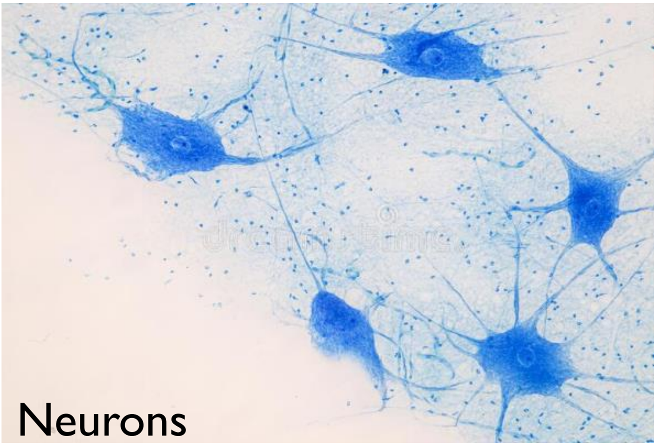
Example-based approaches

Super-resolution relates to image interpolation—how should we interpolate between the digital samples of a photograph? Researchers have long studied this problem, although only recently with machine learning or sampling approaches. (See the "Related Approaches"

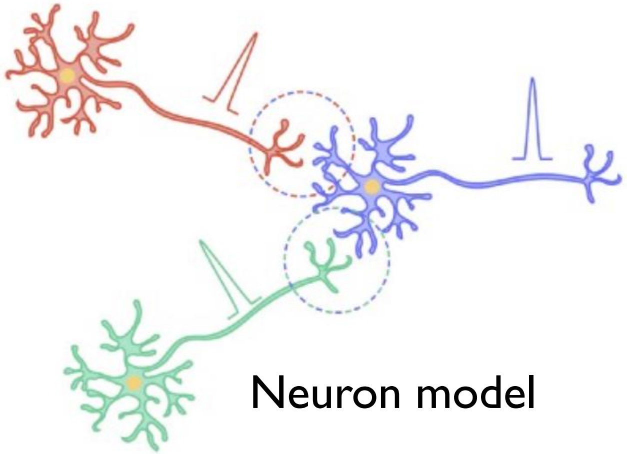
Freeman et al – IEEE Computer Graphics and Applications [2002]

Enter deep learning

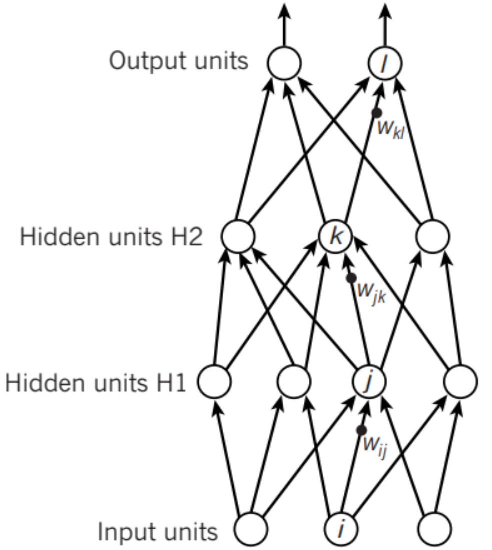
Approach ML problems with learnable processing graphs inspired by biological neurons.



Neurons



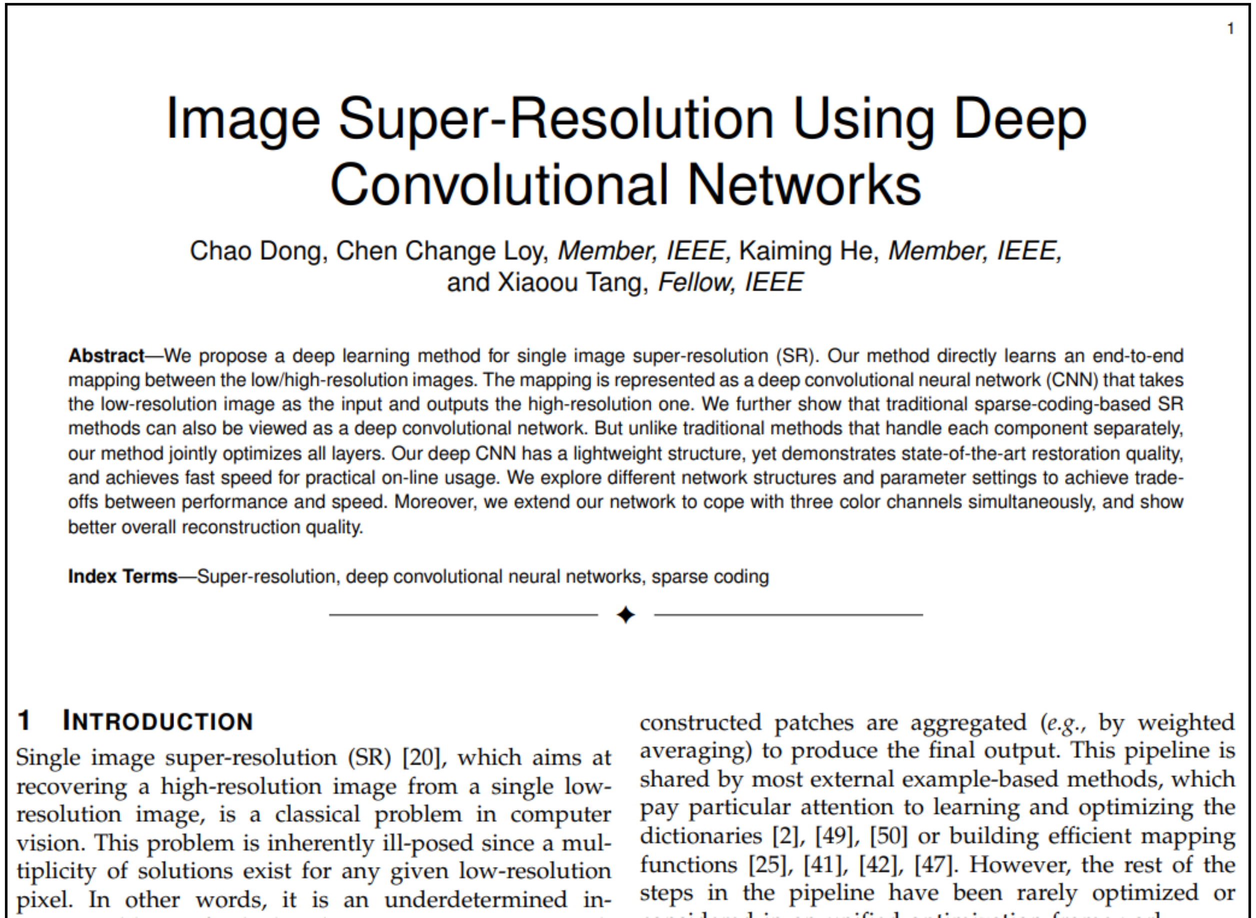
Neuron model



Artificial "neural" network graph

Super-resolution was target of early CNNs

Dong, Loy, He, Tang [ECCV'14]



1

Image Super-Resolution Using Deep Convolutional Networks

Chao Dong, Chen Change Loy, *Member, IEEE*, Kaiming He, *Member, IEEE*, and Xiaoou Tang, *Fellow, IEEE*

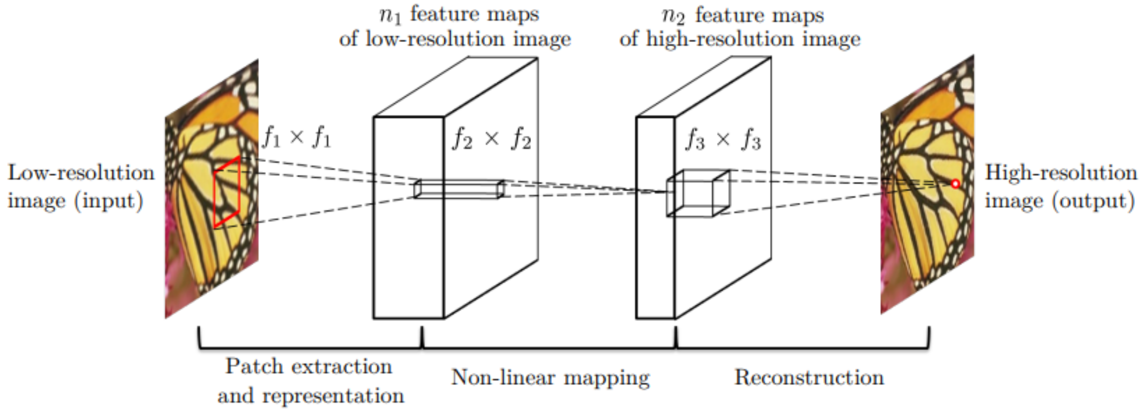
Abstract—We propose a deep learning method for single image super-resolution (SR). Our method directly learns an end-to-end mapping between the low/high-resolution images. The mapping is represented as a deep convolutional neural network (CNN) that takes the low-resolution image as the input and outputs the high-resolution one. We further show that traditional sparse-coding-based SR methods can also be viewed as a deep convolutional network. But unlike traditional methods that handle each component separately, our method jointly optimizes all layers. Our deep CNN has a lightweight structure, yet demonstrates state-of-the-art restoration quality, and achieves fast speed for practical on-line usage. We explore different network structures and parameter settings to achieve trade-offs between performance and speed. Moreover, we extend our network to cope with three color channels simultaneously, and show better overall reconstruction quality.

Index Terms—Super-resolution, deep convolutional neural networks, sparse coding

1 Introduction

Single image super-resolution (SR) [20], which aims at recovering a high-resolution image from a single low-resolution image, is a classical problem in computer vision. This problem is inherently ill-posed since a multiplicity of solutions exist for any given low-resolution pixel. In other words, it is an underdetermined in-

constructed patches are aggregated (*e.g.*, by weighted averaging) to produce the final output. This pipeline is shared by most external example-based methods, which pay particular attention to learning and optimizing the dictionaries [2], [49], [50] or building efficient mapping functions [25], [41], [42], [47]. However, the rest of the steps in the pipeline have been rarely optimized or



Let network learn "feature."
Let network learn how to reconstruct.

CNN performance

Eval. Mat	Scale	Bicubic	SC [50]	NE+LLE [4]	KK [25]	ANR [41]	A+ [41]	SRCNN
PSNR	2	30.23	-	31.76	32.11	31.80	32.28	**32.45**
	3	27.54	28.31	28.60	28.94	28.65	29.13	**29.30**
	4	26.00	-	26.81	27.14	26.85	27.32	**27.50**
SSIM	2	0.8687	-	0.8993	0.9026	0.9004	0.9056	**0.9067**
	3	0.7736	0.7954	0.8076	0.8132	0.8093	0.8188	**0.8215**
	4	0.7019	-	0.7331	0.7419	0.7352	0.7491	**0.7513**
IFC	2	6.09	-	7.59	6.83	7.81	**8.11**	7.76
	3	3.41	2.98	4.14	3.83	4.23	**4.45**	4.26
	4	2.23	-	2.71	2.57	2.78	**2.94**	2.74
NQM	2	40.98	-	41.34	38.86	41.79	**42.61**	38.95
	3	33.15	29.06	37.12	35.23	37.22	**38.24**	35.25
	4	26.15	-	31.17	29.18	31.27	**32.31**	30.46
WPSNR	2	47.64	-	54.47	53.85	54.57	**55.62**	55.39
	3	39.72	41.66	43.22	43.56	43.36	44.25	**44.32**
	4	35.71	-	37.75	38.26	37.85	38.72	**38.87**
MSSSIM	2	0.9813	-	0.9886	0.9890	0.9888	0.9896	**0.9897**
	3	0.9512	0.9595	0.9643	0.9653	0.9647	0.9669	**0.9675**
	4	0.9134	-	0.9317	0.9338	0.9326	0.9371	**0.9376**

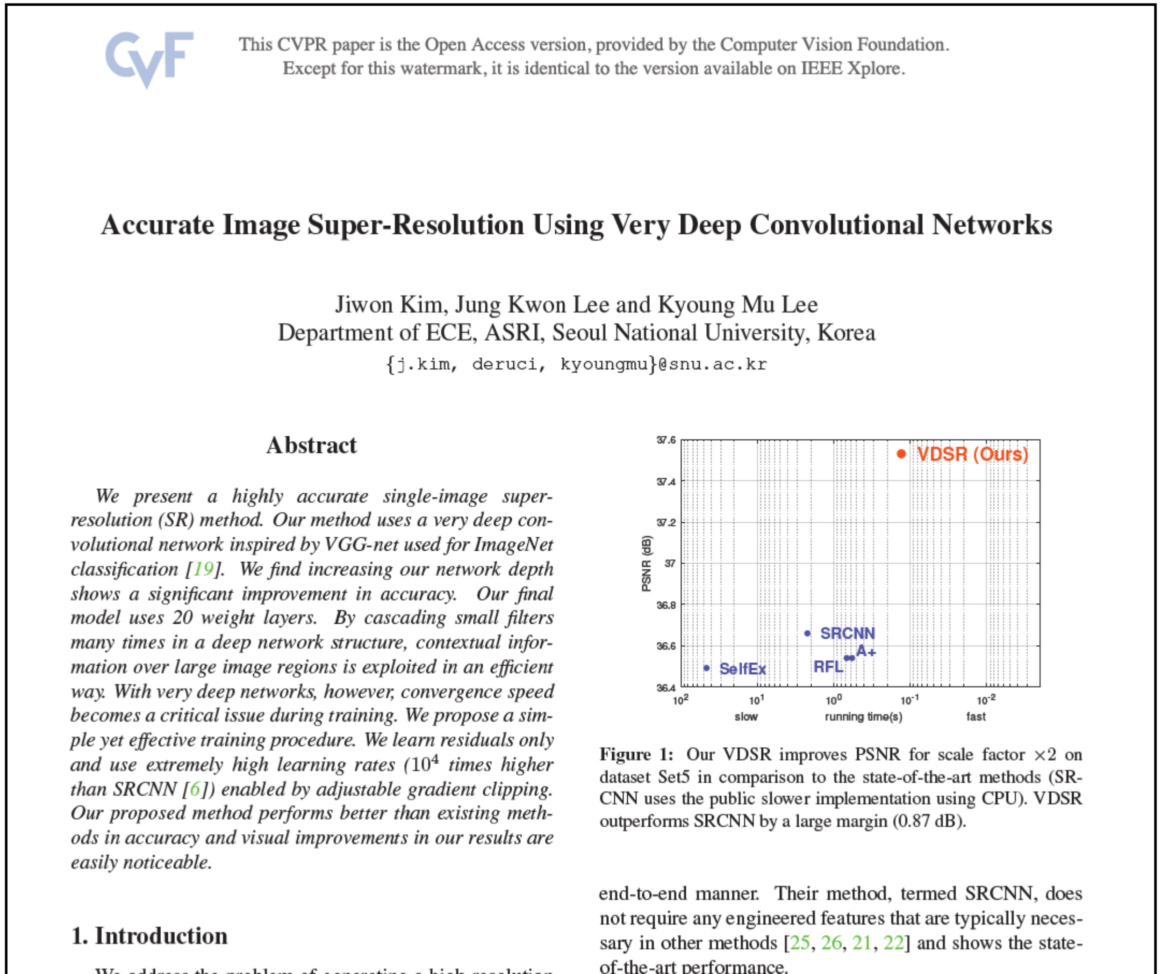
Early CNN approaches were not always the "best."

Highly hand-crafted methods still worked well.

But, CNN learned everything.
(The care now was in optimizing the CNN).

Super-resolution with very deep networks

Kim, Lee, Lee CVPR'16



Accurate Image Super-Resolution Using Very Deep Convolutional Networks

Jiwon Kim, Jung Kwon Lee and Kyoung Mu Lee
Department of ECE, ASRI, Seoul National University, Korea
{j.kim, deruci, kyoungmu}@snu.ac.kr

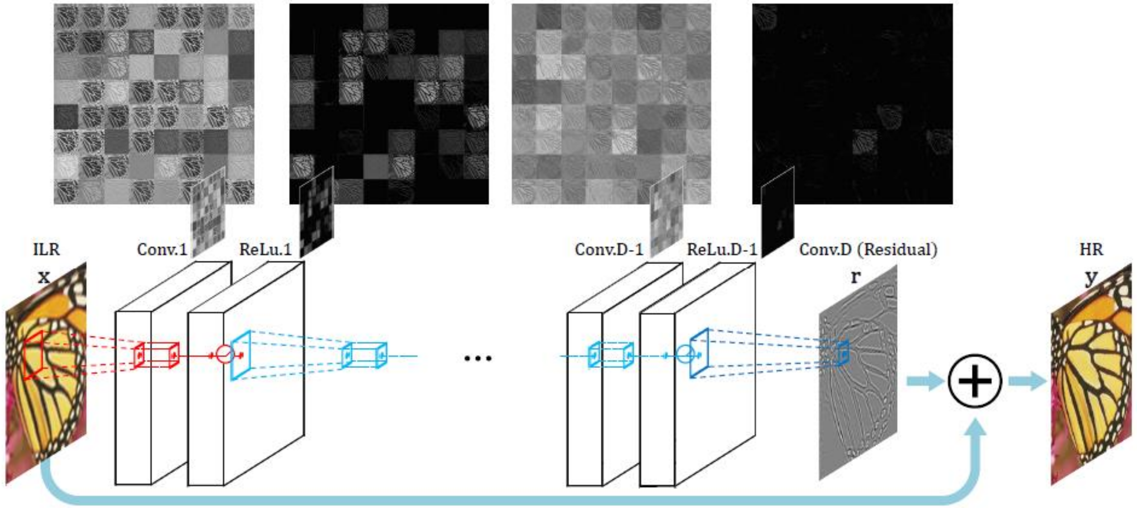
Abstract

We present a highly accurate single-image super-resolution (SR) method. Our method uses a very deep convolutional network inspired by VGG-net used for ImageNet classification [19]. We find increasing our network depth shows a significant improvement in accuracy. Our final model uses 20 weight layers. By cascading small filters many times in a deep network structure, contextual information over large image regions is exploited in an efficient way. With very deep networks, however, convergence speed becomes a critical issue during training. We propose a simple yet effective training procedure. We learn residuals only and use extremely high learning rates (10^4 times higher than SRCNN [6]) enabled by adjustable gradient clipping. Our proposed method performs better than existing methods in accuracy and visual improvements in our results are easily noticeable.

Figure 1: Our VDSR improves PSNR for scale factor ×2 on dataset Set5 in comparison to the state-of-the-art methods (SRCNN uses the public slower implementation using CPU). VDSR outperforms SRCNN by a large margin (0.87 dB).

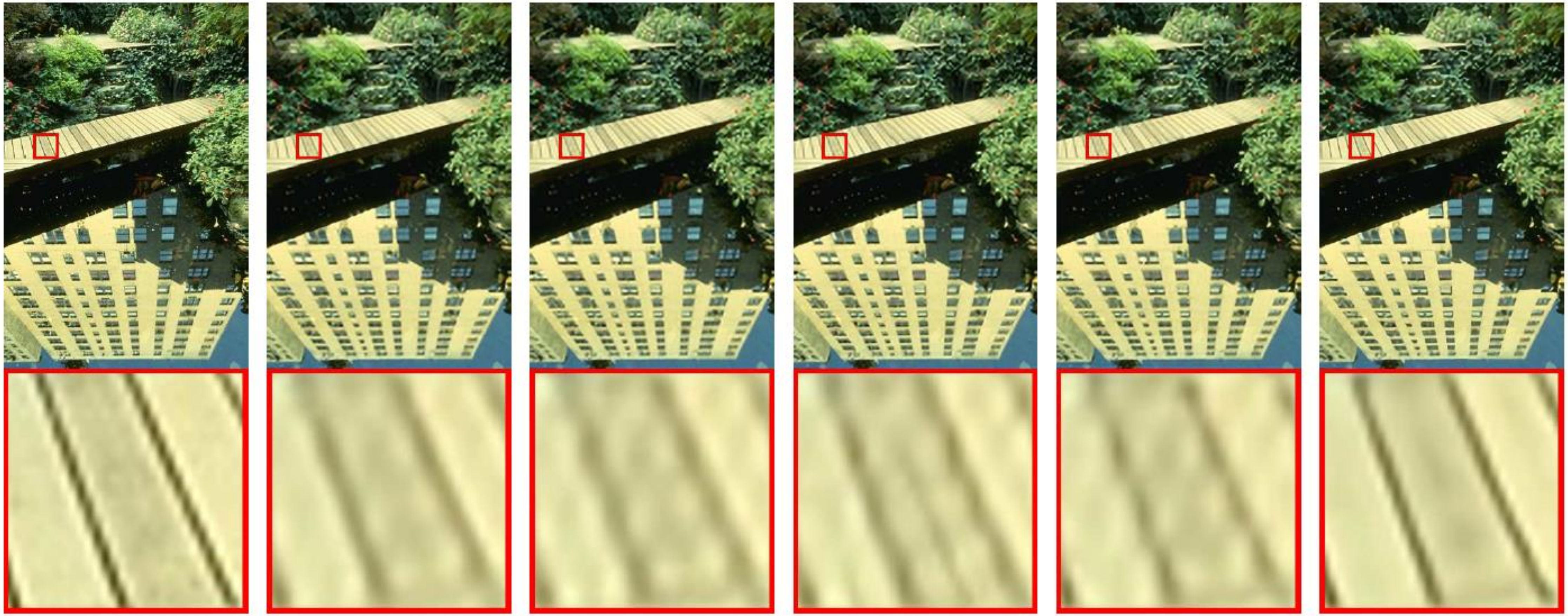
1. Introduction

end-to-end manner. Their method, termed SRCNN, does not require any engineered features that are typically necessary in other methods [25, 26, 21, 22] and shows the state-of-the-art performance.



- Pairs of convolution layers + nonlinear activations
- Prediction is added to upsampled low-res input
- Special care for gradient clipping

Took "SR" to the next level visually



Ground Truth	A+ [22]	RFL [18]	SelfEx [11]	SRCNN [5]	VDSR (Ours)
(PSNR, SSIM)	(22.92, 0.7379)	(22.90, 0.7332)	(23.00, 0.7439)	(23.15, 0.7487)	(23.50, 0.7777)

Adding adversial loss to SR

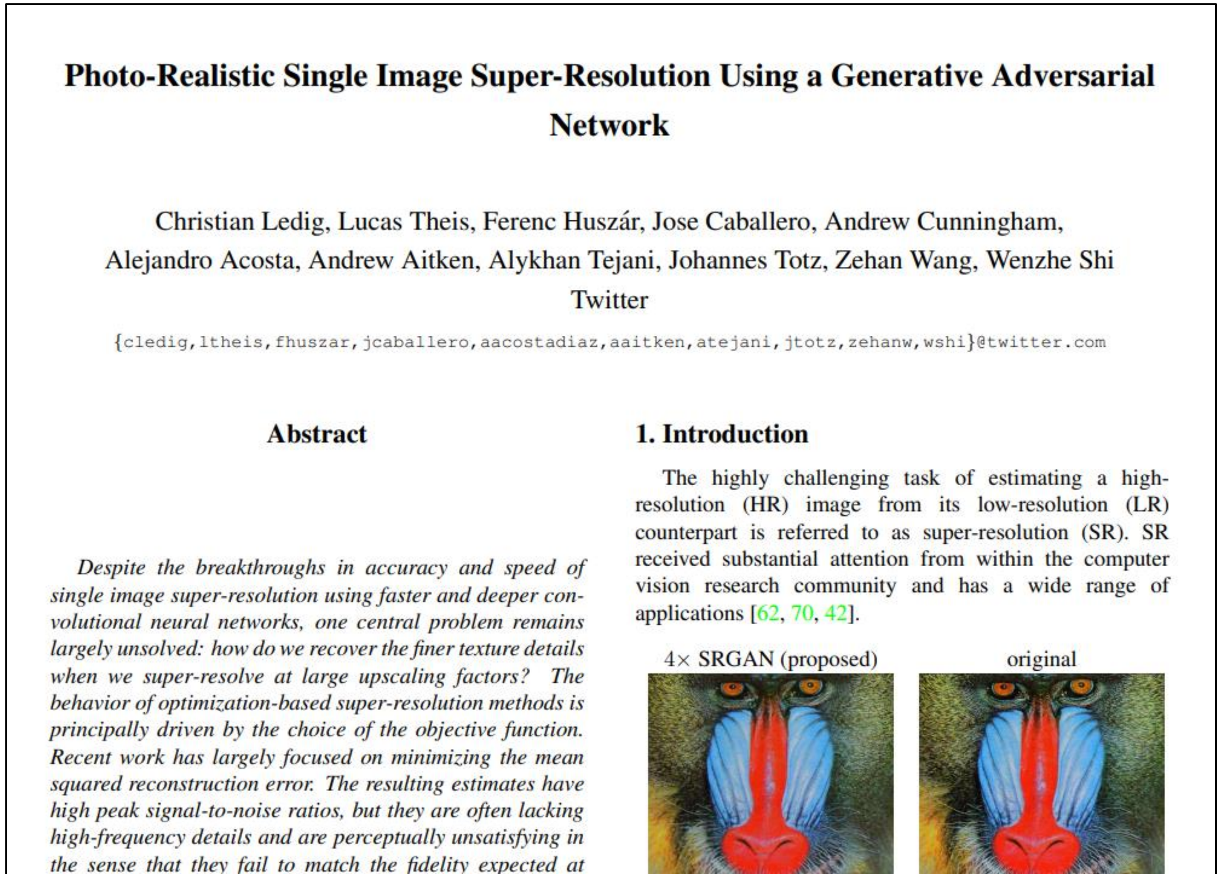


Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham,
Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi
Twitter
{cledig,ltheis,fhuszar,jcaballero,aacostadiaz,aaitken,atejani,jtotz,zehanw,wshi}@twitter.com

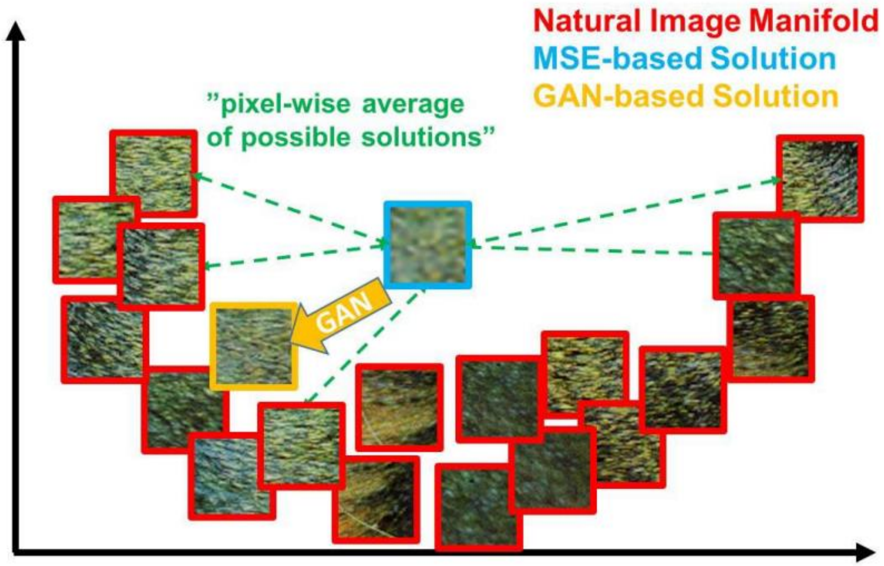
Abstract

Despite the breakthroughs in accuracy and speed of single image super-resolution using faster and deeper convolutional neural networks, one central problem remains largely unsolved: how do we recover the finer texture details when we super-resolve at large upscaling factors? The behavior of optimization-based super-resolution methods is principally driven by the choice of the objective function. Recent work has largely focused on minimizing the mean squared reconstruction error. The resulting estimates have high peak signal-to-noise ratios, but they are often lacking high-frequency details and are perceptually unsatisfying in the sense that they fail to match the fidelity expected at

1. Introduction

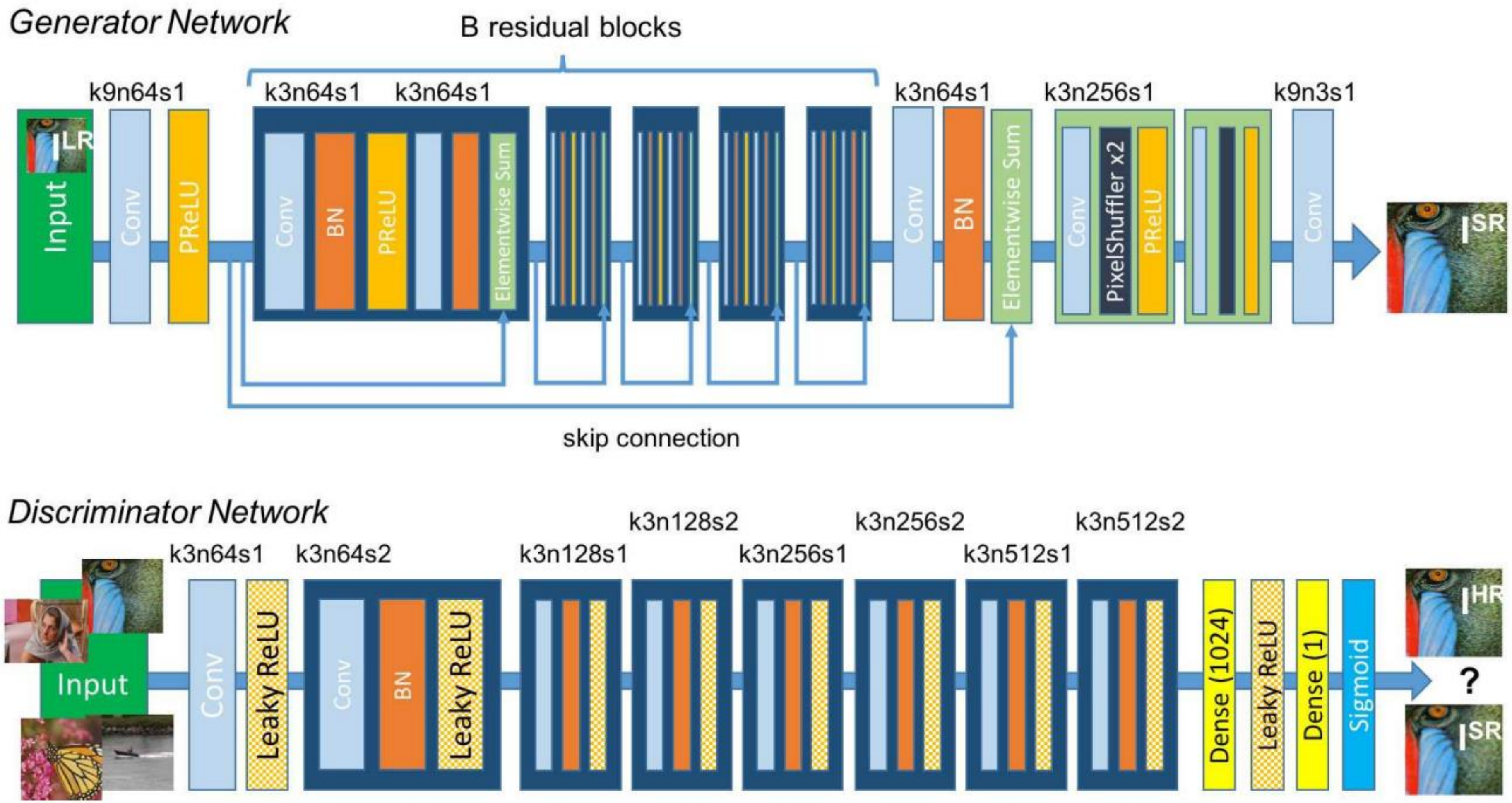
The highly challenging task of estimating a high-resolution (HR) image from its low-resolution (LR) counterpart is referred to as super-resolution (SR). SR received substantial attention from within the computer vision research community and has a wide range of applications [62, 70, 42].

4× SRGAN (proposed) original

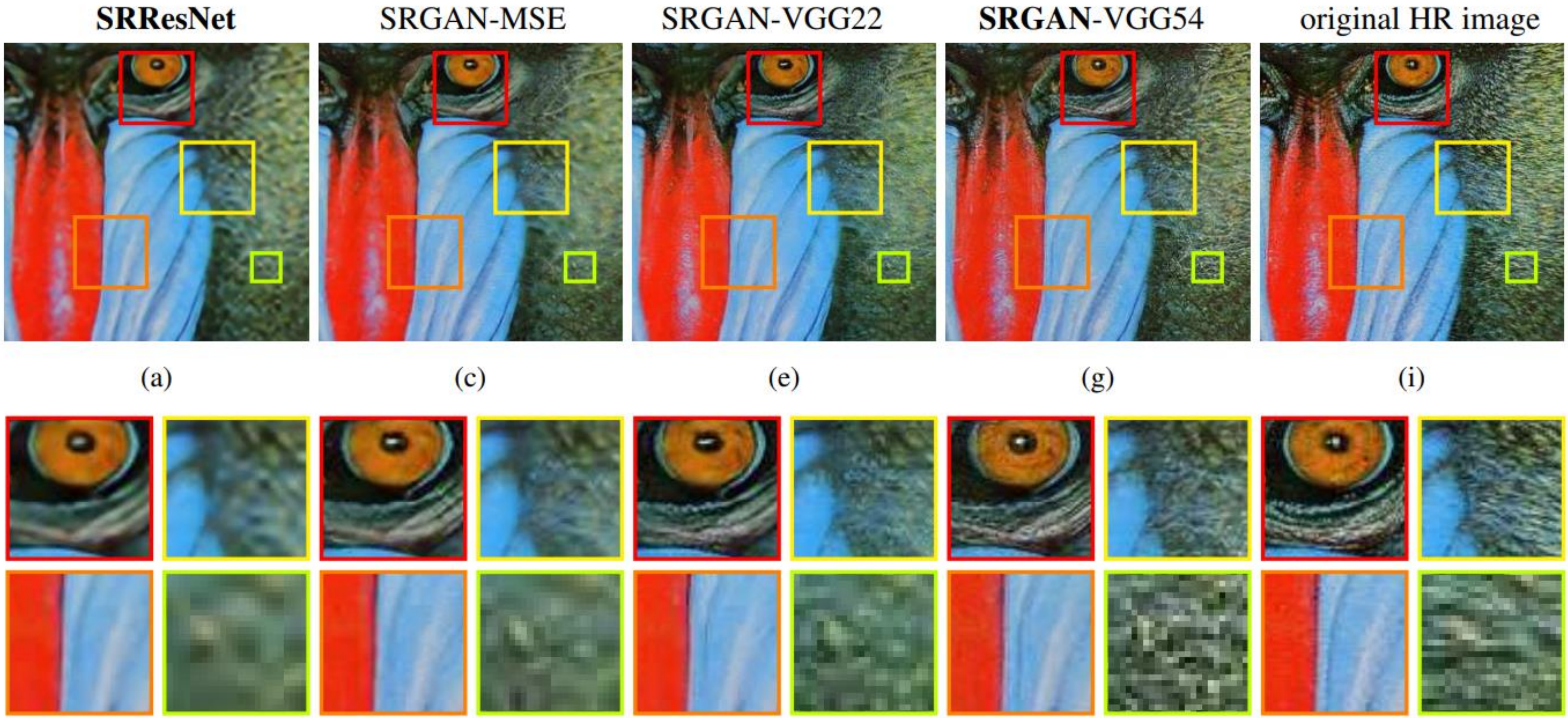


In corporate additional loss that considers how realistic the solution is with respect to image distribution.

SRGAN structure



GAN loss adds another visual level!



AIM/NTIRE SR challenges

New **T**rends in **I**mage **R**estoration and **E**nhancement (NTIRE)

Advances in **I**mage **M**anipulation (AIM)

- These workshops run regular challenges on super resolution

NTIRE 2022 Challenge on Efficient Super-Resolution: Methods and Results

Yawei Li* Kai Zhang* Radu Timofte* Luc Van Gool* Fangyuan Kong
Mingxi Li Songwei Liu Zongcai Du Ding Liu Chenhui Zhou Jingyi Chen
Qingrui Han Zheyuan Li Yingqi Liu Xiangyu Chen Haoming Cai Yu Qiao
Chao Dong Long Sun Jinshan Pan Yi Zhu Zhikai Zong Xiaoxiao Liu
Zheng Hui Tao Yang Peiran Ren Xuansong Xie Xian-Sheng Hua Yanbo Wang
Xiaozhong Ji Chuming Lin Donghao Luo Ying Tai Chengjie Wang
Zhizhong Zhang Yuan Xie Shen Cheng Ziwei Luo Lei Yu Zhihong Wen
Qi Wu1 Youwei Li Haoqiang Fan Jian Sun Shuaicheng Liu Yuanfei Huang
Meiguang Jin Hua Huang Jing Liu Xinjian Zhang Yan Wang Lingshun Long
Gen Li Yuanfan Zhang Zuowei Cao Lei Sun Panaetov Alexander
Yucong Wang Minjie Cai Li Wang Lu Tian Zheyuan Wang Hongbing Ma
Jie Liu Chao Chen Yidong Cai Jie Tang Gangshan Wu Weiran Wang
Shirui Huang Honglei Lu Huan Liu Keyan Wang Jun Chen Shi Chen
Yuchun Miao Zimo Huang Lefei Zhang Mustafa Ayazoğlu Wei Xiong
Chengyi Xiong Fei Wang Hao Li Ruimian Wen Zhijing Yang Wenbin Zou
Weixin Zheng Tian Ye Yuncheng Zhang Xiangzhen Kong Aditya Arora
Syed Waqas Zamir Salman Khan Munawar Hayat Fahad Shahbaz Khan
Dandan Gao Dengwen Zhou Qian Ning Jingzhu Tang Han Huang Yufei Wang
Zhangheng Peng Haobo Li Wenxue Guan Shenghua Gong Xin Li Jun Liu
Wanjun Wang Dengwen Zhou Kun Zeng Hanjiang Lin Xinyu Chen
Jinsheng Fang

- Current winning solutions are transformer-based.

- Interestingly, solutions do not use GANs!
  GANs often help with visual appearance.
  GANs do not necessarily beat benchmarks
  Benchmarks are based on RMSE/SIMM losses.

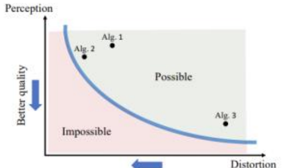
See CVPR'18 paper – Perception-Distortion Tradeoff

The Perception-Distortion Tradeoff

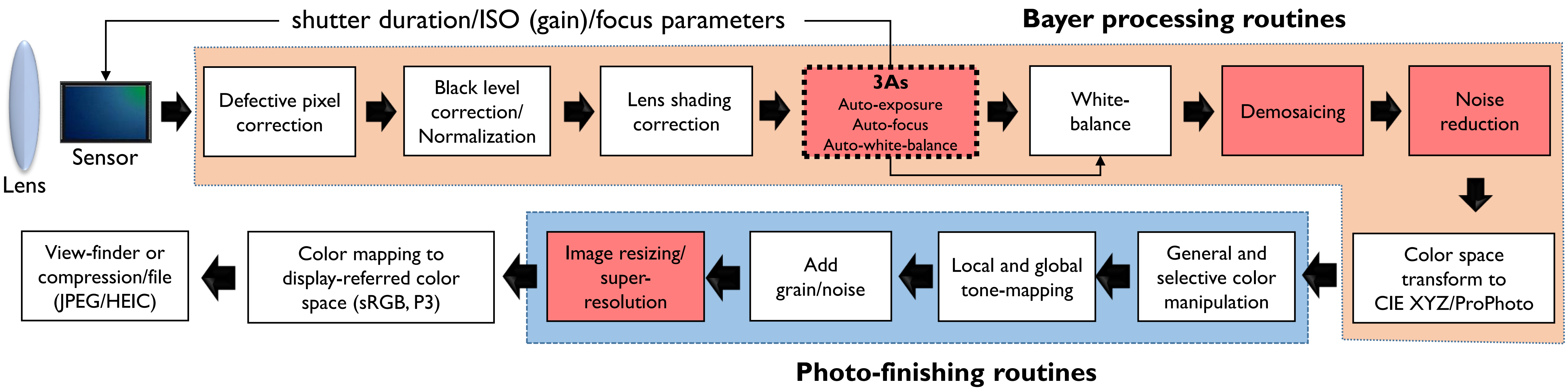
Yochai Blau and Tomer Michaeli
Technion–Israel Institute of Technology, Haifa, Israel
{yochai@campus,tomer.m@ee}.technion.ac.il

Abstract

Image restoration algorithms are typically evaluated by some distortion measure (e.g. PSNR, SSIM, IFC, VIF) or by human opinion scores that quantify perceived perceptual quality. In this paper, we prove mathematically that distortion and perceptual quality are at odds with each other. Specifically, we study the optimal probability for correctly discriminating the outputs of an image restoration algorithm from real images. We show that as the mean dis-



Use deep learning for hard problems



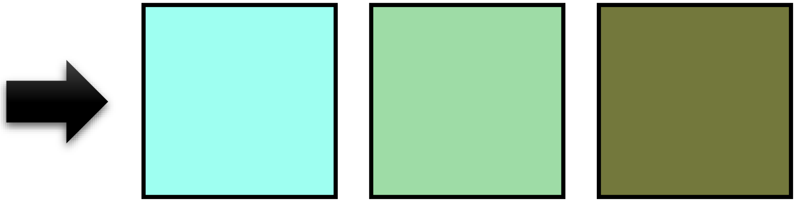
The highlighted components are camera pipeline steps that are challenging and areas AI can make notable gains:

AWB (illumination estimation)
Demosaicing
Noise reduction
Super-resolution

Recall why illumination estimation is hard

RAW sensor image

What is the sensor's response to illumination?

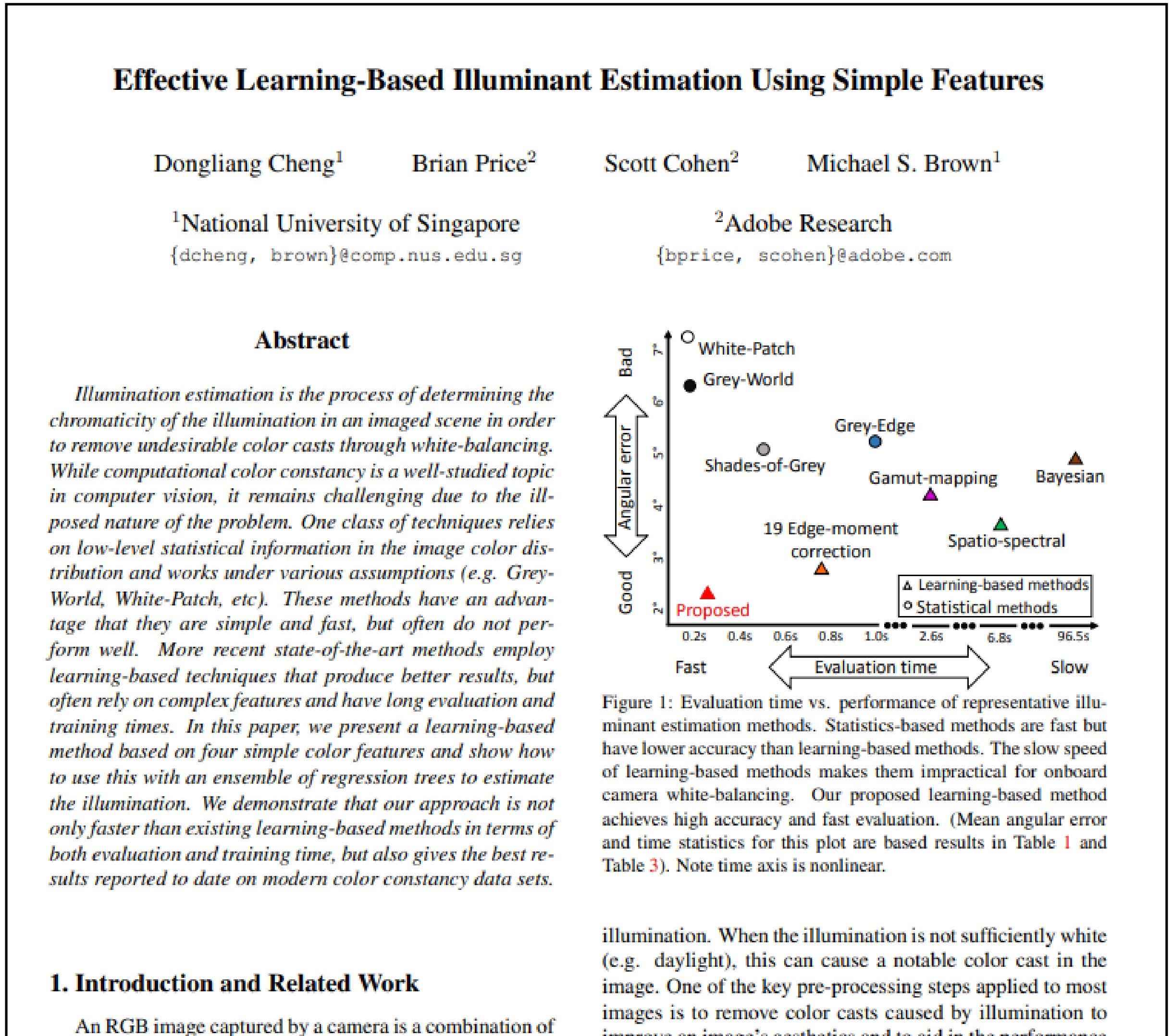


Given an arbitrary input image, predict the scene illumination.

Getting this *incorrect* has significant impact on image quality/color reproduction.

Many ML approaches before deep learning

Cheng et al CVPR'15



Effective Learning-Based Illuminant Estimation Using Simple Features

Dongliang Cheng[1] Brian Price[2] Scott Cohen[2] Michael S. Brown[1]

[1]National University of Singapore
{dcheng, brown}@comp.nus.edu.sg

[2]Adobe Research
{bprice, scohen}@adobe.com

Abstract

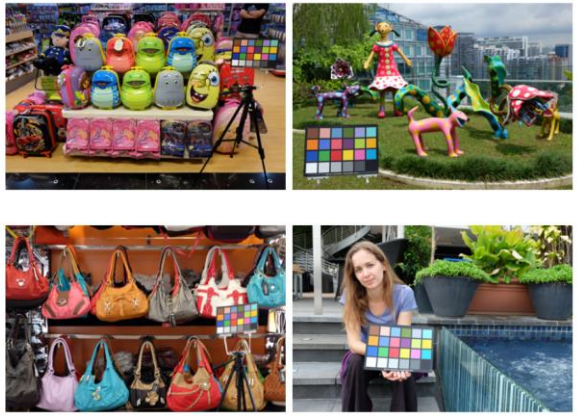
Illumination estimation is the process of determining the chromaticity of the illumination in an imaged scene in order to remove undesirable color casts through white-balancing. While computational color constancy is a well-studied topic in computer vision, it remains challenging due to the ill-posed nature of the problem. One class of techniques relies on low-level statistical information in the image color distribution and works under various assumptions (e.g. Grey-World, White-Patch, etc). These methods have an advantage that they are simple and fast, but often do not perform well. More recent state-of-the-art methods employ learning-based techniques that produce better results, but often rely on complex features and have long evaluation and training times. In this paper, we present a learning-based method based on four simple color features and show how to use this with an ensemble of regression trees to estimate the illumination. We demonstrate that our approach is not only faster than existing learning-based methods in terms of both evaluation and training time, but also gives the best results reported to date on modern color constancy data sets.

1. Introduction and Related Work

An RGB image captured by a camera is a combination of

Figure 1: Evaluation time vs. performance of representative illuminant estimation methods. Statistics-based methods are fast but have lower accuracy than learning-based methods. The slow speed of learning-based methods makes them impractical for onboard camera white-balancing. Our proposed learning-based method achieves high accuracy and fast evaluation. (Mean angular error and time statistics for this plot are based results in Table 1 and Table 3). Note time axis is nonlinear.

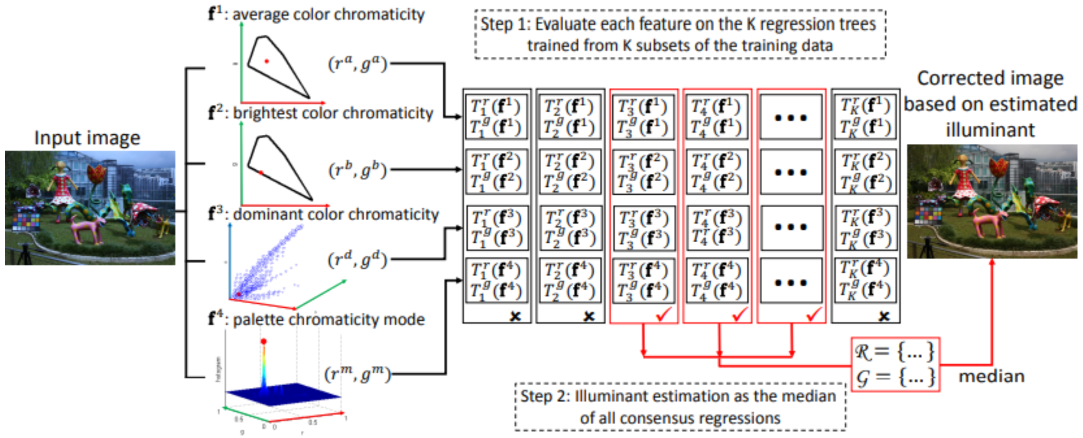
illumination. When the illumination is not sufficiently white (e.g. daylight), this can cause a notable color cast in the image. One of the key pre-processing steps applied to most images is to remove color casts caused by illumination to



Training images (sensor specific)

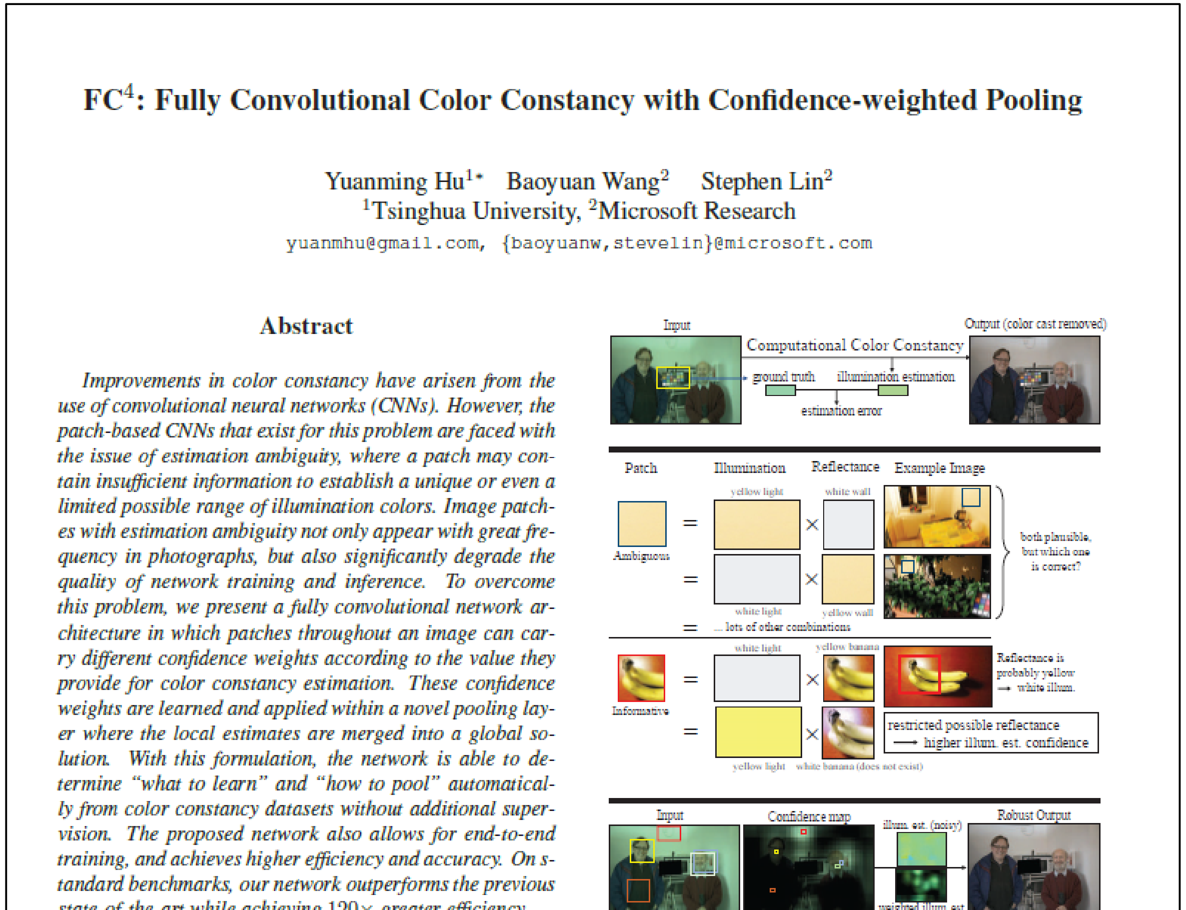
Derive some features (usually histogram statistics)

Apply ML method to predict illumination of scene.



Improving AWB with CNN

Hu et al. CVPR'17

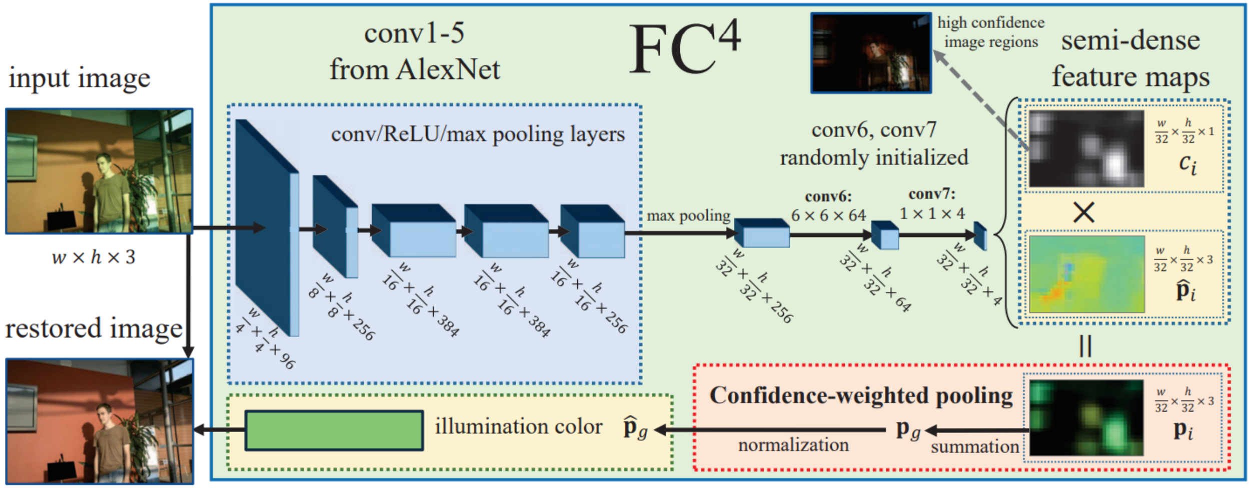


FC[4]: Fully Convolutional Color Constancy with Confidence-weighted Pooling

Yuanming Hu[1*] Baoyuan Wang[2] Stephen Lin[2]
[1]Tsinghua University, [2]Microsoft Research
yuanmhu@gmail.com, {baoyuanw,stevelin}@microsoft.com

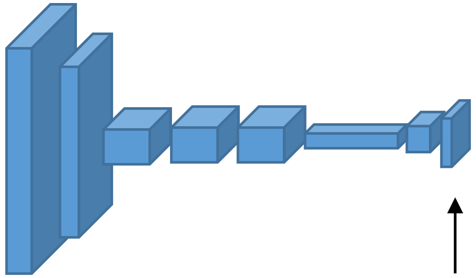
Abstract

Improvements in color constancy have arisen from the use of convolutional neural networks (CNNs). However, the patch-based CNNs that exist for this problem are faced with the issue of estimation ambiguity, where a patch may contain insufficient information to establish a unique or even a limited possible range of illumination colors. Image patches with estimation ambiguity not only appear with great frequency in photographs, but also significantly degrade the quality of network training and inference. To overcome this problem, we present a fully convolutional network architecture in which patches throughout an image can carry different confidence weights according to the value they provide for color constancy estimation. These confidence weights are learned and applied within a novel pooling layer where the local estimates are merged into a global solution. With this formulation, the network is able to determine "what to learn" and "how to pool" automatically from color constancy datasets without additional supervision. The proposed network also allows for end-to-end training, and achieves higher efficiency and accuracy. On standard benchmarks, our network outperforms the previous state of the art while achieving 120× greater efficiency

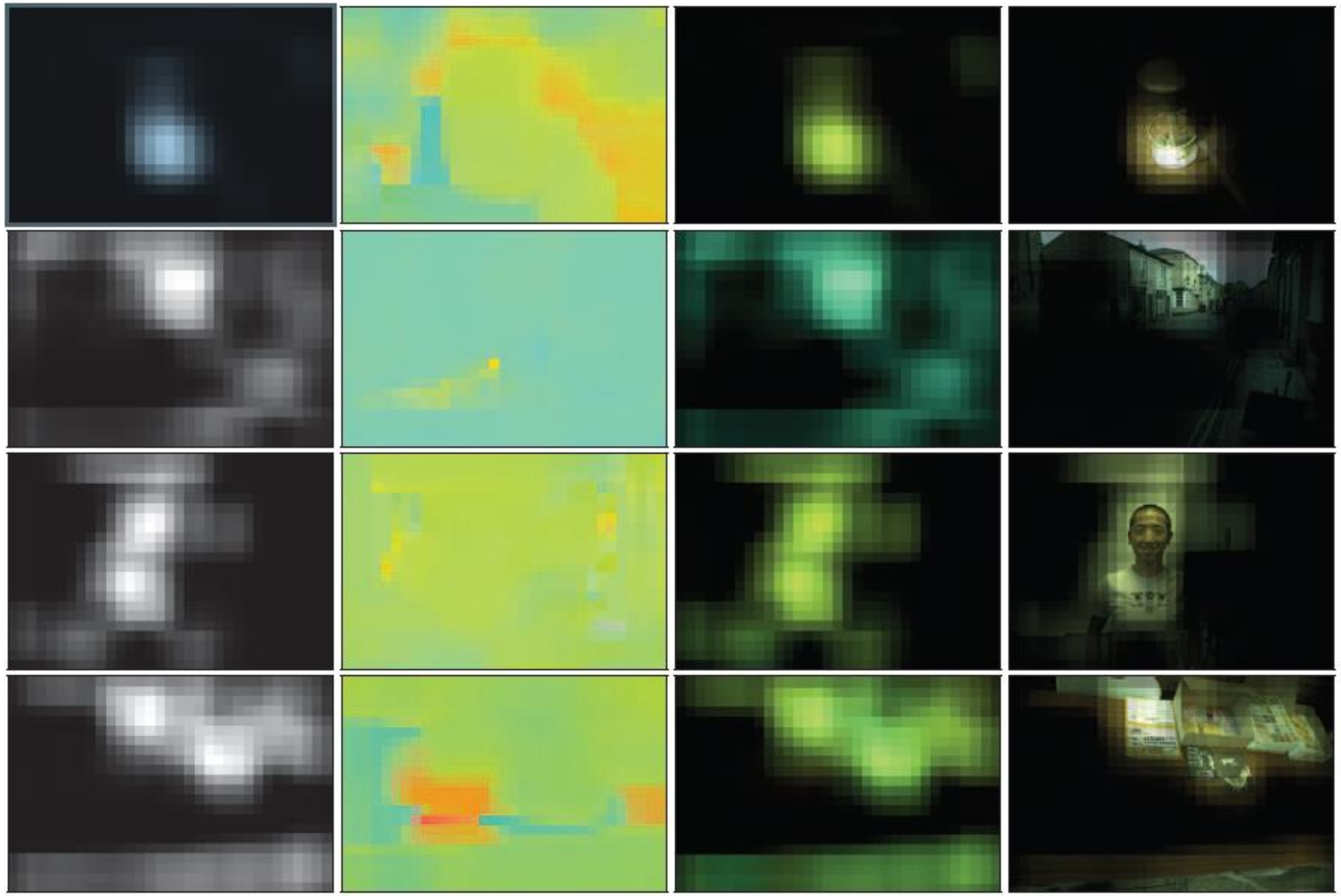


Predicts local estimates over the image and their confidence.
Pools confident weighted estimates for final result.

What did it learn?



Predicts low-res
4 channel ouput
(estimation r,g,b
+ confidence map)

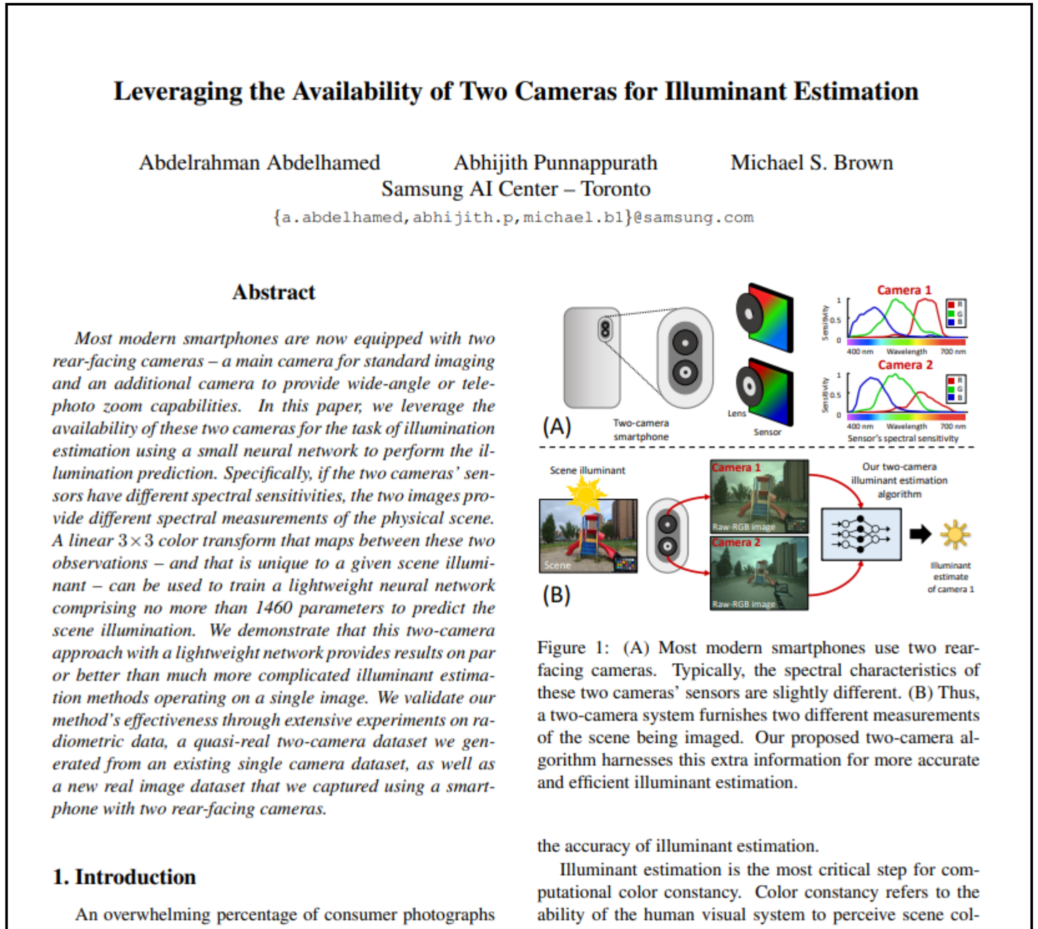


confidence map | estimation map | weighted estimation | image × confidence

The method appears to learns to identify pixels that are most likely "neutral/achromatic" scene patches.

Exploiting two cameras for AWB

Abdelhamed et al CVPR'21



Leveraging the Availability of Two Cameras for Illuminant Estimation

Abdelrahman Abdelhamed Abhijith Punnappurath Michael S. Brown
Samsung AI Center – Toronto
{a.abdelhamed,abhijith.p,michael.b1}@samsung.com

Abstract

Most modern smartphones are now equipped with two rear-facing cameras – a main camera for standard imaging and an additional camera to provide wide-angle or telephoto zoom capabilities. In this paper, we leverage the availability of these two cameras for the task of illumination estimation using a small neural network to perform the illumination prediction. Specifically, if the two cameras' sensors have different spectral sensitivities, the two images provide different spectral measurements of the physical scene. A linear 3×3 color transform that maps between these two observations – and that is unique to a given scene illuminant – can be used to train a lightweight neural network comprising no more than 1460 parameters to predict the scene illumination. We demonstrate that this two-camera approach with a lightweight network provides results on par or better than much more complicated illuminant estimation methods operating on a single image. We validate our method's effectiveness through extensive experiments on radiometric data, a quasi-real two-camera dataset we generated from an existing single camera dataset, as well as a new real image dataset that we captured using a smartphone with two rear-facing cameras.

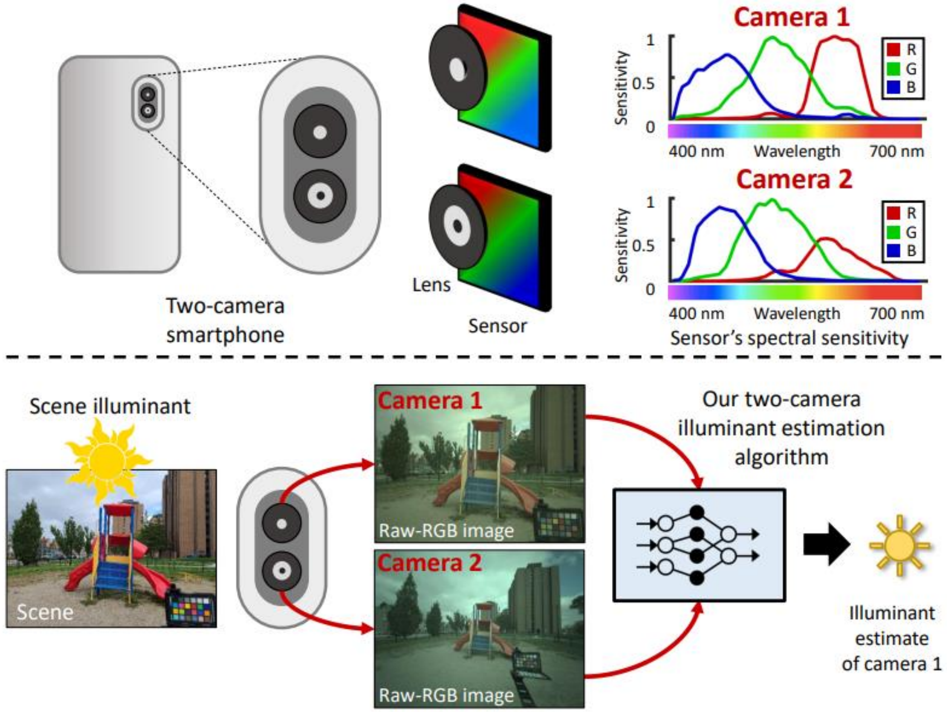
1. Introduction

An overwhelming percentage of consumer photographs

Figure 1: (A) Most modern smartphones use two rear-facing cameras. Typically, the spectral characteristics of these two cameras' sensors are slightly different. (B) Thus, a two-camera system furnishes two different measurements of the scene being imaged. Our proposed two-camera algorithm harnesses this extra information for more accurate and efficient illuminant estimation.

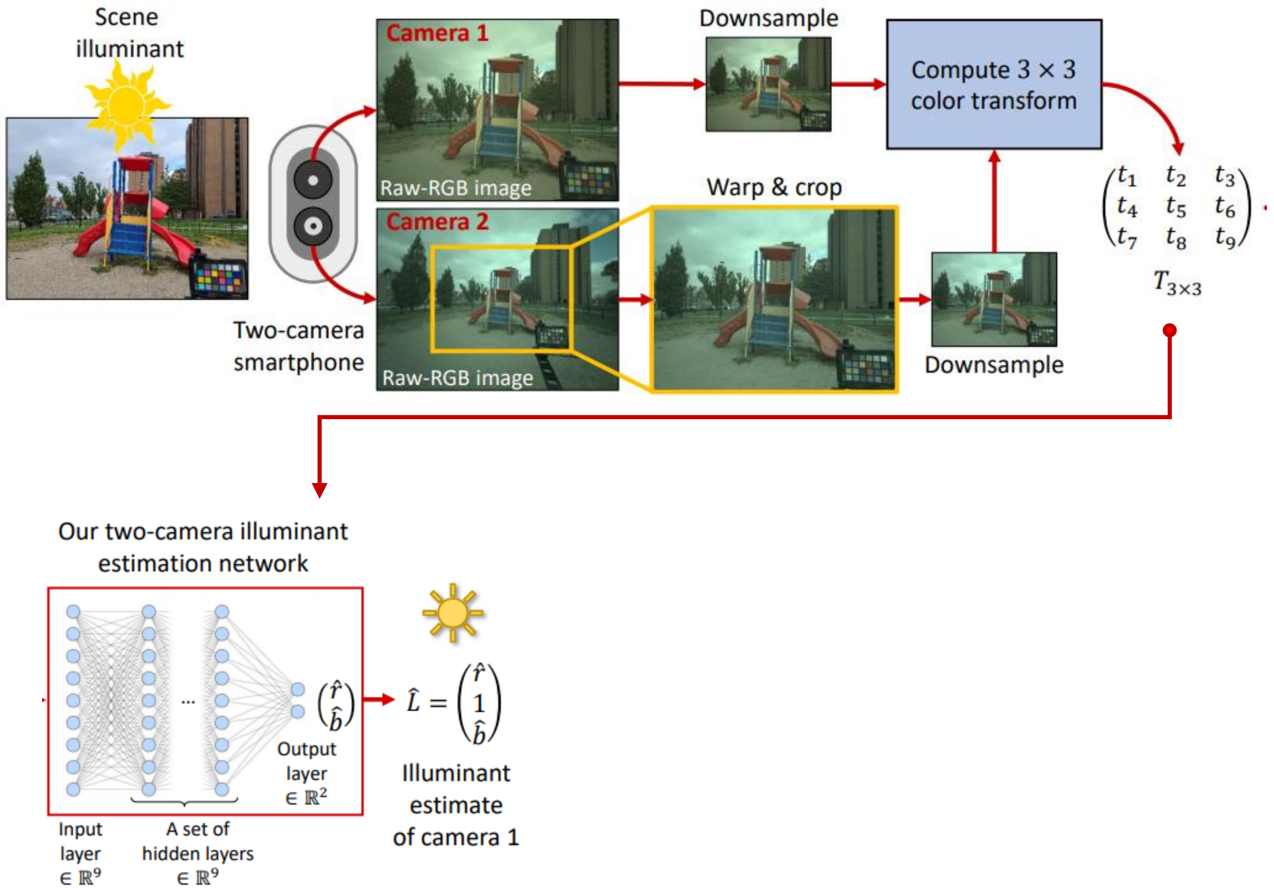
the accuracy of illuminant estimation.

Illuminant estimation is the most critical step for computational color constancy. Color constancy refers to the ability of the human visual system to perceive scene col-



Two views of the same scene with different sensors – is essentially a 6-channel camera.

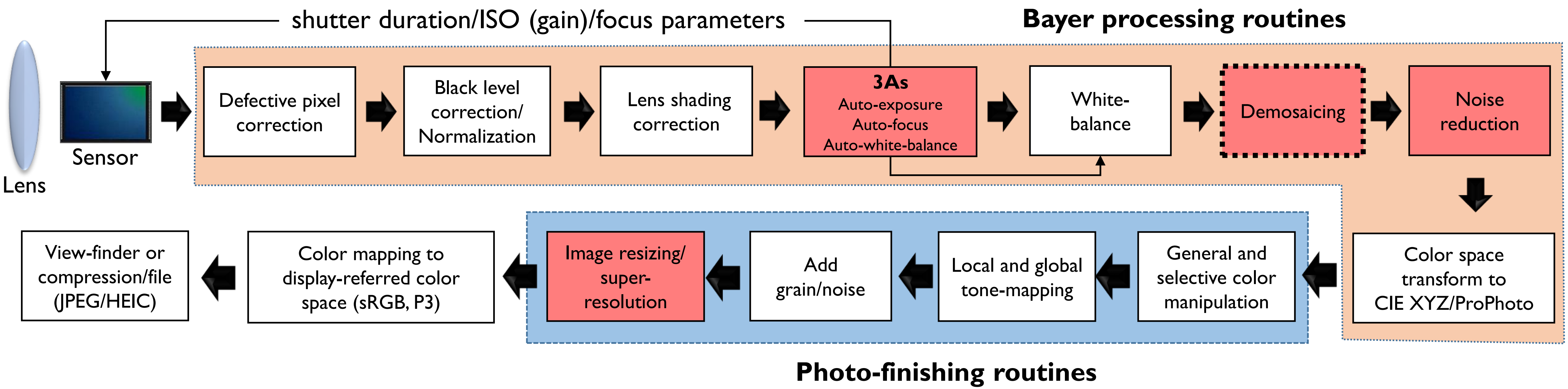
Exploiting two cameras for AWB



Method	Mean	Med	B25%	W25%	Q1	Q3
GW [15]	4.43	3.42	0.90	9.82	1.54	6.11
SoG [24]	3.31	2.63	0.70	7.20	1.18	4.17
GE-1 [46]	4.49	3.03	0.87	10.38	1.40	6.34
GE-2 [46]	4.99	3.28	0.94	11.83	1.54	6.65
WGE [29]	5.77	3.11	0.77	14.75	1.38	7.89
PCA [16]	4.01	2.68	0.69	9.20	1.22	6.07
WP [14]	4.49	3.47	0.93	9.99	1.42	6.09
Gamut Pixel [28]	5.99	3.70	0.90	14.95	1.41	8.65
Gamut Edge [28]	4.99	3.38	0.85	11.63	1.72	7.22
CM [18]	2.80	2.09	0.66	6.12	1.21	3.67
Homography [19] (SoG)	2.70	1.95	0.69	5.88	1.06	3.71
Homography [19] (PCA)	2.97	2.16	0.72	6.47	1.14	4.22
APAP [4] (GW)	2.64	2.00	0.60	5.99	1.02	3.26
APAP [4] (SoG)	2.49	1.75	0.60	5.61	0.88	3.14
APAP [4] (PCA)	2.77	1.83	0.60	6.45	0.94	3.49
SIIE [3]	2.04	1.55	0.51	4.41	0.80	2.80
Quasi U CC [10]	3.57	2.77	0.62	8.04	1.09	5.06
Quasi U CC finetuned [10]	2.68	1.72	0.57	6.25	0.98	3.67
FC4 [32]	2.65	2.06	0.67	5.69	1.12	3.49
FFCC [9]	2.44	1.50	0.40	5.87	0.75	3.19
Ours (200 params)	2.39	1.44	0.46	5.95	0.81	2.81
Ours (470 params)	1.91	1.24	0.36	4.78	0.62	2.22
Ours (1460 params)	**1.69**	**1.09**	**0.37**	**4.02**	**0.59**	**2.02**

State-of-the-art results
with very lightweight network

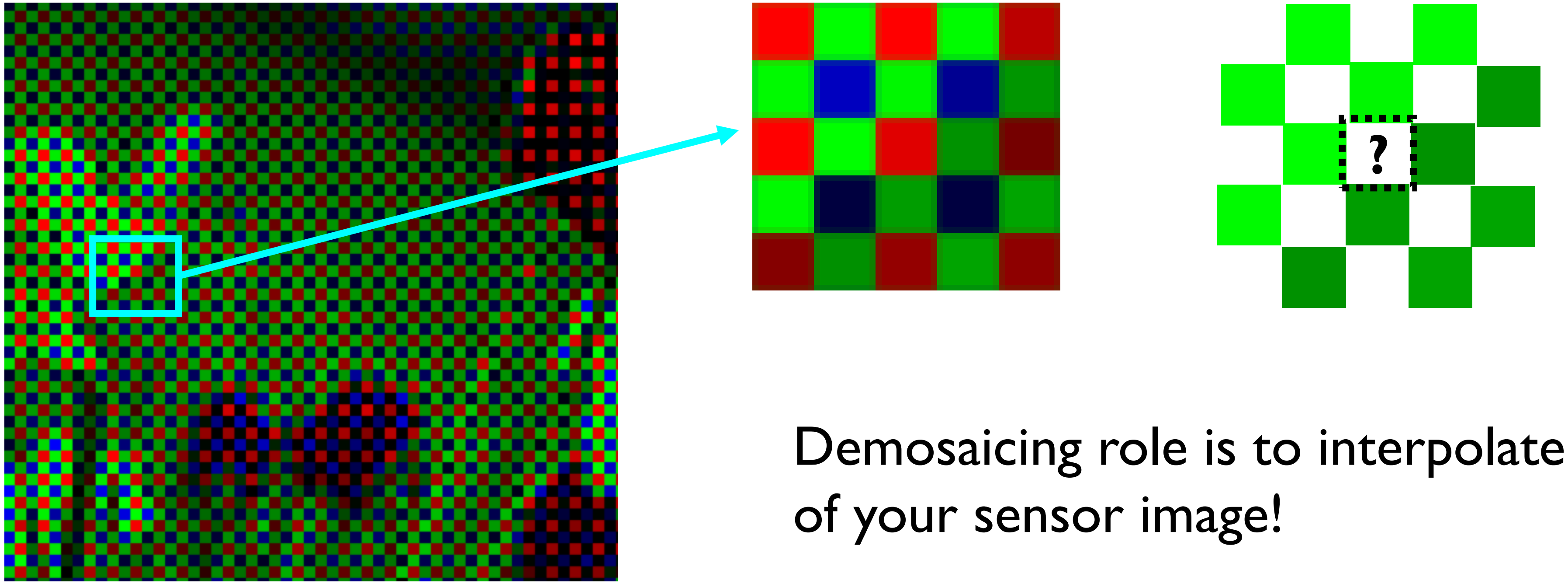
Use deep learning for hard problems



The highlighted components are camera pipeline steps that are challenging and areas AI can make notable gains:

AWB (illumination estimation)
Demosaicing
Noise reduction
Super-resolution

Demosaicing

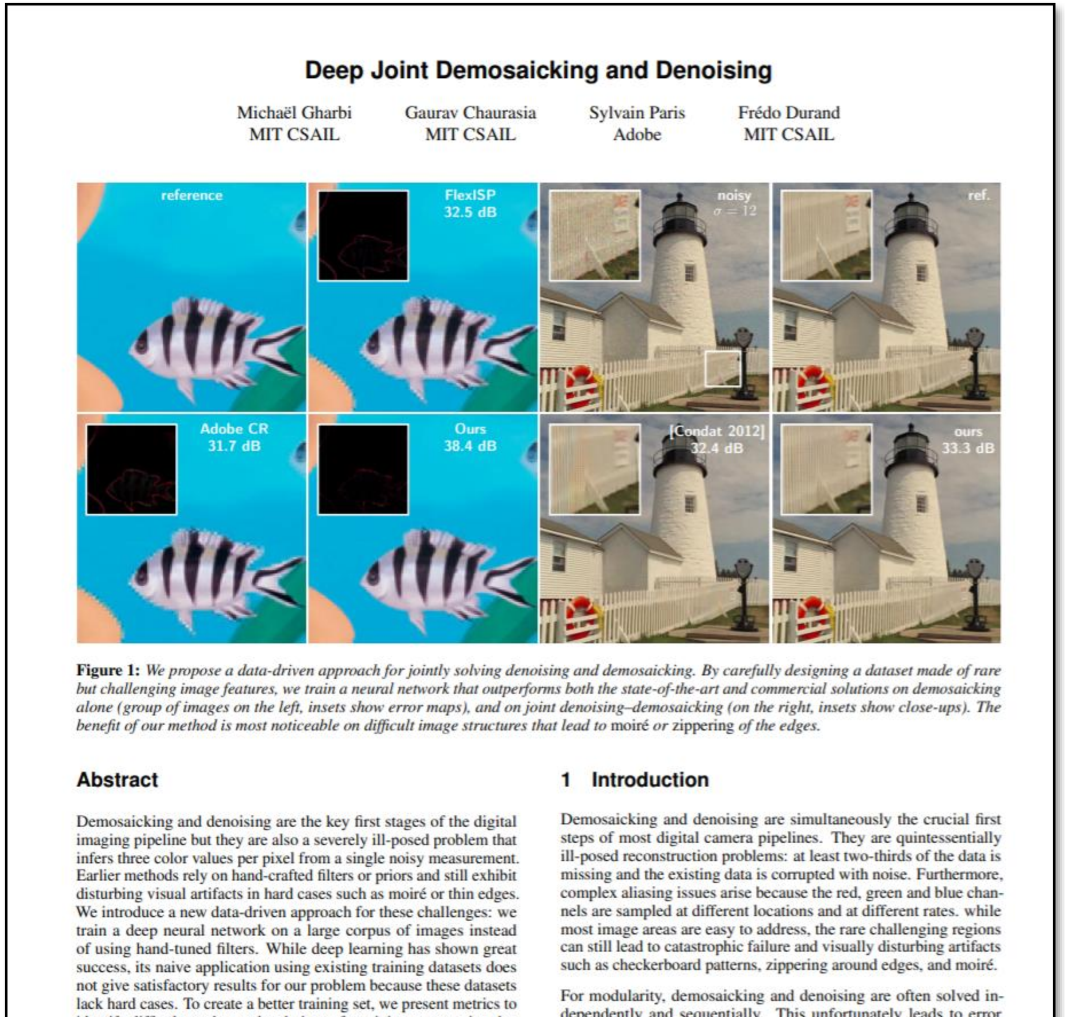


Captured raw-Bayer image

Demosaicing role is to interpolate 2/3 (66%) of your sensor image!

DNN for demosaicing (and denoising)

Gharbi et al SIGGRAPH Asia, 2016



Deep Joint Demosaicking and Denoising

Michaël Gharbi, MIT CSAIL — Gaurav Chaurasia, MIT CSAIL — Sylvain Paris, Adobe — Frédo Durand, MIT CSAIL

reference | FlexISP 32.5 dB | noisy $\sigma = 12$ | ref.

Adobe CR 31.7 dB | Ours 38.4 dB | [Condat 2012] 32.4 dB | ours 33.3 dB

Figure 1: *We propose a data-driven approach for jointly solving denoising and demosaicking. By carefully designing a dataset made of rare but challenging image features, we train a neural network that outperforms both the state-of-the-art and commercial solutions on demosaicking alone (group of images on the left, insets show error maps), and on joint denoising–demosaicking (on the right, insets show close-ups). The benefit of our method is most noticeable on difficult image structures that lead to moiré or zippering of the edges.*

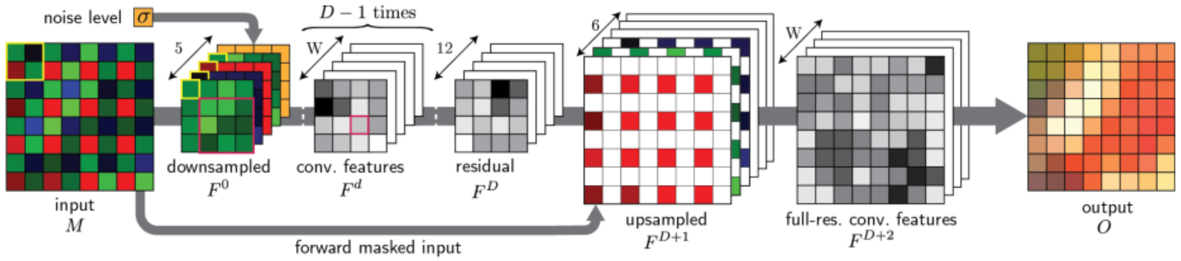
Abstract

Demosaicking and denoising are the key first stages of the digital imaging pipeline but they are also a severely ill-posed problem that infers three color values per pixel from a single noisy measurement. Earlier methods rely on hand-crafted filters or priors and still exhibit disturbing visual artifacts in hard cases such as moiré or thin edges. We introduce a new data-driven approach for these challenges: we train a deep neural network on a large corpus of images instead of using hand-tuned filters. While deep learning has shown great success, its naive application using existing training datasets does not give satisfactory results for our problem because these datasets lack hard cases. To create a better training set, we present metrics to

1 Introduction

Demosaicking and denoising are simultaneously the crucial first steps of most digital camera pipelines. They are quintessentially ill-posed reconstruction problems: at least two-thirds of the data is missing and the existing data is corrupted with noise. Furthermore, complex aliasing issues arise because the red, green and blue channels are sampled at different locations and at different rates. while most image areas are easy to address, the rare challenging regions can still lead to catastrophic failure and visually disturbing artifacts such as checkerboard patterns, zippering around edges, and moiré.

For modularity, demosaicking and denoising are often solved independently and sequentially. This unfortunately leads to error



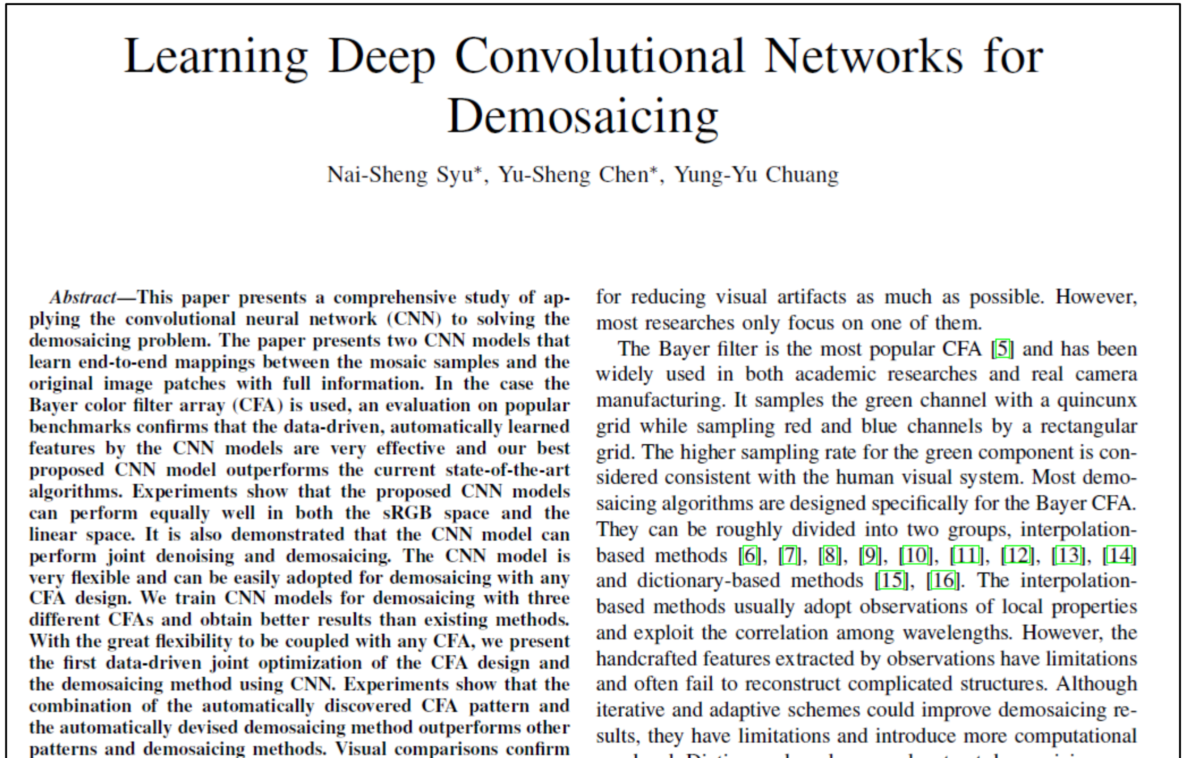
- Early work found that you got denoising for free.

- Network similar to SR-residual.

- Training data/experimental data 100% synthetic.

Deep CNN for demosiacing

Syu et at al - arXiv 2018



Learning Deep Convolutional Networks for Demosaicing

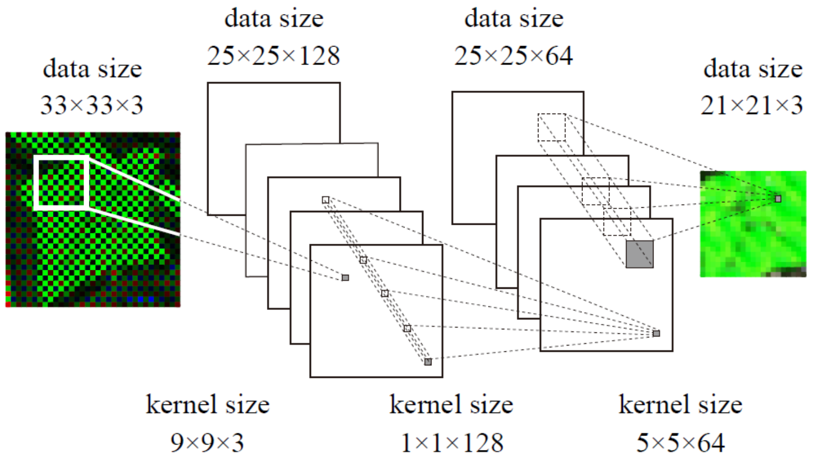
Nai-Sheng Syu*, Yu-Sheng Chen*, Yung-Yu Chuang

Abstract—This paper presents a comprehensive study of applying the convolutional neural network (CNN) to solving the demosaicing problem. The paper presents two CNN models that learn end-to-end mappings between the mosaic samples and the original image patches with full information. In the case the Bayer color filter array (CFA) is used, an evaluation on popular benchmarks confirms that the data-driven, automatically learned features by the CNN models are very effective and our best proposed CNN model outperforms the current state-of-the-art algorithms. Experiments show that the proposed CNN models can perform equally well in both the sRGB space and the linear space. It is also demonstrated that the CNN model can perform joint denoising and demosaicing. The CNN model is very flexible and can be easily adopted for demosaicing with any CFA design. We train CNN models for demosaicing with three different CFAs and obtain better results than existing methods. With the great flexibility to be coupled with any CFA, we present the first data-driven joint optimization of the CFA design and the demosaicing method using CNN. Experiments show that the combination of the automatically discovered CFA pattern and the automatically devised demosaicing method outperforms other patterns and demosaicing methods. Visual comparisons confirm

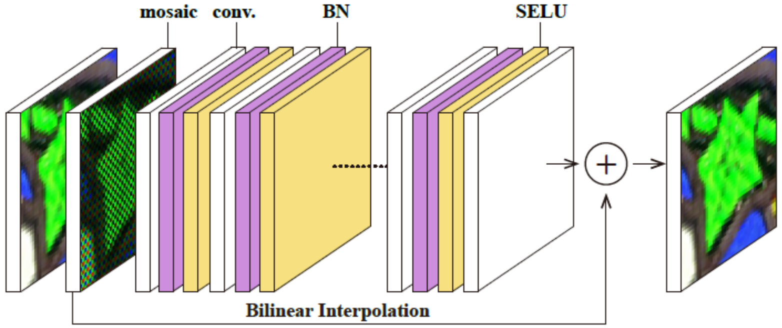
for reducing visual artifacts as much as possible. However, most researches only focus on one of them.

The Bayer filter is the most popular CFA [5] and has been widely used in both academic researches and real camera manufacturing. It samples the green channel with a quincunx grid while sampling red and blue channels by a rectangular grid. The higher sampling rate for the green component is considered consistent with the human visual system. Most demosaicing algorithms are designed specifically for the Bayer CFA. They can be roughly divided into two groups, interpolation-based methods [6], [7], [8], [9], [10], [11], [12], [13], [14] and dictionary-based methods [15], [16]. The interpolation-based methods usually adopt observations of local properties and exploit the correlation among wavelengths. However, the handcrafted features extracted by observations have limitations and often fail to reconstruct complicated structures. Although iterative and adaptive schemes could improve demosaicing results, they have limitations and introduce more computational

Examined #1 – shallow network with deep channels (SRNet)



Examined #2 - deep layers with residual (SR-ResNet)



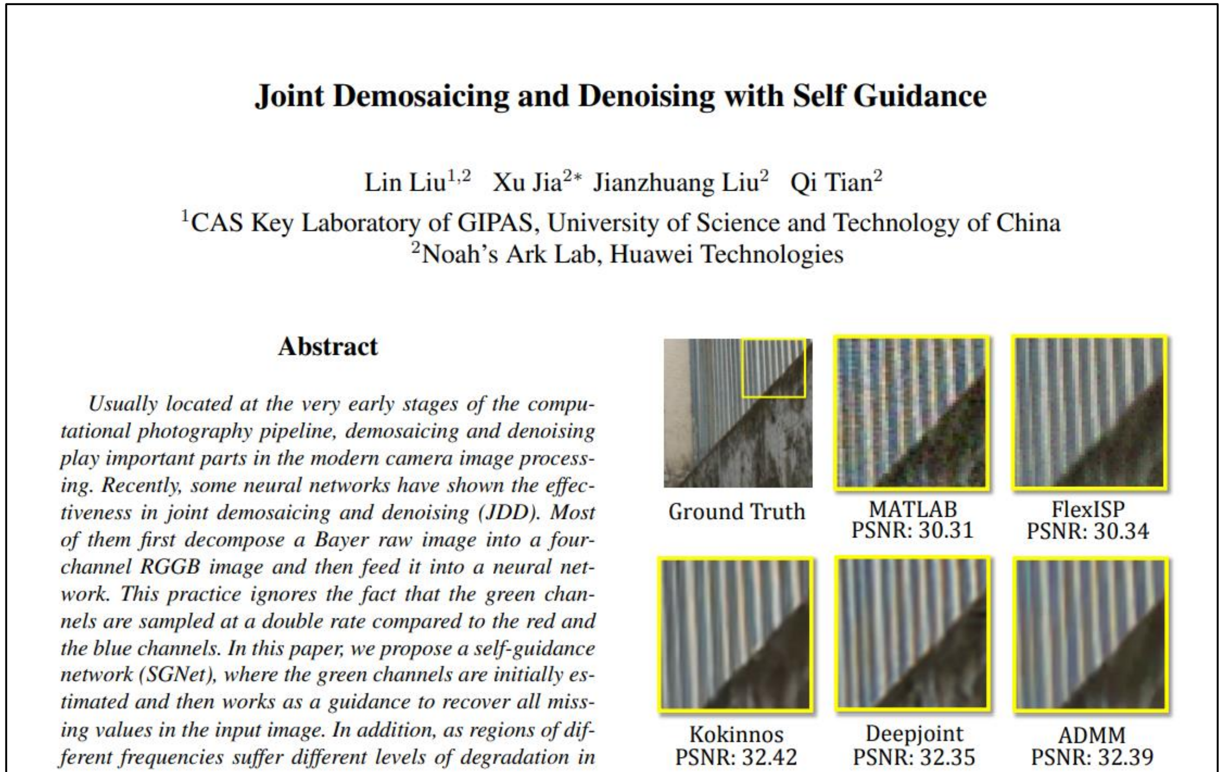
Deep CNN for demosiacing

Algorithm	Kodak (12 photos)				McM (18 photos)				Kodak+McM (30 photos)			
	PSNR			CPSNR	PSNR			CPSNR	PSNR			CPSNR
	R	G	B		R	G	B		R	G	B	
SA [44]	39.8	43.31	39.5	40.54	32.73	34.73	32.1	32.98	35.56	38.16	35.06	36.01
SSD [14]	38.83	40.51	39.08	39.4	35.02	38.27	33.8	35.23	36.54	39.16	35.91	36.9
NLS [16]	42.34	45.68	41.57	42.85	36.02	38.81	34.71	36.15	38.55	41.56	37.46	38.83
CS [45]	41.01	44.17	40.12	41.43	35.56	38.84	34.58	35.92	37.74	40.97	36.8	38.12
ECC [46]	39.87	42.17	39	40.14	36.67	39.99	35.31	36.78	37.95	40.86	36.79	38.12
RI [8]	39.64	42.17	38.87	39.99	36.07	39.99	35.35	36.48	37.5	40.86	36.76	37.88
MLRI [9]	40.53	42.91	39.82	40.88	36.32	39.87	35.35	36.60	38.00	41.08	37.13	38.32
ARI [10]	40.75	43.59	40.16	41.25	37.39	40.68	36.03	37.49	38.73	41.84	37.68	39.00
PAMD [47]	41.88	45.21	41.23	42.44	34.12	36.88	33.31	34.48	37.22	40.21	36.48	37.66
AICC [48]	42.04	44.51	40.57	42.07	35.66	39.21	34.34	35.86	38.21	41.33	36.83	38.34
DMCNN	39.86	42.97	39.18	40.37	36.50	39.34	35.21	36.62	37.85	40.79	36.79	38.12
DMCNN-VD	43.28	46.10	41.99	43.45	39.69	42.53	37.76	39.45	41.13	43.96	39.45	41.05

Paper showed that Deep SR-ResNet is a good DNN for demosiacing too.

More recent approach

Liu et al CVPR'21 [Huawei]


Joint Demosaicing and Denoising with Self Guidance

Lin Liu[1,2] Xu Jia[2*] Jianzhuang Liu[2] Qi Tian[2]
[1]CAS Key Laboratory of GIPAS, University of Science and Technology of China
[2]Noah's Ark Lab, Huawei Technologies

Abstract

Usually located at the very early stages of the computational photography pipeline, demosaicing and denoising play important parts in the modern camera image processing. Recently, some neural networks have shown the effectiveness in joint demosaicing and denoising (JDD). Most of them first decompose a Bayer raw image into a four-channel RGGB image and then feed it into a neural network. This practice ignores the fact that the green channels are sampled at a double rate compared to the red and the blue channels. In this paper, we propose a self-guidance network (SGNet), where the green channels are initially estimated and then works as a guidance to recover all missing values in the input image. In addition, as regions of different frequencies suffer different levels of degradation in

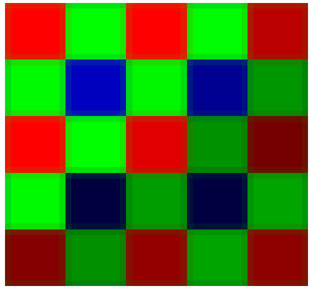
Ground Truth
MATLAB PSNR: 30.31
FlexISP PSNR: 30.34
Kokinnos PSNR: 32.42
Deepjoint PSNR: 32.35
ADMM PSNR: 32.39

Considers the following:

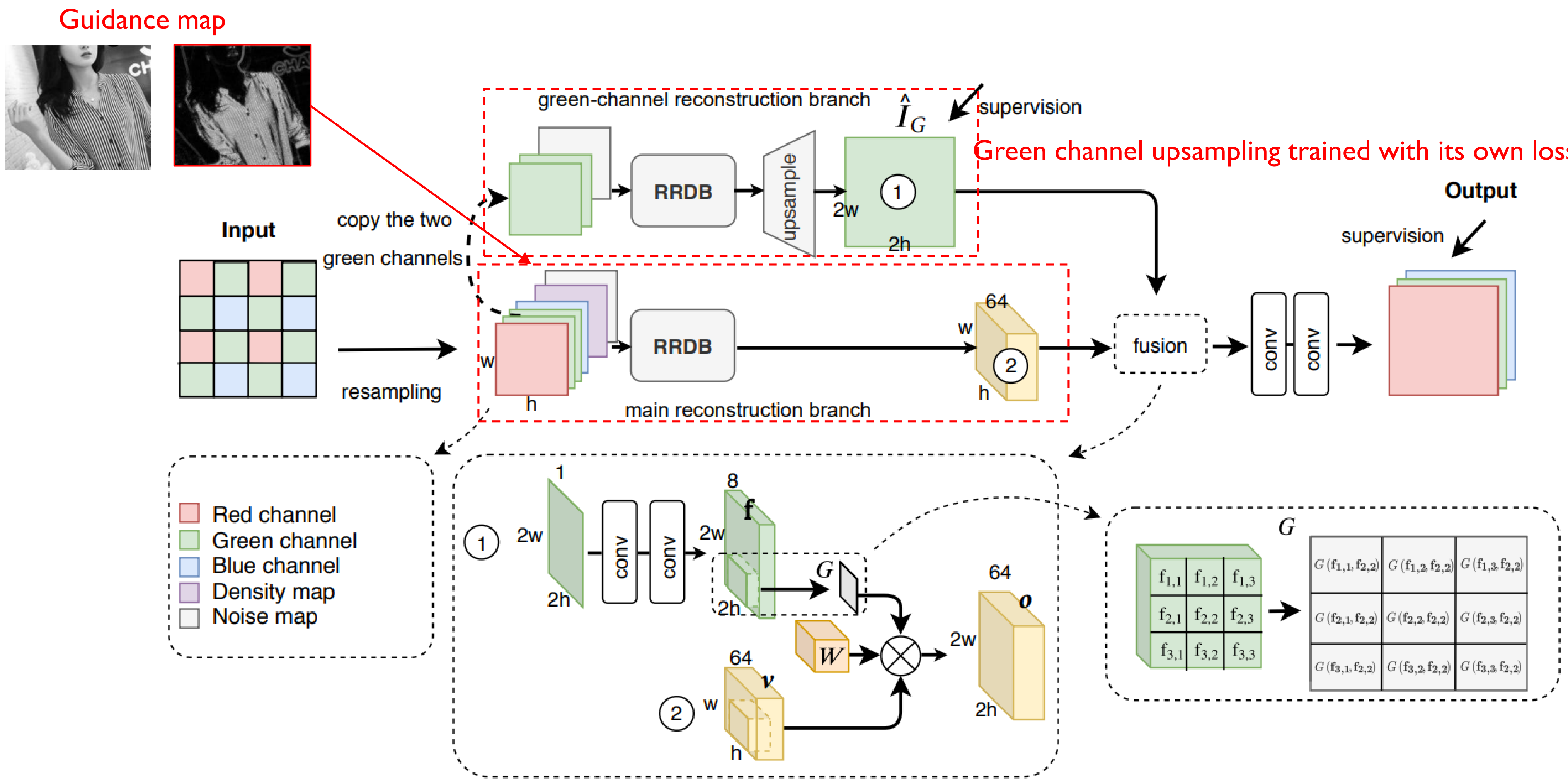
(1) Demosaicing is hardest in high-frequency areas

Predicting
hard regions

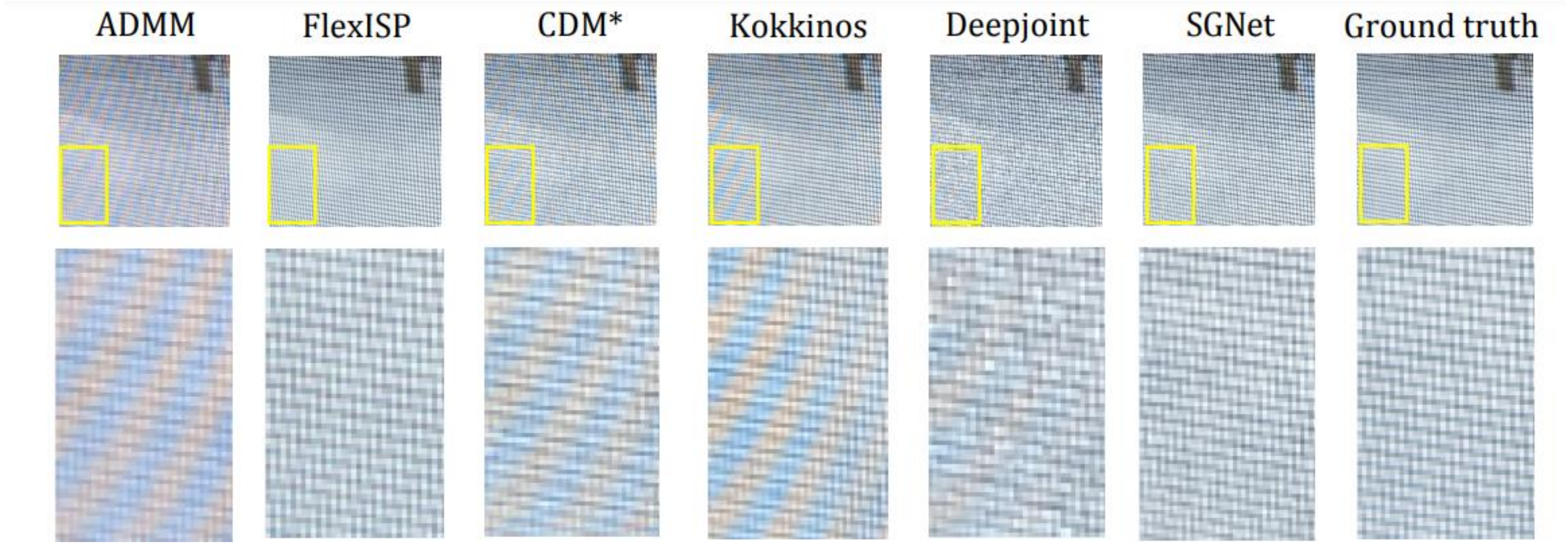
(2) Green channel
most reliable (since
we have a lot of green)



Demosiacing with "self guidance"

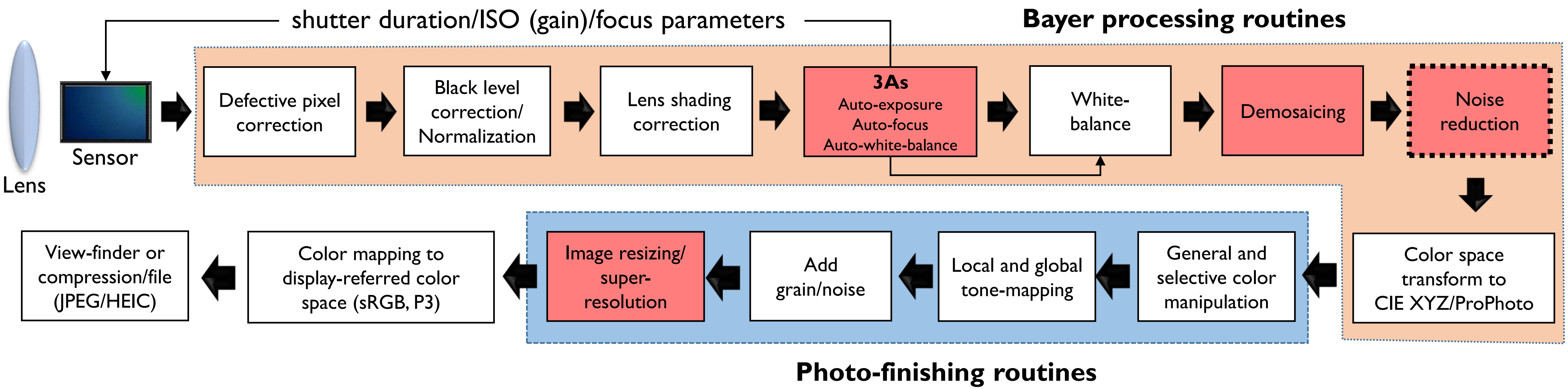


Produces nice visual results



Self-guidance NET

Use deep learning for hard problems



The highlighted components are camera pipeline steps that are challenging and areas AI can make notable gains:

- AWB (illumination estimation)
- Demosaicing
- **Noise reduction**
- Super-resolution

Non-deep-learning noise reduction

- One of the best-performing methods was based on non-local means (2007).
- Block-matching with 3D filtering [BM3D]
- It is slow, but works well.

Dabov et al TIP'07

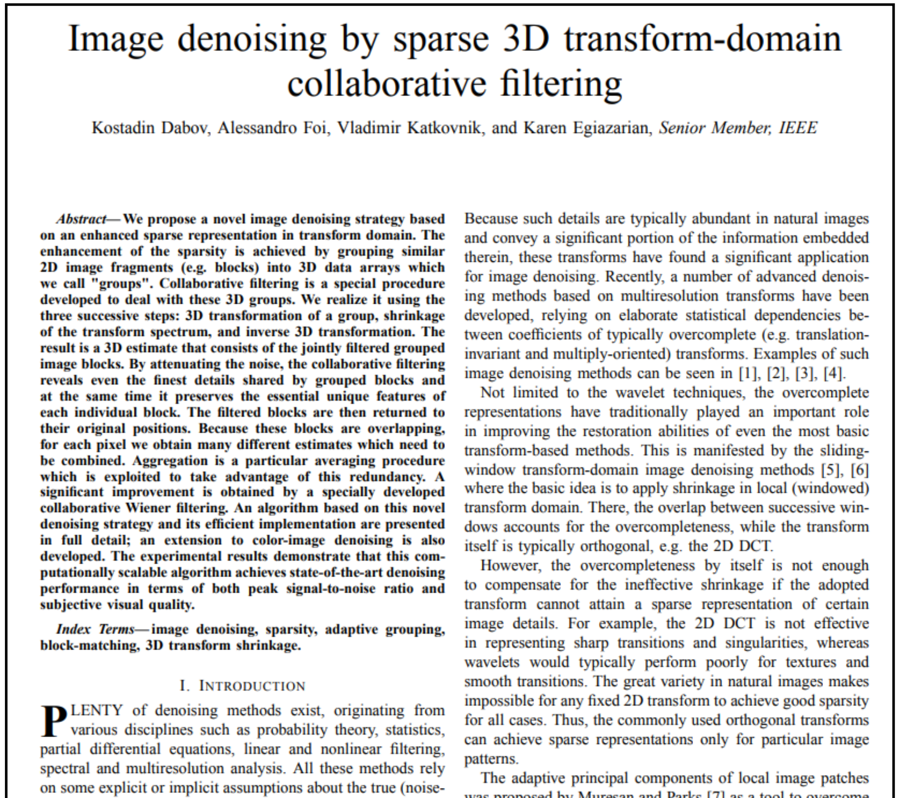


Image denoising by sparse 3D transform-domain collaborative filtering

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, *Senior Member, IEEE*

***Abstract*—We propose a novel image denoising strategy based on an enhanced sparse representation in transform domain. The enhancement of the sparsity is achieved by grouping similar 2D image fragments (e.g. blocks) into 3D data arrays which we call "groups". Collaborative filtering is a special procedure developed to deal with these 3D groups. We realize it using the three successive steps: 3D transformation of a group, shrinkage of the transform spectrum, and inverse 3D transformation. The result is a 3D estimate that consists of the jointly filtered grouped image blocks. By attenuating the noise, the collaborative filtering reveals even the finest details shared by grouped blocks and at the same time it preserves the essential unique features of each individual block. The filtered blocks are then returned to their original positions. Because these blocks are overlapping, for each pixel we obtain many different estimates which need to be combined. Aggregation is a particular averaging procedure which is exploited to take advantage of this redundancy. A significant improvement is obtained by a specially developed collaborative Wiener filtering. An algorithm based on this novel denoising strategy and its efficient implementation are presented in full detail; an extension to color-image denoising is also developed. The experimental results demonstrate that this computationally scalable algorithm achieves state-of-the-art denoising performance in terms of both peak signal-to-noise ratio and subjective visual quality.**

***Index Terms*—image denoising, sparsity, adaptive grouping, block-matching, 3D transform shrinkage.**

I. Introduction

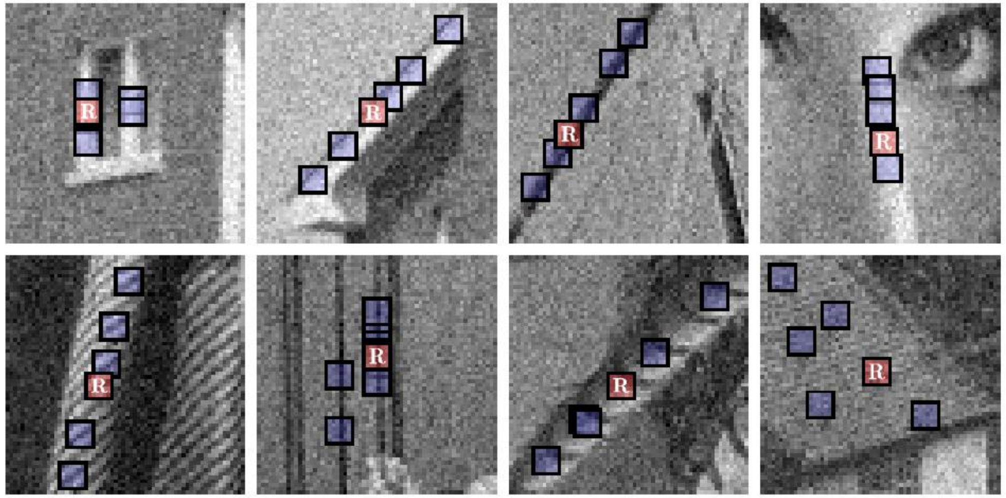
PLENTY of denoising methods exist, originating from various disciplines such as probability theory, statistics, partial differential equations, linear and nonlinear filtering, spectral and multiresolution analysis. All these methods rely on some explicit or implicit assumptions about the true (noise-

Because such details are typically abundant in natural images and convey a significant portion of the information embedded therein, these transforms have found a significant application for image denoising. Recently, a number of advanced denoising methods based on multiresolution transforms have been developed, relying on elaborate statistical dependencies between coefficients of typically overcomplete (e.g. translation-invariant and multiply-oriented) transforms. Examples of such image denoising methods can be seen in [1], [2], [3], [4].

Not limited to the wavelet techniques, the overcomplete representations have traditionally played an important role in improving the restoration abilities of even the most basic transform-based methods. This is manifested by the sliding-window transform-domain image denoising methods [5], [6] where the basic idea is to apply shrinkage in local (windowed) transform domain. There, the overlap between successive windows accounts for the overcompleteness, while the transform itself is typically orthogonal, e.g. the 2D DCT.

However, the overcompleteness by itself is not enough to compensate for the ineffective shrinkage if the adopted transform cannot attain a sparse representation of certain image details. For example, the 2D DCT is not effective in representing sharp transitions and singularities, whereas wavelets would typically perform poorly for textures and smooth transitions. The great variety in natural images makes impossible for any fixed 2D transform to achieve good sparsity for all cases. Thus, the commonly used orthogonal transforms can achieve sparse representations only for particular image patterns.

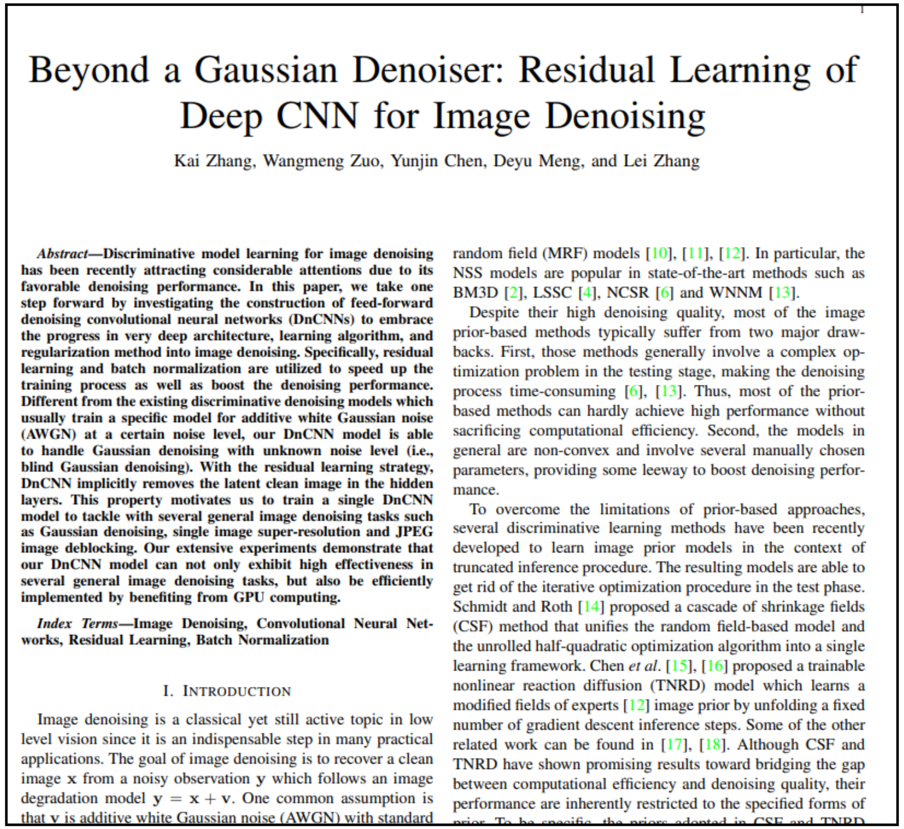
The adaptive principal components of local image patches



For small reference patch R, find similar patches.
Average the patches.

DNN for denoising (DnDNN)

Zhang et a. TIP'17



1

Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang

Abstract—**Discriminative model learning for image denoising has been recently attracting considerable attentions due to its favorable denoising performance. In this paper, we take one step forward by investigating the construction of feed-forward denoising convolutional neural networks (DnCNNs) to embrace the progress in very deep architecture, learning algorithm, and regularization method into image denoising. Specifically, residual learning and batch normalization are utilized to speed up the training process as well as boost the denoising performance. Different from the existing discriminative denoising models which usually train a specific model for additive white Gaussian noise (AWGN) at a certain noise level, our DnCNN model is able to handle Gaussian denoising with unknown noise level (i.e., blind Gaussian denoising). With the residual learning strategy, DnCNN implicitly removes the latent clean image in the hidden layers. This property motivates us to train a single DnCNN model to tackle with several general image denoising tasks such as Gaussian denoising, single image super-resolution and JPEG image deblocking. Our extensive experiments demonstrate that our DnCNN model can not only exhibit high effectiveness in several general image denoising tasks, but also be efficiently implemented by benefiting from GPU computing.**

Index Terms—**Image Denoising, Convolutional Neural Networks, Residual Learning, Batch Normalization**

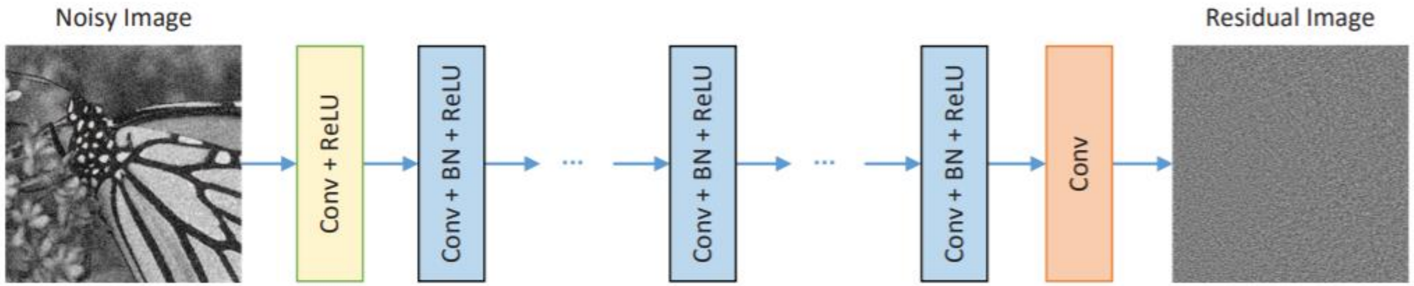
I. Introduction

Image denoising is a classical yet still active topic in low level vision since it is an indispensable step in many practical applications. The goal of image denoising is to recover a clean image $\mathbf{x}$ from a noisy observation $\mathbf{y}$ which follows an image degradation model $\mathbf{y} = \mathbf{x} + \mathbf{v}$. One common assumption is that $\mathbf{v}$ is additive white Gaussian noise (AWGN) with standard

random field (MRF) models [10], [11], [12]. In particular, the NSS models are popular in state-of-the-art methods such as BM3D [2], LSSC [4], NCSR [6] and WNNM [13].

Despite their high denoising quality, most of the image prior-based methods typically suffer from two major drawbacks. First, those methods generally involve a complex optimization problem in the testing stage, making the denoising process time-consuming [6], [13]. Thus, most of the prior-based methods can hardly achieve high performance without sacrificing computational efficiency. Second, the models in general are non-convex and involve several manually chosen parameters, providing some leeway to boost denoising performance.

To overcome the limitations of prior-based approaches, several discriminative learning methods have been recently developed to learn image prior models in the context of truncated inference procedure. The resulting models are able to get rid of the iterative optimization procedure in the test phase. Schmidt and Roth [14] proposed a cascade of shrinkage fields (CSF) method that unifies the random field-based model and the unrolled half-quadratic optimization algorithm into a single learning framework. Chen *et al*. [15], [16] proposed a trainable nonlinear reaction diffusion (TNRD) model which learns a modified fields of experts [12] image prior by unfolding a fixed number of gradient descent inference steps. Some of the other related work can be found in [17], [18]. Although CSF and TNRD have shown promising results toward bridging the gap between computational efficiency and denoising quality, their performance are inherently restricted to the specified forms of



- Straight-forward network based on deep residual learning (Kim SR-ResNet).

- Introduced batch normalization to the network.

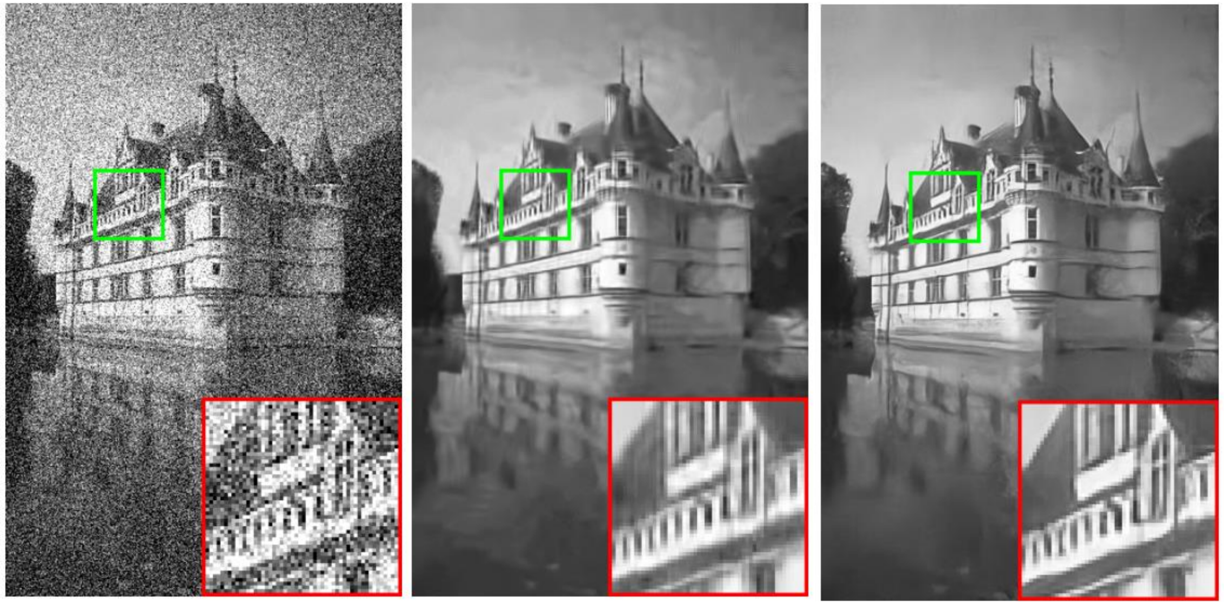
- Predicts the residual noise layer.

DnCNN result

Methods	BM3D	WNNM	EPLL	MLP	CSF	TNRD	DnCNN-S	DnCNN-B
$\sigma = 15$	31.07	31.37	31.21	-	31.24	31.42	**31.73**	31.61
$\sigma = 25$	28.57	28.83	28.68	28.96	28.74	28.92	**29.23**	29.16
$\sigma = 50$	25.62	25.87	25.67	26.03	-	25.97	**26.23**	**26.23**

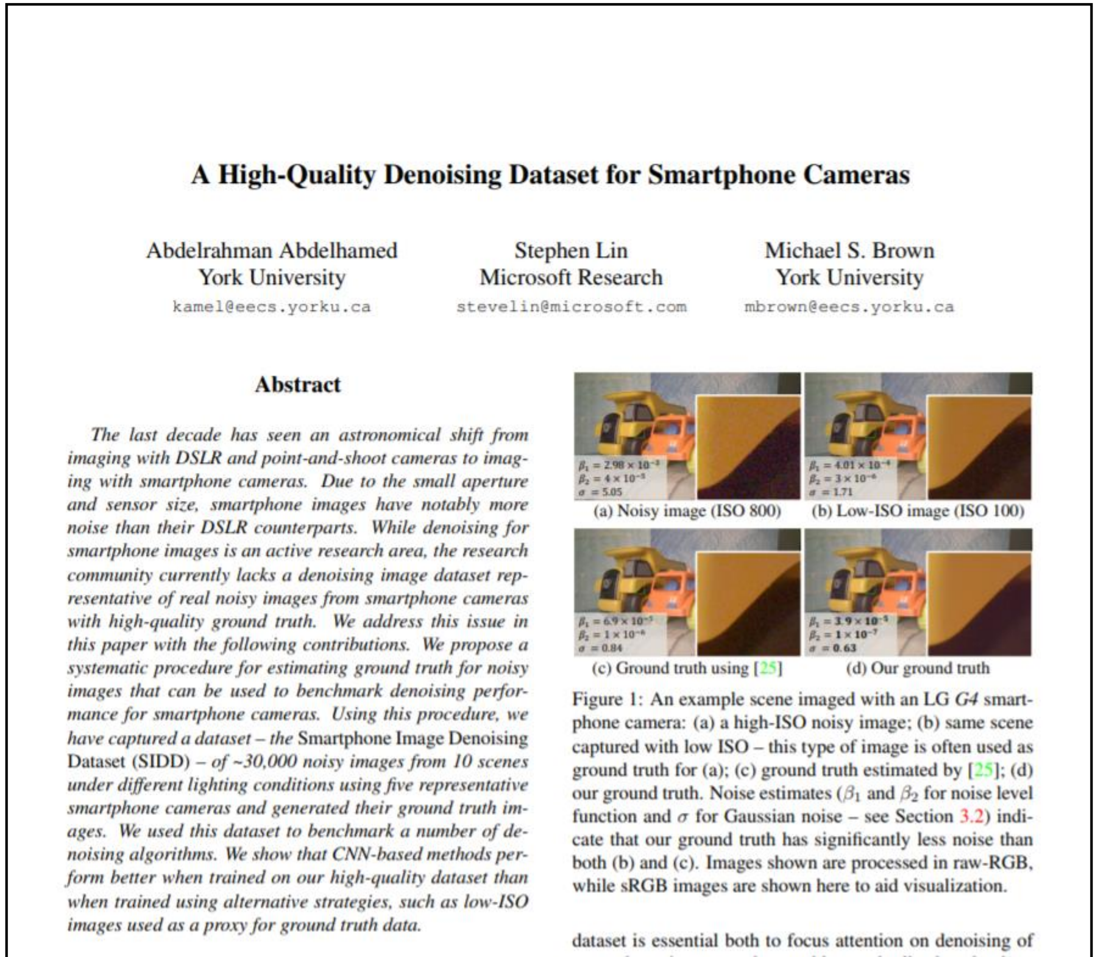
- Method trained on synthetic noise data.
- Beats BM3D and is much faster.
- BM3D does not require training data!

BM3D DnCNN



Need for real denoising dataset

Abdelhamed et al CVPR 2018



A High-Quality Denoising Dataset for Smartphone Cameras

Abdelrahman Abdelhamed
York University
kamel@eecs.yorku.ca

Stephen Lin
Microsoft Research
stevelin@microsoft.com

Michael S. Brown
York University
mbrown@eecs.yorku.ca

Abstract

The last decade has seen an astronomical shift from imaging with DSLR and point-and-shoot cameras to imaging with smartphone cameras. Due to the small aperture and sensor size, smartphone images have notably more noise than their DSLR counterparts. While denoising for smartphone images is an active research area, the research community currently lacks a denoising image dataset representative of real noisy images from smartphone cameras with high-quality ground truth. We address this issue in this paper with the following contributions. We propose a systematic procedure for estimating ground truth for noisy images that can be used to benchmark denoising performance for smartphone cameras. Using this procedure, we have captured a dataset – the Smartphone Image Denoising Dataset (SIDD) *– of ~30,000 noisy images from 10 scenes under different lighting conditions using five representative smartphone cameras and generated their ground truth images. We used this dataset to benchmark a number of denoising algorithms. We show that CNN-based methods perform better when trained on our high-quality dataset than when trained using alternative strategies, such as low-ISO images used as a proxy for ground truth data.*

(a) Noisy image (ISO 800) (b) Low-ISO image (ISO 100)

(c) Ground truth using [25] (d) Our ground truth

Figure 1: An example scene imaged with an LG *G4* smartphone camera: (a) a high-ISO noisy image; (b) same scene captured with low ISO – this type of image is often used as ground truth for (a); (c) ground truth estimated by [25]; (d) our ground truth. Noise estimates (β_1 and β_2 for noise level function and σ for Gaussian noise – see Section 3.2) indicate that our ground truth has significantly less noise than both (b) and (c). Images shown are processed in raw-RGB, while sRGB images are shown here to aid visualization.

dataset is essential both to focus attention on denoising of

SIDD: Smartphone Image Denoising Dataset

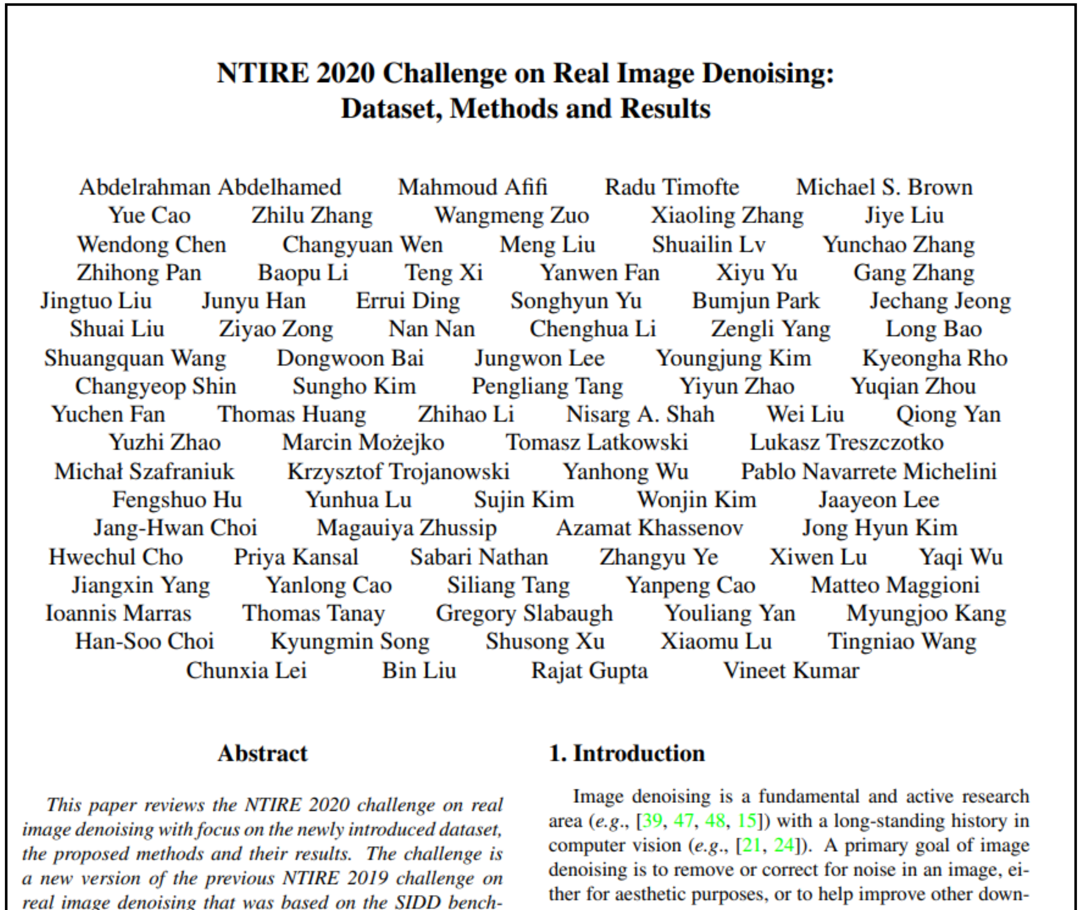
- -30,000 images
- -5 cameras
- -160 scene instances
- -15 ISO settings
- -Direct current lighting
- -Three illuminations

Interesting finding

- When trained on synthetic only, BM3D beat DnCNN
- When trained on real data, DnCNN wins
- Implies noise models in literature are not accurate

Denoising contest at CVPR'20

NTIRE denoising contest at CVPR'20



NTIRE 2020 Challenge on Real Image Denoising:
Dataset, Methods and Results

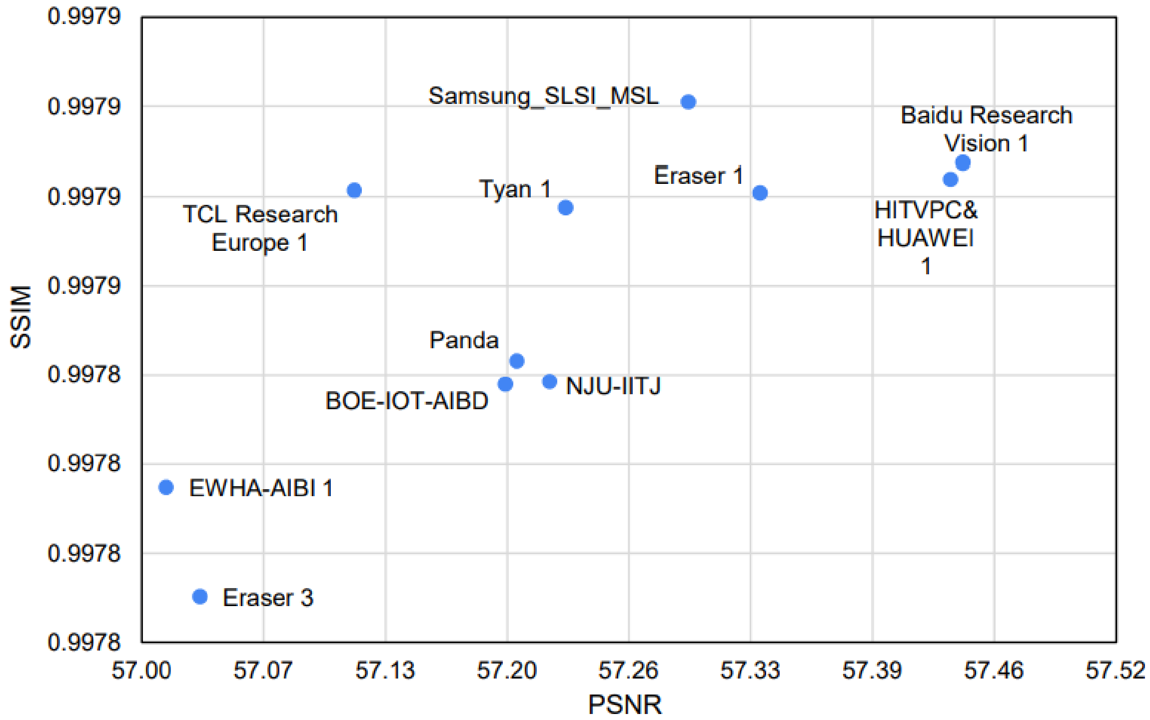
Abdelrahman Abdelhamed Mahmoud Afifi Radu Timofte Michael S. Brown
Yue Cao Zhilu Zhang Wangmeng Zuo Xiaoling Zhang Jiye Liu
Wendong Chen Changyuan Wen Meng Liu Shuailin Lv Yunchao Zhang
Zhihong Pan Baopu Li Teng Xi Yanwen Fan Xiyu Yu Gang Zhang
Jingtuo Liu Junyu Han Errui Ding Songhyun Yu Bumjun Park Jechang Jeong
Shuai Liu Ziyao Zong Nan Nan Chenghua Li Zengli Yang Long Bao
Shuangquan Wang Dongwoon Bai Jungwon Lee Youngjung Kim Kyeongha Rho
Changyeop Shin Sungho Kim Pengliang Tang Yiyun Zhao Yuqian Zhou
Yuchen Fan Thomas Huang Zhihao Li Nisarg A. Shah Wei Liu Qiong Yan
Yuzhi Zhao Marcin Możejko Tomasz Latkowski Lukasz Treszczotko
Michał Szafraniuk Krzysztof Trojanowski Yanhong Wu Pablo Navarrete Michelini
Fengshuo Hu Yunhua Lu Sujin Kim Wonjin Kim Jaayeon Lee
Jang-Hwan Choi Magauiya Zhussip Azamat Khassenov Jong Hyun Kim
Hwechul Cho Priya Kansal Sabari Nathan Zhangyu Ye Xiwen Lu Yaqi Wu
Jiangxin Yang Yanlong Cao Siliang Tang Yanpeng Cao Matteo Maggioni
Ioannis Marras Thomas Tanay Gregory Slabaugh Youliang Yan Myungjoo Kang
Han-Soo Choi Kyungmin Song Shusong Xu Xiaomu Lu Tingniao Wang
Chunxia Lei Bin Liu Rajat Gupta Vineet Kumar

Abstract

This paper reviews the NTIRE 2020 challenge on real image denoising with focus on the newly introduced dataset, the proposed methods and their results. The challenge is a new version of the previous NTIRE 2019 challenge on real image denoising that was based on the SIDD bench-

1. Introduction

Image denoising is a fundamental and active research area (*e.g.*, [39, 47, 48, 15]) with a long-standing history in computer vision (*e.g.*, [21, 24]). A primary goal of image denoising is to remove or correct for noise in an image, either for aesthetic purposes, or to help improve other down-



- Winning solutions (Baidu) relied on neural architecture search
- Samsung method used a U-net + multi-scale residual nets

ISPs with multi-frame (burst) imaging

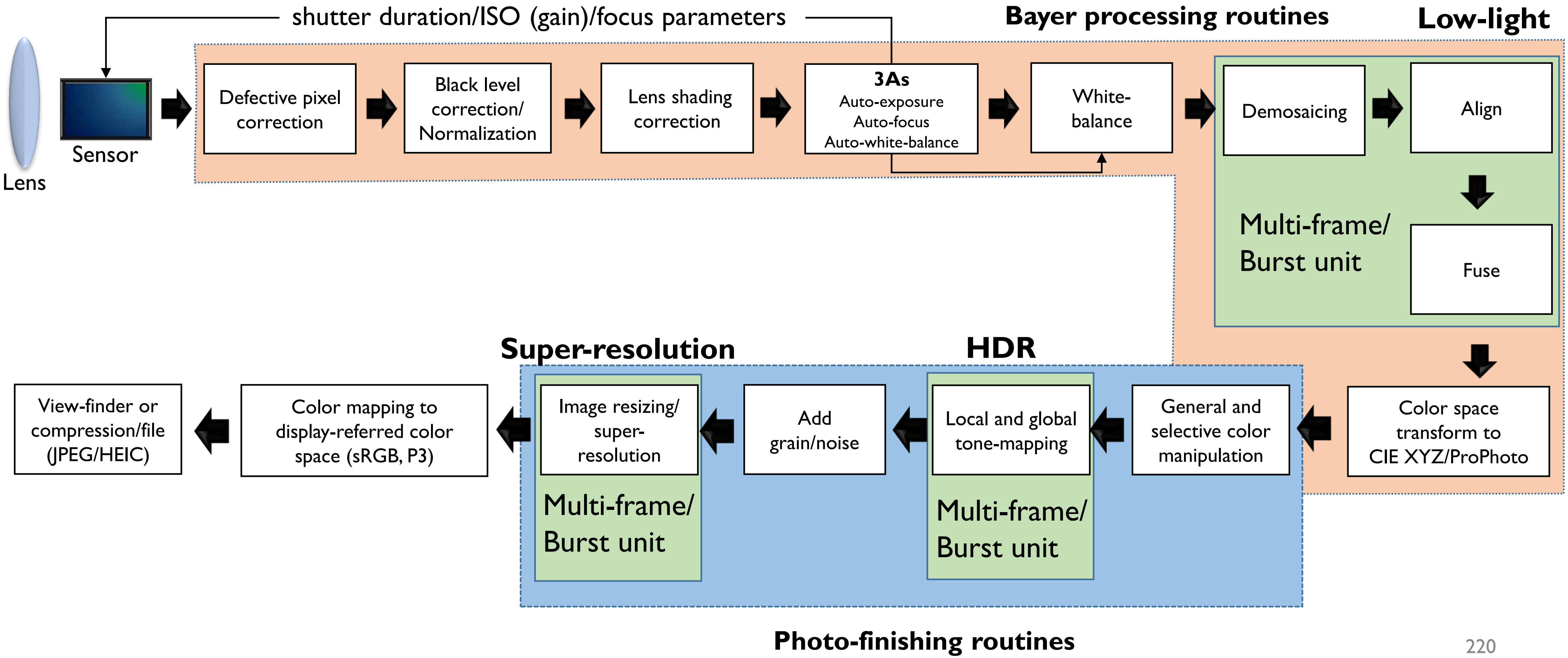
Why multi-frame?

- Two primary applications where is currently multi-frame is used
  - Low-light imaging (or "night mode" or "extended ISO")
  - High-dynamic-range (HDR) imaging

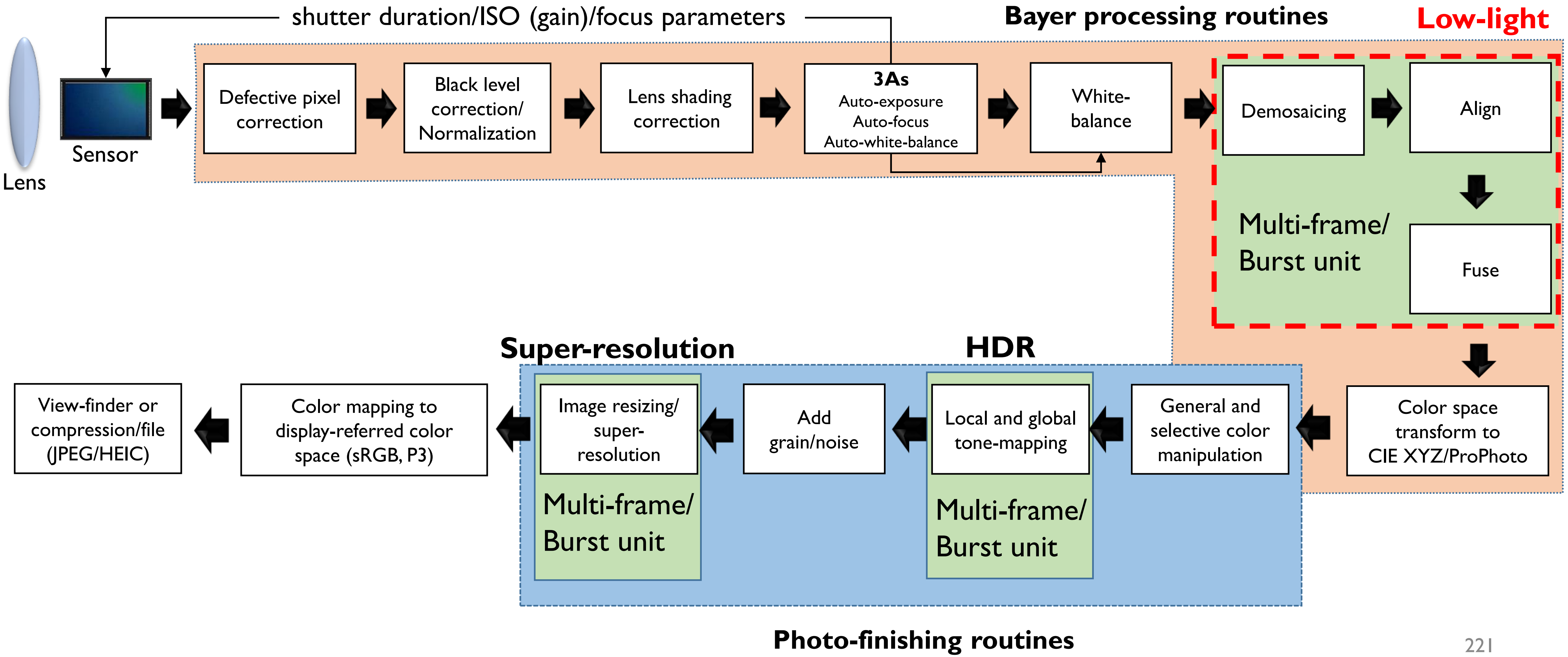
- Many ISP now support multi-frame image.

- Possible to do super-res with multi-frame.
  - For time sake, I won't discuss those methods

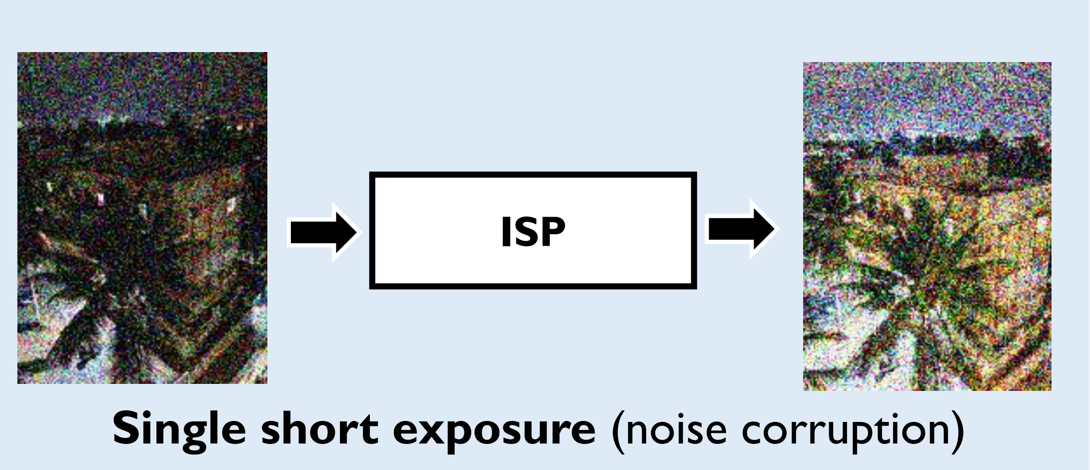
ISP with multi-frame

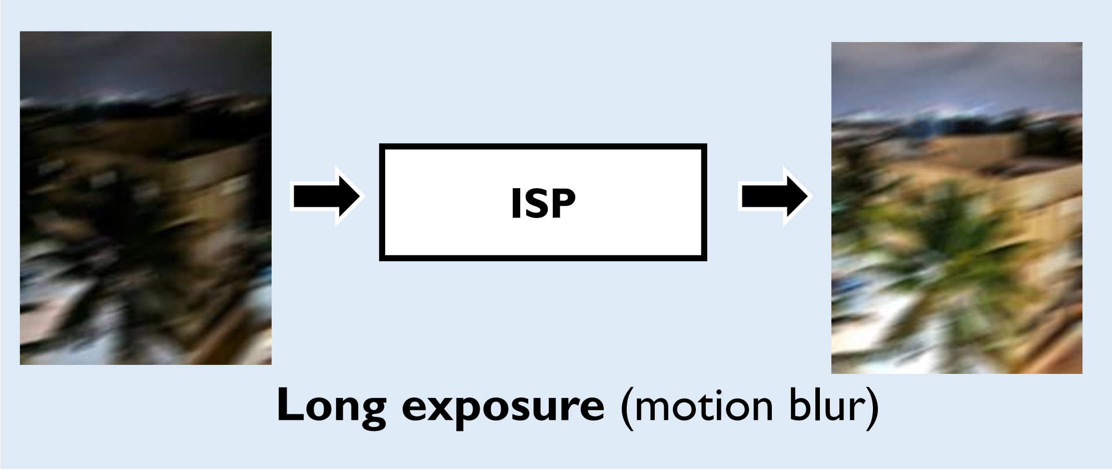


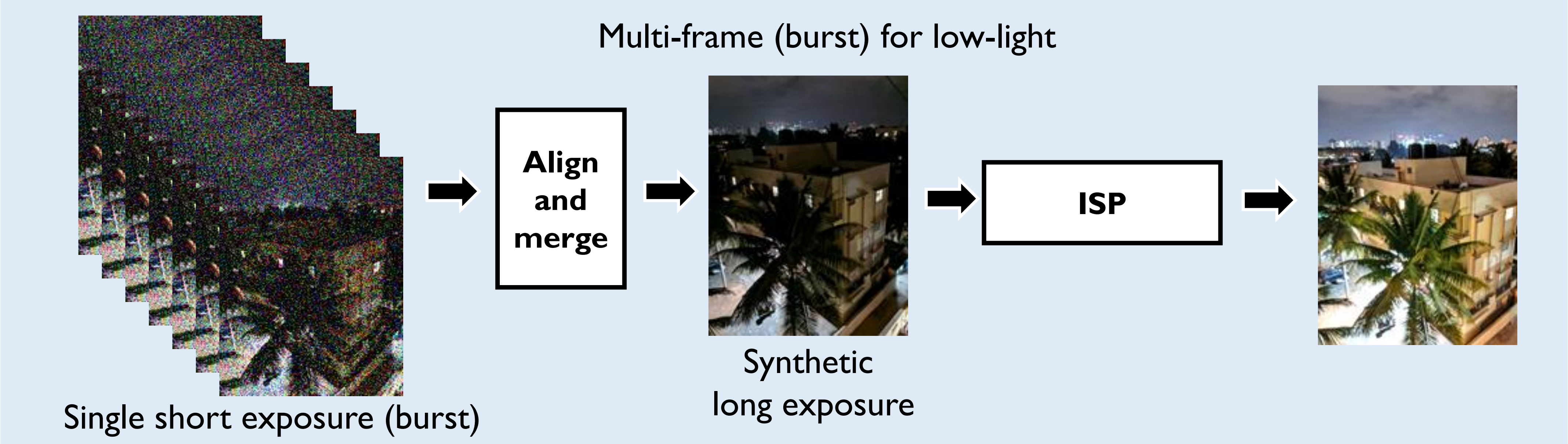
ISP with multi-frame



Low-light imaging







Low-light imaging is essentially a noise-reduction problem.

Multi-frame for low-light

Early work on low-light image is from Samsung SAIT

Moon et al – ICCE 2013

A Fast Low-Light Multi-Image Fusion with Online Image Restoration

Young-Su Moon, Shi-Hwa Lee, Yong-Min Tai, and Junguk Cho
Samsung Advanced Institute of Technology, Samsung Electronics, Korea

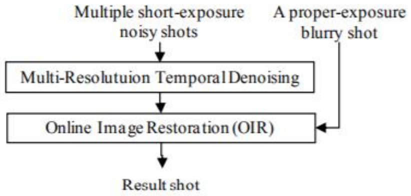
***Abstract*—This paper presents a new low-light multi-frame fusion algorithm to get a bright and clear shot even under dark conditions. To this end, using multiple short-exposure images and one proper-exposure blurry image as an input, a new hierarchical block-wise temporal noise filtering is done. Finally, an online image restoration of the denoising result is conducted along with the blurry image input. Test results on real low-light scene show its effectiveness like fast processing speed and satisfactory visual quality.**

I. INTRODUCTION

Digital camera photos taken under a low-light condition reveal significant image artifacts such as motion blur by long-exposure shooting or strong noise corruption by High-ISO setting. Furthermore, as camera sensor's resolution increases, such artifacts are getting worse due to lack of incoming lights on each sensor cell.

To solve it, many research works have been studied. In a

shooting mode need to be geometrically aligned. To achieve this effectively, global image motion between a reference short-exposure input image and other short-exposure input images is estimated with a fast and effective method using a translation model [4]. For convenience, the first short-exposure input image is selected as the reference. Since actual image motion between the input frames is complicated, subsequent block-based local motion estimation is required.



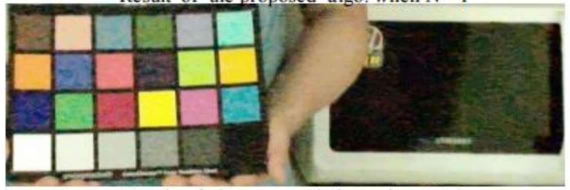
Proposed alignment
(1) Works on a Laplacian pyramid.
(2) global motion alignment
(3) local motion correction
(4) Temporal fusion (to avoid ghosting)


Just blurry proper-exposure input image


Result of (just TA+OIR) of the four noisy input images


Result of just (OIR) of just one noisy input image

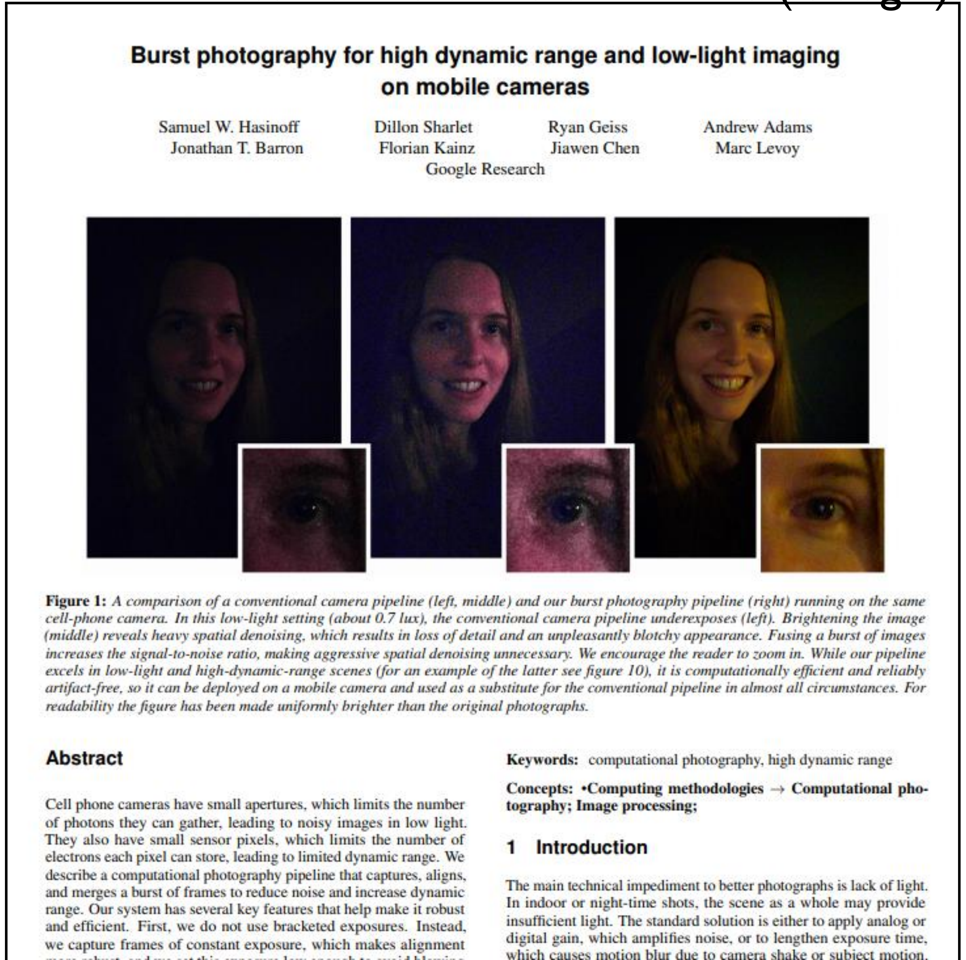

Result of the proposed algo. when N =1


Result of the proposed algo. when N =3

Burst for low-light and HDR

No deep learning – this paper describes a fast/robust alignment estimation based on bilateral fitter.
Robust merging is guided by a reference frame.
All is done in the Bayer/RAW frame.

Hassinoff et al SIGGRAPH'16 (Google)



Burst photography for high dynamic range and low-light imaging on mobile cameras

Samuel W. Hasinoff, Jonathan T. Barron, Dillon Sharlet, Florian Kainz, Ryan Geiss, Jiawen Chen, Andrew Adams, Marc Levoy
Google Research

Figure 1: *A comparison of a conventional camera pipeline (left, middle) and our burst photography pipeline (right) running on the same cell-phone camera. In this low-light setting (about 0.7 lux), the conventional camera pipeline underexposes (left). Brightening the image (middle) reveals heavy spatial denoising, which results in loss of detail and an unpleasantly blotchy appearance. Fusing a burst of images increases the signal-to-noise ratio, making aggressive spatial denoising unnecessary. We encourage the reader to zoom in. While our pipeline excels in low-light and high-dynamic-range scenes (for an example of the latter see figure 10), it is computationally efficient and reliably artifact-free, so it can be deployed on a mobile camera and used as a substitute for the conventional pipeline in almost all circumstances. For readability the figure has been made uniformly brighter than the original photographs.*

Abstract

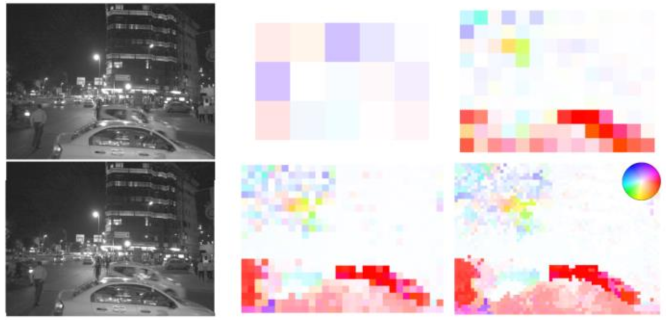
Cell phone cameras have small apertures, which limits the number of photons they can gather, leading to noisy images in low light. They also have small sensor pixels, which limits the number of electrons each pixel can store, leading to limited dynamic range. We describe a computational photography pipeline that captures, aligns, and merges a burst of frames to reduce noise and increase dynamic range. Our system has several key features that help make it robust and efficient. First, we do not use bracketed exposures. Instead, we capture frames of constant exposure, which makes alignment

Keywords: computational photography, high dynamic range

Concepts: •**Computing methodologies** → **Computational photography; Image processing;**

1 Introduction

The main technical impediment to better photographs is lack of light. In indoor or night-time shots, the scene as a whole may provide insufficient light. The standard solution is either to apply analog or digital gain, which amplifies noise, or to lengthen exposure time, which causes motion blur due to camera shake or subject motion.



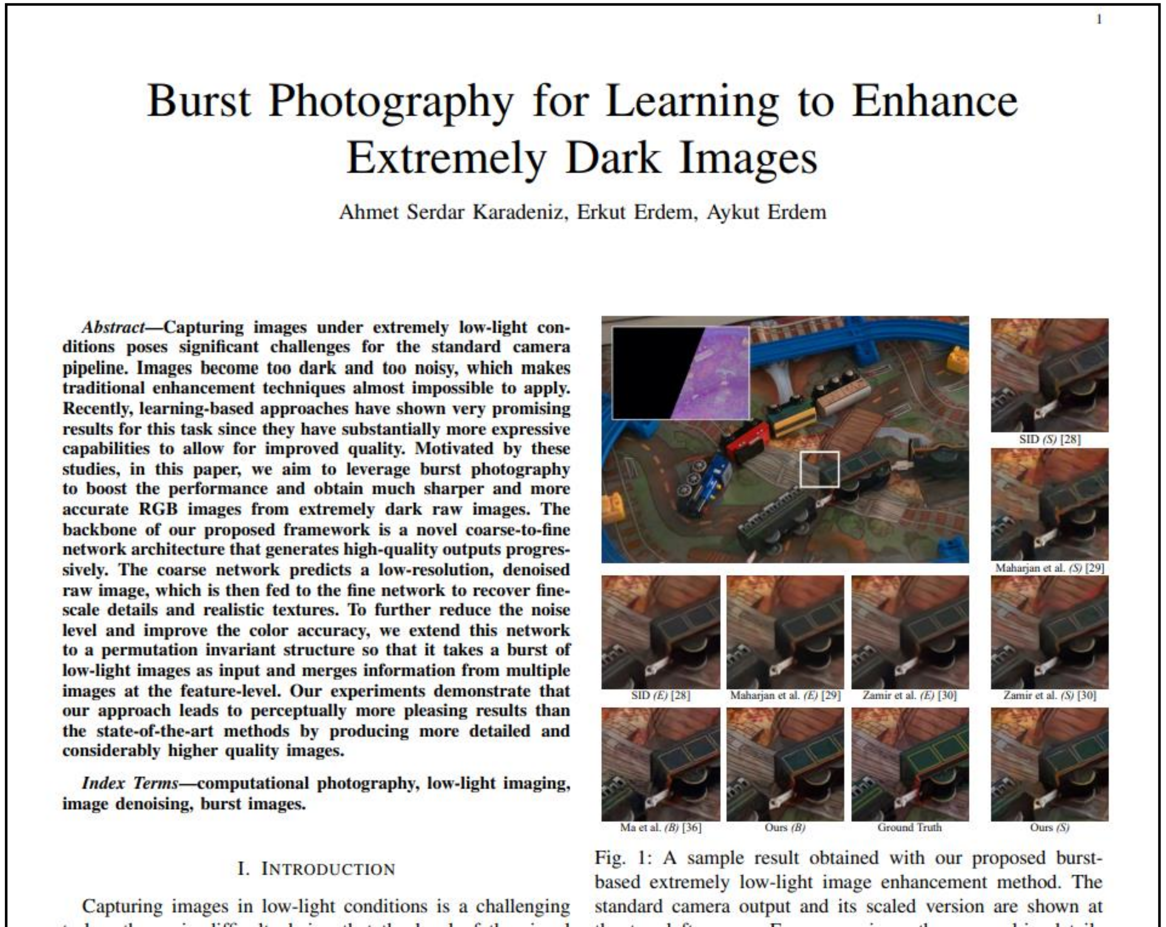
(a) *Image pair* **(b)** *Intermediate alignment fields*

(a) *Reference frame* **(b)** *Temporal mean with alignment* **(c)** *Robust merge*

High-dynamic-range is in terms of bit depth. This paper claims the denoising/fusion can upsample from 10bit to 12bit.

Multi-frame for low light

Senhar et al TIP 2021



1

Burst Photography for Learning to Enhance Extremely Dark Images

Ahmet Serdar Karadeniz, Erkut Erdem, Aykut Erdem

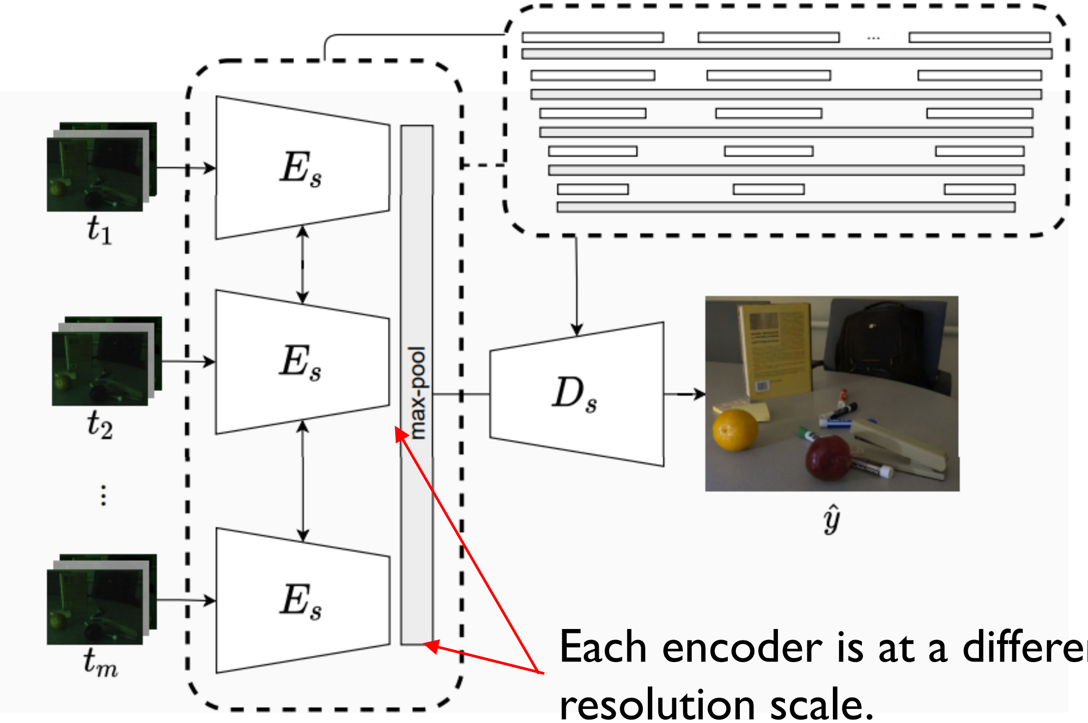
***Abstract*—Capturing images under extremely low-light conditions poses significant challenges for the standard camera pipeline. Images become too dark and too noisy, which makes traditional enhancement techniques almost impossible to apply. Recently, learning-based approaches have shown very promising results for this task since they have substantially more expressive capabilities to allow for improved quality. Motivated by these studies, in this paper, we aim to leverage burst photography to boost the performance and obtain much sharper and more accurate RGB images from extremely dark raw images. The backbone of our proposed framework is a novel coarse-to-fine network architecture that generates high-quality outputs progressively. The coarse network predicts a low-resolution, denoised raw image, which is then fed to the fine network to recover fine-scale details and realistic textures. To further reduce the noise level and improve the color accuracy, we extend this network to a permutation invariant structure so that it takes a burst of low-light images as input and merges information from multiple images at the feature-level. Our experiments demonstrate that our approach leads to perceptually more pleasing results than the state-of-the-art methods by producing more detailed and considerably higher quality images.**

***Index Terms*—computational photography, low-light imaging, image denoising, burst images.**

Fig. 1: A sample result obtained with our proposed burst-based extremely low-light image enhancement method. The standard camera output and its scaled version are shown at

I. INTRODUCTION

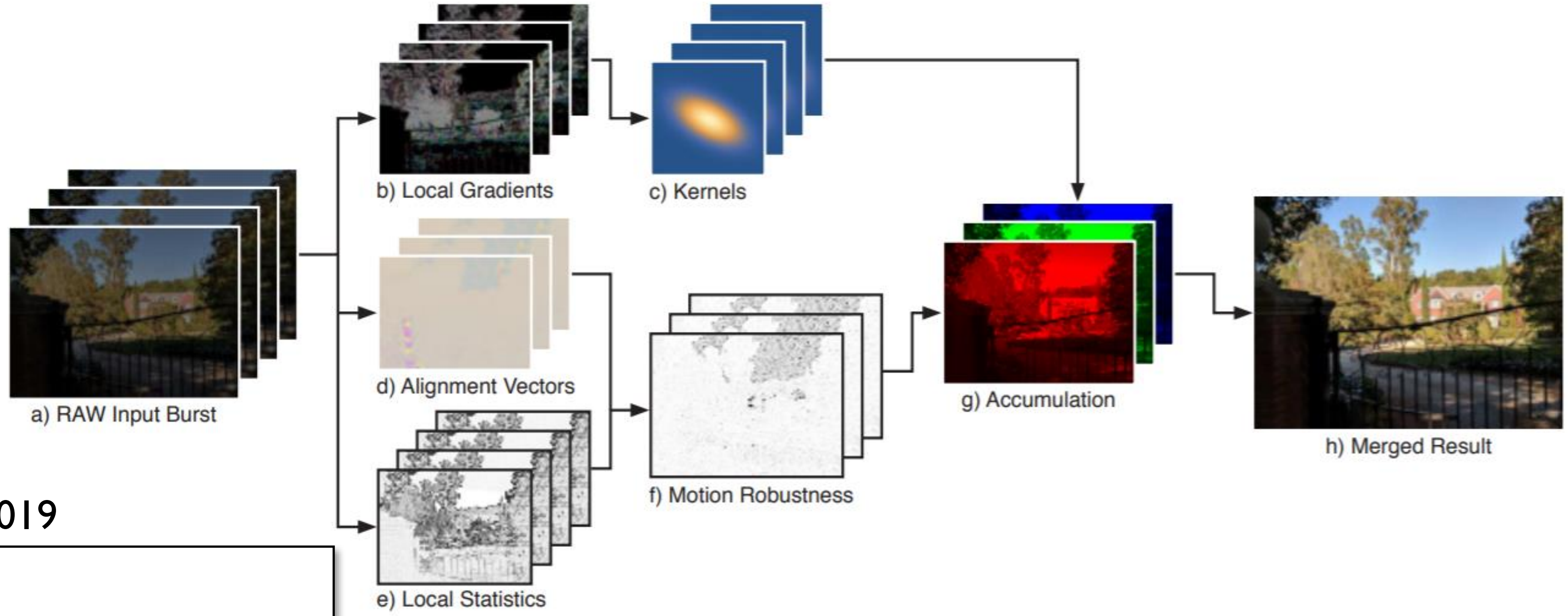
Capturing images in low-light conditions is a challenging



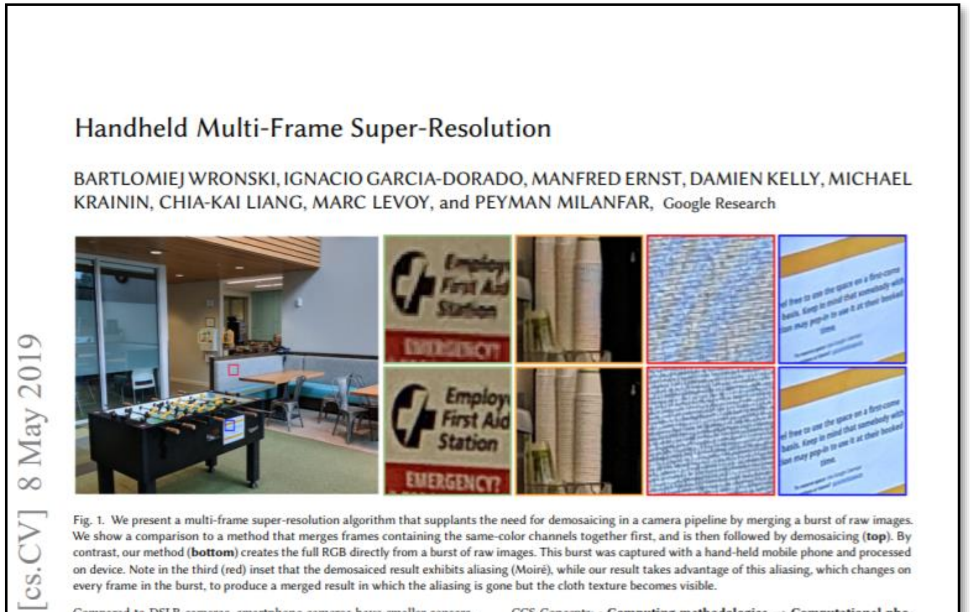
Each encoder is at a different resolution scale.

Learns a multi-scale burst encoder/decoder framework. Input is RAW, output is sRGB.

Google pixel phones multi-frame



Wronski et al SIGGRAPH 2019

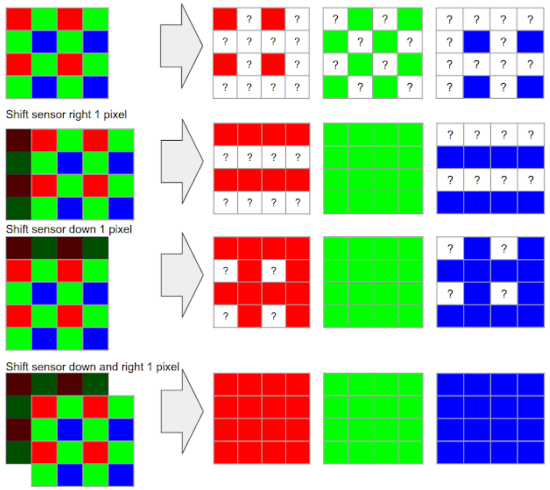


[cs.CV] 8 May 2019

Handheld Multi-Frame Super-Resolution

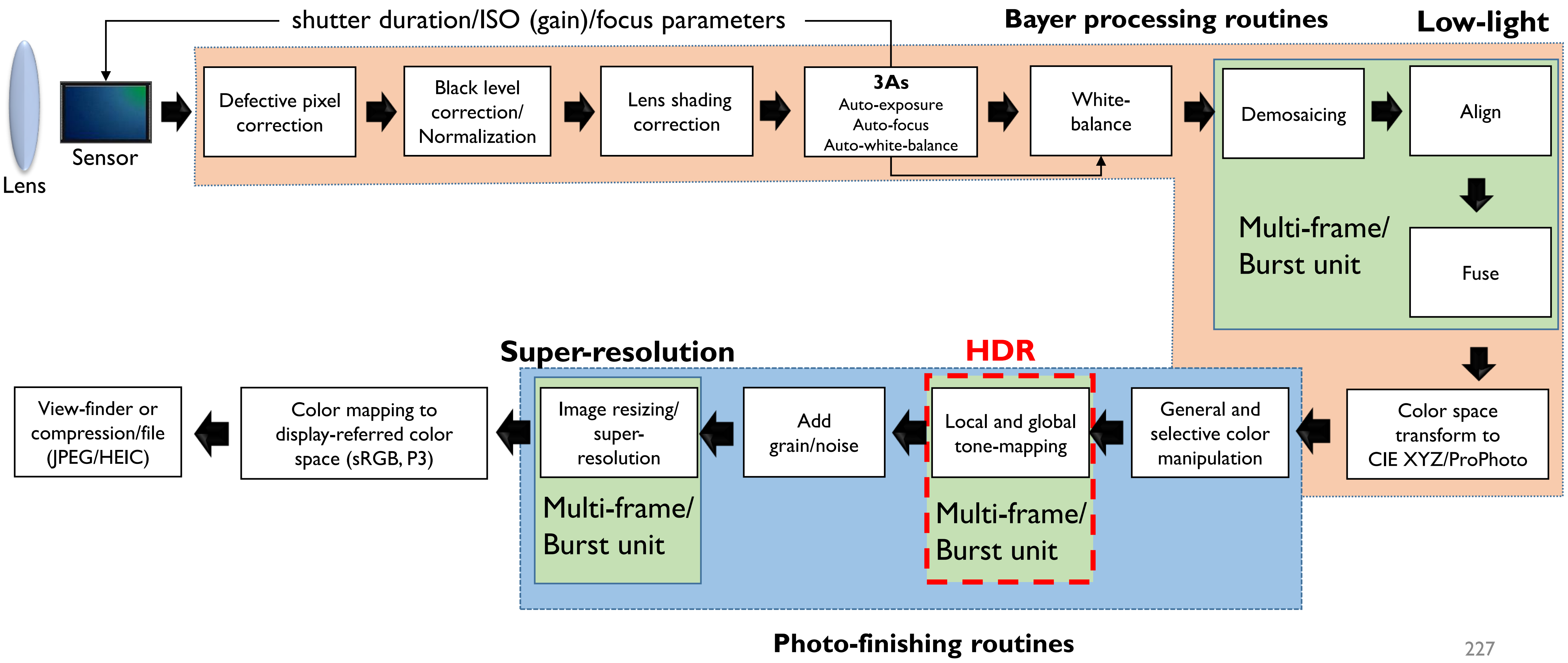
BARTLOMIEJ WRONSKI, IGNACIO GARCIA-DORADO, MANFRED ERNST, DAMIEN KELLY, MICHAEL KRAININ, CHIA-KAI LIANG, MARC LEVOY, and PEYMAN MILANFAR, Google Research

Fig. 1. We present a multi-frame super-resolution algorithm that supplants the need for demosaicing in a camera pipeline by merging a burst of raw images. We show a comparison to a method that merges frames containing the same-color channels together first, and is then followed by demosaicing (**top**). By contrast, our method (**bottom**) creates the full RGB directly from a burst of raw images. This burst was captured with a hand-held mobile phone and processed on device. Note in the third (red) inset that the demosaiced result exhibits aliasing (Moiré), while our result takes advantage of this aliasing, which changes on every frame in the burst, to produce a merged result in which the aliasing is gone but the cloth texture becomes visible.



Not necessarily for low-light, but does target RAW.
This paper uses multiple frames and very small camera motion (from hand tremors) to perform demosaicing and super-resolution. By exploiting motion, they can fill in missing Bayer data too.

ISP with multi-frame



High dynamic range imaging

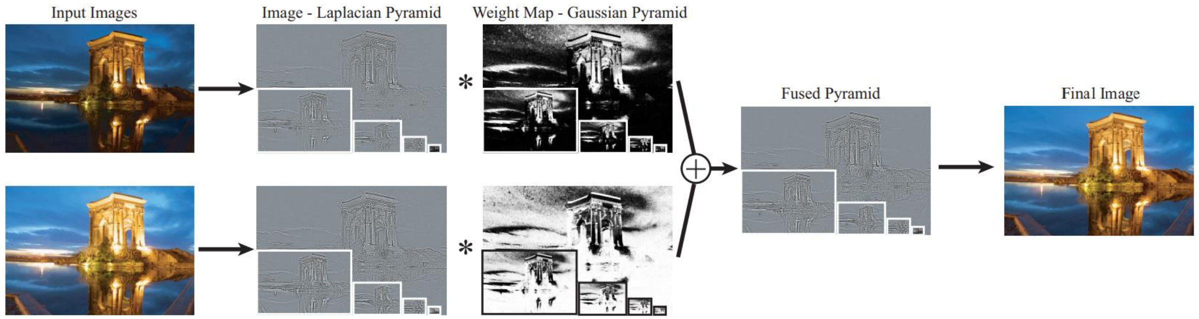
An f-stop adjusts the amount of light that falls on the sensor generally by a factor of 2. So, a +1 f-stop increases the amount of light by two times. An -1 f-stop reduces the amount of light by ½. We assume the ISO is not adjusted.

This is often called an Exposure Value adjustment. Change of EV is a change in stop.

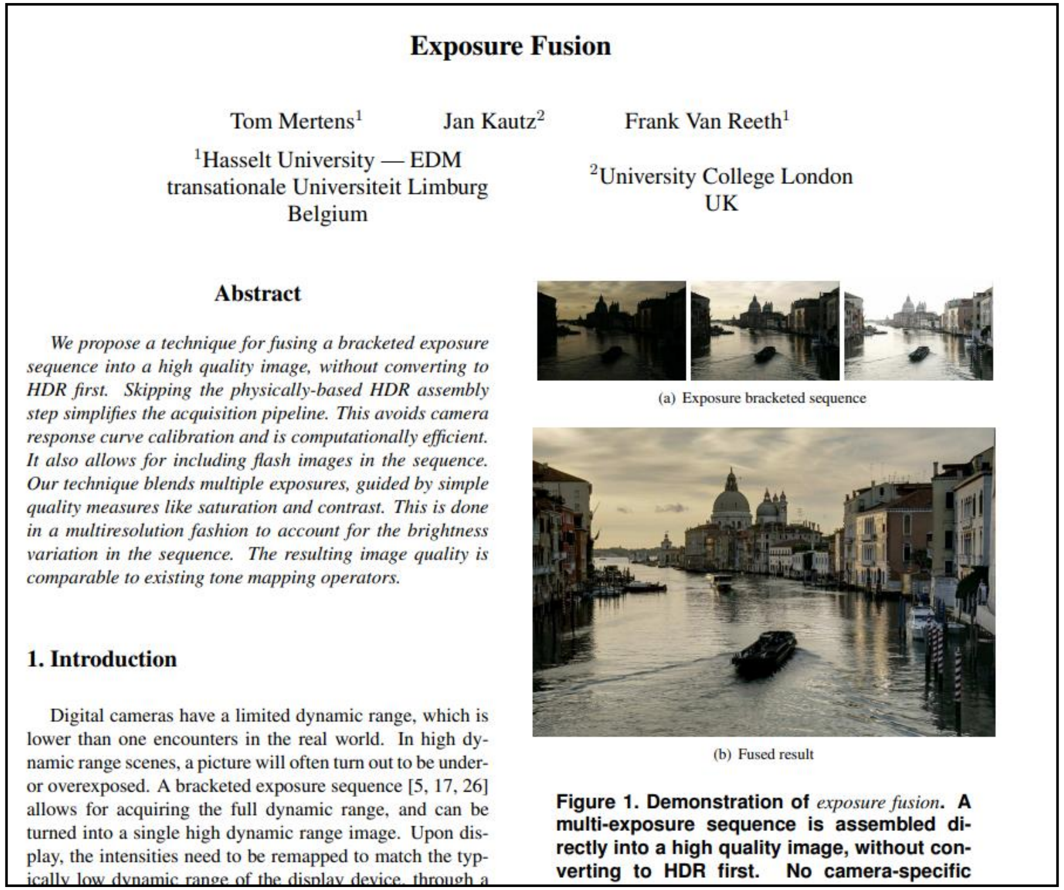
Fused result

Fused result + local tone mapping

Exposure fusion



Mertens et al – Pacific Graphics 2007



Exposure Fusion

Tom Mertens[1] Jan Kautz[2] Frank Van Reeth[1]

[1]Hasselt University — EDM
transnationale Universiteit Limburg
Belgium

[2]University College London
UK

Abstract

We propose a technique for fusing a bracketed exposure sequence into a high quality image, without converting to HDR first. Skipping the physically-based HDR assembly step simplifies the acquisition pipeline. This avoids camera response curve calibration and is computationally efficient. It also allows for including flash images in the sequence. Our technique blends multiple exposures, guided by simple quality measures like saturation and contrast. This is done in a multiresolution fashion to account for the brightness variation in the sequence. The resulting image quality is comparable to existing tone mapping operators.

(a) Exposure bracketed sequence

(b) Fused result

Figure 1. Demonstration of *exposure fusion*. A multi-exposure sequence is assembled directly into a high quality image, without converting to HDR first. No camera-specific

1. Introduction

Digital cameras have a limited dynamic range, which is lower than one encounters in the real world. In high dynamic range scenes, a picture will often turn out to be under- or overexposed. A bracketed exposure sequence [5, 17, 26] allows for acquiring the full dynamic range, and can be turned into a single high dynamic range image. Upon display, the intensities need to be remapped to match the typically low dynamic range of the display device, through a

Simple method that fused multiple exposed (and rendered) images to a single 'fused' output.

Works on Laplacian pyramid.

Proposed heuristics for determining weights for fusion.
 - Namely: saturation, contrast, "exposedness" at each level

Exposure fusion

(a) Input sequence

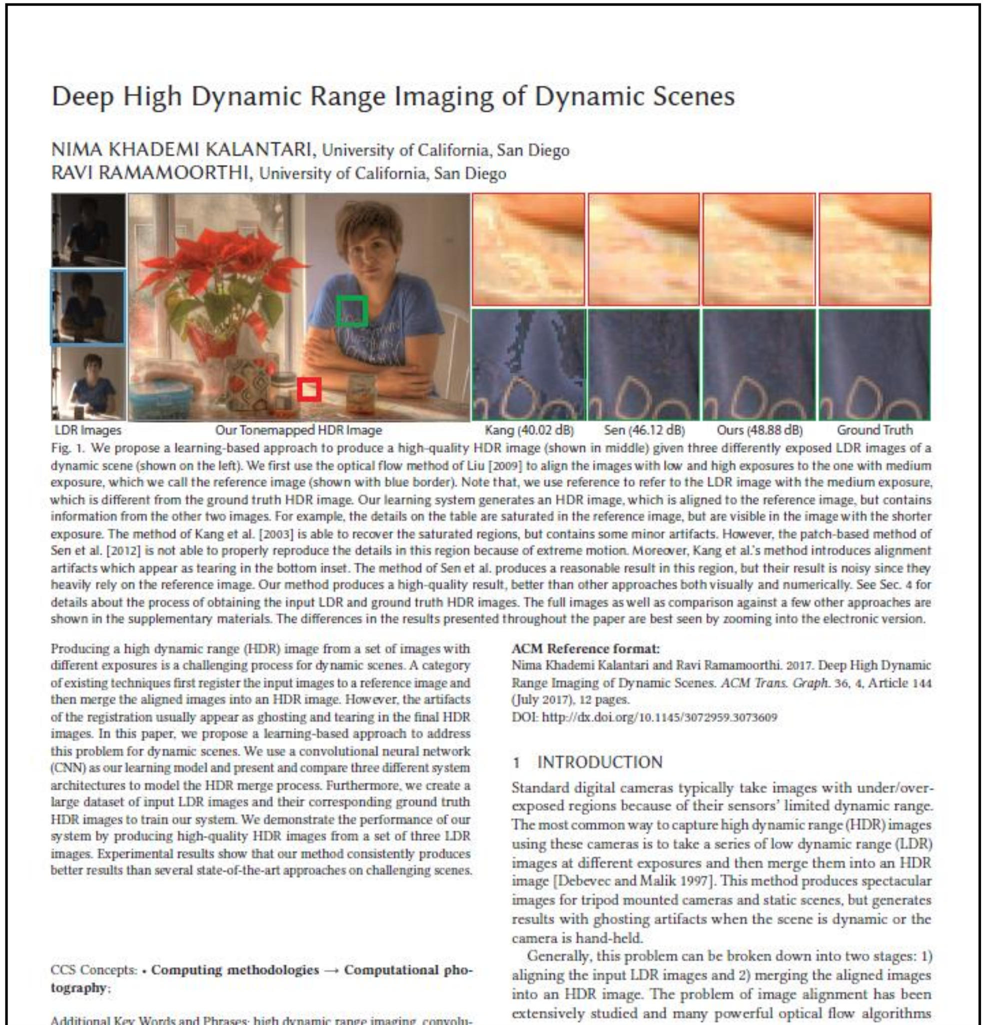
(b) Ogden et al. [19] (c) Burt et al. [4] (d) Our technique

Exposure fusion gave spectacular results compared to existing methods in 2007.

Simple algorithm
makes it suitable
for real-time deployment
on device.

DNN-based multi-frame HDR

Kalantari and Ramanoorthi SIGGRAPH'17



Deep High Dynamic Range Imaging of Dynamic Scenes

NIMA KHADEMI KALANTARI, University of California, San Diego
RAVI RAMAMOORTHI, University of California, San Diego

LDR Images | Our Tonemapped HDR Image | Kang (40.02 dB) | Sen (46.12 dB) | Ours (48.88 dB) | Ground Truth

Fig. 1. We propose a learning-based approach to produce a high-quality HDR image (shown in middle) given three differently exposed LDR images of a dynamic scene (shown on the left). We first use the optical flow method of Liu [2009] to align the images with low and high exposures to the one with medium exposure, which we call the reference image (shown with blue border). Note that, we use reference to refer to the LDR image with the medium exposure, which is different from the ground truth HDR image. Our learning system generates an HDR image, which is aligned to the reference image, but contains information from the other two images. For example, the details on the table are saturated in the reference image, but are visible in the image with the shorter exposure. The method of Kang et al. [2003] is able to recover the saturated regions, but contains some minor artifacts. However, the patch-based method of Sen et al. [2012] is not able to properly reproduce the details in this region because of extreme motion. Moreover, Kang et al.'s method introduces alignment artifacts which appear as tearing in the bottom inset. The method of Sen et al. produces a reasonable result in this region, but their result is noisy since they heavily rely on the reference image. Our method produces a high-quality result, better than other approaches both visually and numerically. See Sec. 4 for details about the process of obtaining the input LDR and ground truth HDR images. The full images as well as comparison against a few other approaches are shown in the supplementary materials. The differences in the results presented throughout the paper are best seen by zooming into the electronic version.

Producing a high dynamic range (HDR) image from a set of images with different exposures is a challenging process for dynamic scenes. A category of existing techniques first register the input images to a reference image and then merge the aligned images into an HDR image. However, the artifacts of the registration usually appear as ghosting and tearing in the final HDR images. In this paper, we propose a learning-based approach to address this problem for dynamic scenes. We use a convolutional neural network (CNN) as our learning model and present and compare three different system architectures to model the HDR merge process. Furthermore, we create a large dataset of input LDR images and their corresponding ground truth HDR images to train our system. We demonstrate the performance of our system by producing high-quality HDR images from a set of three LDR images. Experimental results show that our method consistently produces better results than several state-of-the-art approaches on challenging scenes.

CCS Concepts: • **Computing methodologies → Computational photography**;

Additional Key Words and Phrases: high dynamic range imaging, convolu-

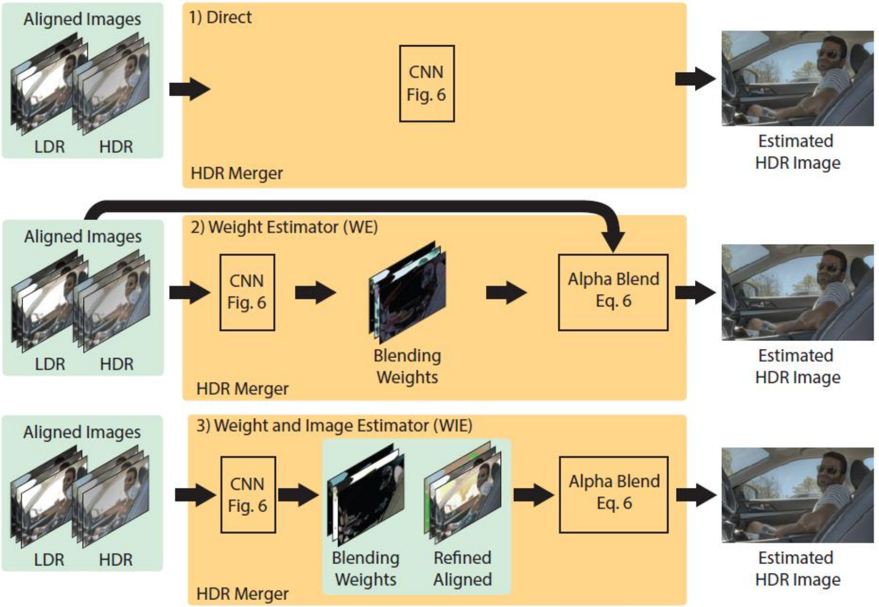
ACM Reference format:
Nima Khademi Kalantari and Ravi Ramamoorthi. 2017. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Trans. Graph.* 36, 4, Article 144 (July 2017), 12 pages.
DOI: http://dx.doi.org/10.1145/3072959.3073609

1 INTRODUCTION

Standard digital cameras typically take images with under/over-exposed regions because of their sensors' limited dynamic range. The most common way to capture high dynamic range (HDR) images using these cameras is to take a series of low dynamic range (LDR) images at different exposures and then merge them into an HDR image [Debevec and Malik 1997]. This method produces spectacular images for tripod mounted cameras and static scenes, but generates results with ghosting artifacts when the scene is dynamic or the camera is hand-held.

Generally, this problem can be broken down into two stages: 1) aligning the input LDR images and 2) merging the aligned images into an HDR image. The problem of image alignment has been extensively studied and many powerful optical flow algorithms

Paper examined three strategies.
(1) Multi-frame and CNN to predict final HDR.
(2) Multi-frame and CNN to predict blending weights, then HDR.
(3) Multi-frame and CNN to predict blending weights and align misaligned regions
Found that (#2) is the best; (3) works for small motions.

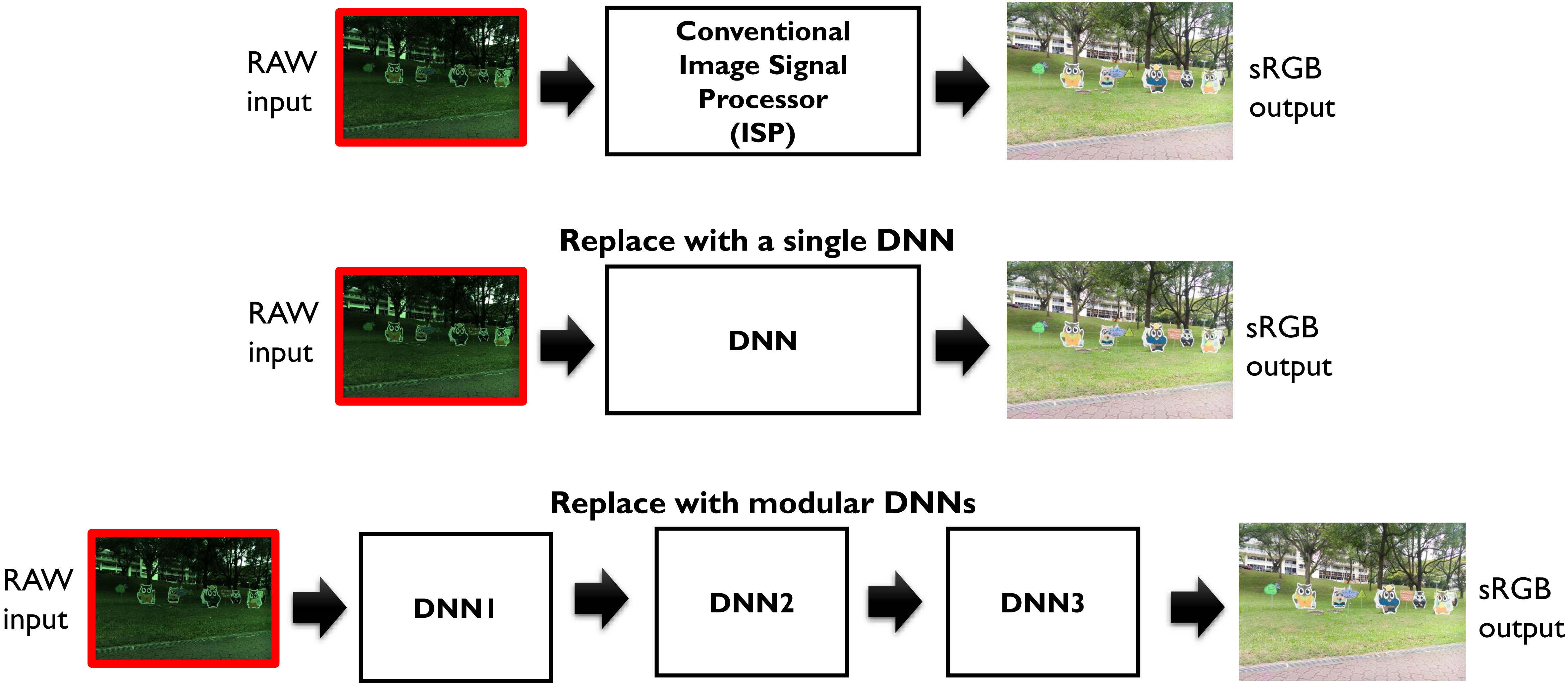


Summary

- Deep learning is good at addressing hard ISP components
  - Demosiacing, denoising, AWB, super-resolution (digital zoom)
  - These are components that are ill-posed problems (many-to-one solutions)
- Hand-crafted solutions still work well
- GANs shows promise for visual results (not necessarily benchmarks)
- Many current SOTA solutions are based on neural architecture search (NAS)

Part 2:
AI-based ISPs

Replacing the conventional ISP



Single DNN replacement

Modeling camera rendering with a DNN

Nam and Kim CVPR'17

Modelling the Scene Dependent Imaging in Cameras with a Deep Neural Network

Seonghyeon Nam
Yonsei University
shnnam@yonsei.ac.kr

Seon Joo Kim
Yonsei University
seonjookim@yonsei.ac.kr

Abstract

We present a novel deep learning framework that models the scene dependent image processing inside cameras. Often called as the radiometric calibration, the process of recovering RAW images from processed images (JPEG format in the sRGB color space) is essential for many computer vision tasks that rely on physically accurate radiance values. All previous works rely on the deterministic imaging model where the color transformation stays the same regardless of the scene and thus they can only be applied for images taken under the manual mode. In this paper, we propose a data-driven approach to learn the scene dependent and locally varying image processing inside cameras under the automode. Our method incorporates both the global and the local scene context into pixel-wise features via multi-scale pyramid of learnable histogram layers. The results show that we can model the imaging pipeline of different cameras that operate under the automode accurately in both direc-

(a) Manual mode (b) Auto-mode

Figure 1. Difference of two images captured (a) under the manual mode and (b) under the auto-mode. The RAW images of both (a) and (b) are identical. In (b), the brightness/contrast and the colors were enhanced automatically by the camera.

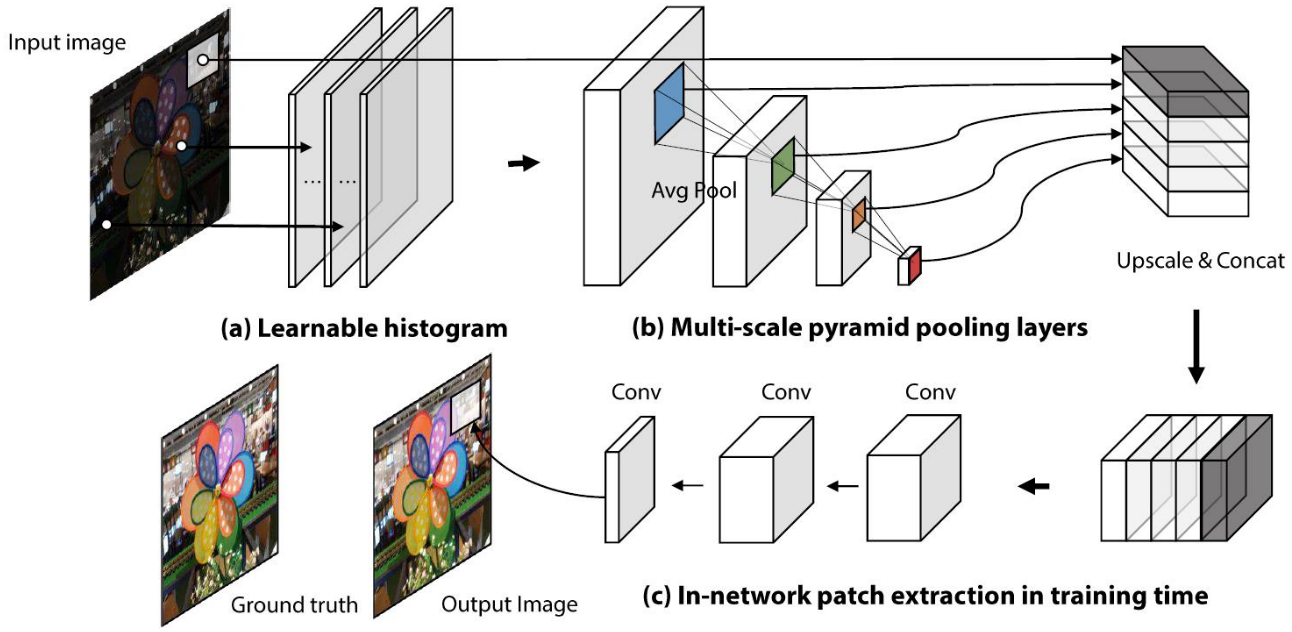
- The paper is motivated by "reversing" the ISP from sRGB to RAW

- Addresses the scene-dependent nature of ISPs for radiometric calibration

- However, the framework can be used for "forward rendering" (RAW to sRGB)

- This work is often overlooked due to the focus on radiometric calibration

Modeling camera with DNN



This paper works on image patches and is trained per camera.

To encode local context information, a learnable histogram feature is used and polled at different scales.

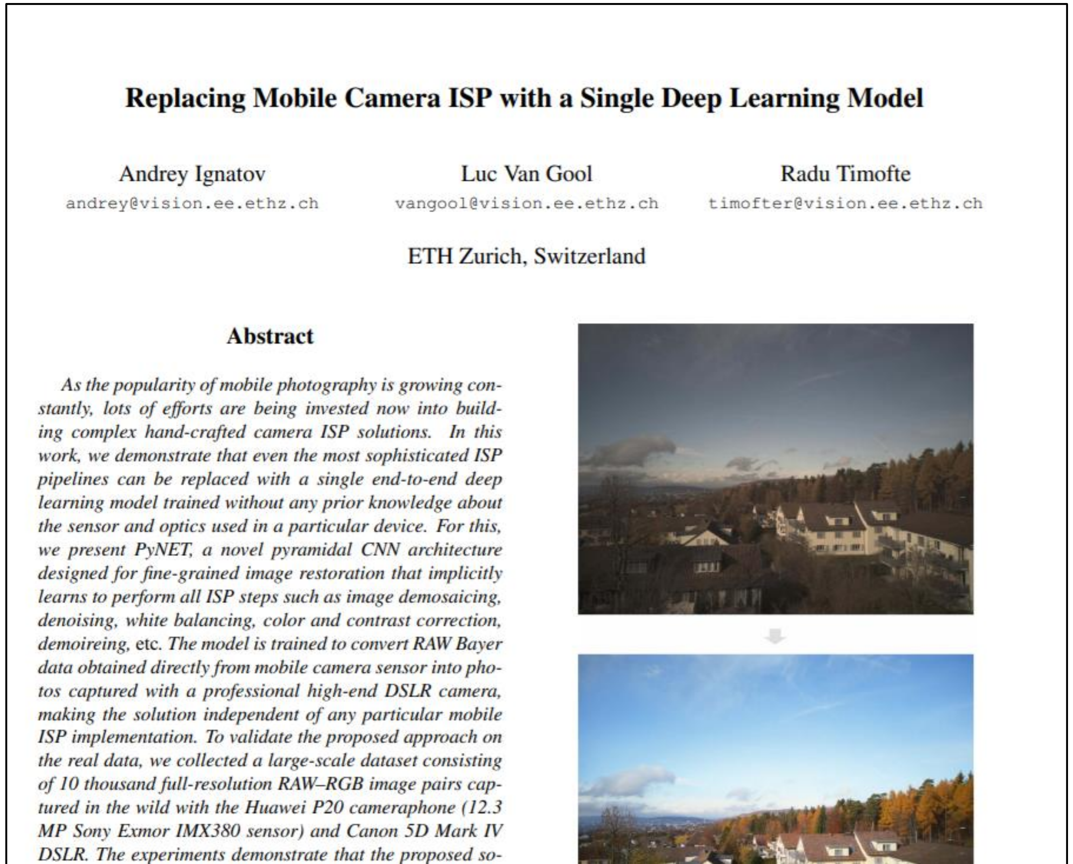
The local histogram feature provides spatial context for converting from RAW to sRGB (or sRGB to RAW).

Modeling camera with DNN

Results of rendering RAW to sRGB.

ISP replacement to mimic better camera

CVPRW'19 (NTIRE workshop)



Replacing Mobile Camera ISP with a Single Deep Learning Model

Andrey Ignatov
andrey@vision.ee.ethz.ch

Luc Van Gool
vangool@vision.ee.ethz.ch

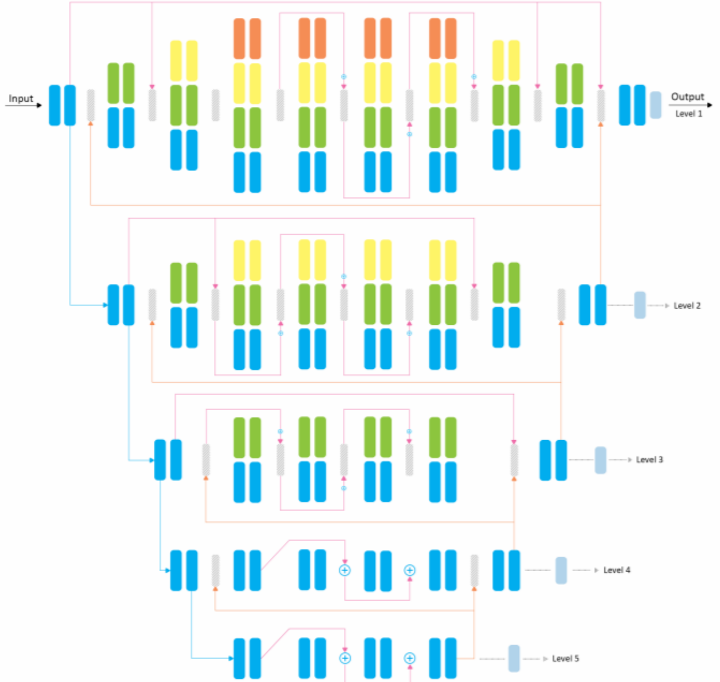
Radu Timofte
timofter@vision.ee.ethz.ch

ETH Zurich, Switzerland

Abstract

As the popularity of mobile photography is growing constantly, lots of efforts are being invested now into building complex hand-crafted camera ISP solutions. In this work, we demonstrate that even the most sophisticated ISP pipelines can be replaced with a single end-to-end deep learning model trained without any prior knowledge about the sensor and optics used in a particular device. For this, we present PyNET, a novel pyramidal CNN architecture designed for fine-grained image restoration that implicitly learns to perform all ISP steps such as image demosaicing, denoising, white balancing, color and contrast correction, demoireing, etc. The model is trained to convert RAW Bayer data obtained directly from mobile camera sensor into photos captured with a professional high-end DSLR camera, making the solution independent of any particular mobile ISP implementation. To validate the proposed approach on the real data, we collected a large-scale dataset consisting of 10 thousand full-resolution RAW–RGB image pairs captured in the wild with the Huawei P20 cameraphone (12.3 MP Sony Exmor IMX380 sensor) and Canon 5D Mark IV DSLR. The experiments demonstrate that the proposed so-

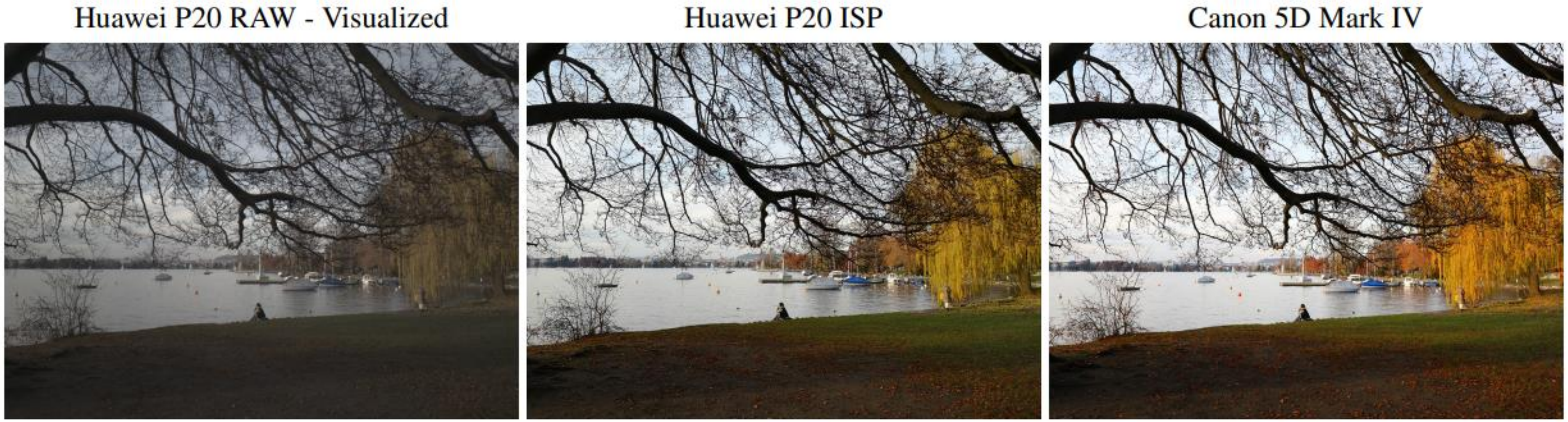
Raw Input (smartphone)



Uses U-net structure (called "PyNet" for pyramid Net)

Processed output (DSLR)

ISP replacement to mimic better camera



Training images
RAW from smartphone
sRGB from DSLR
Images are misaligned!

Images are globally aligned, and then patch wise aligned.

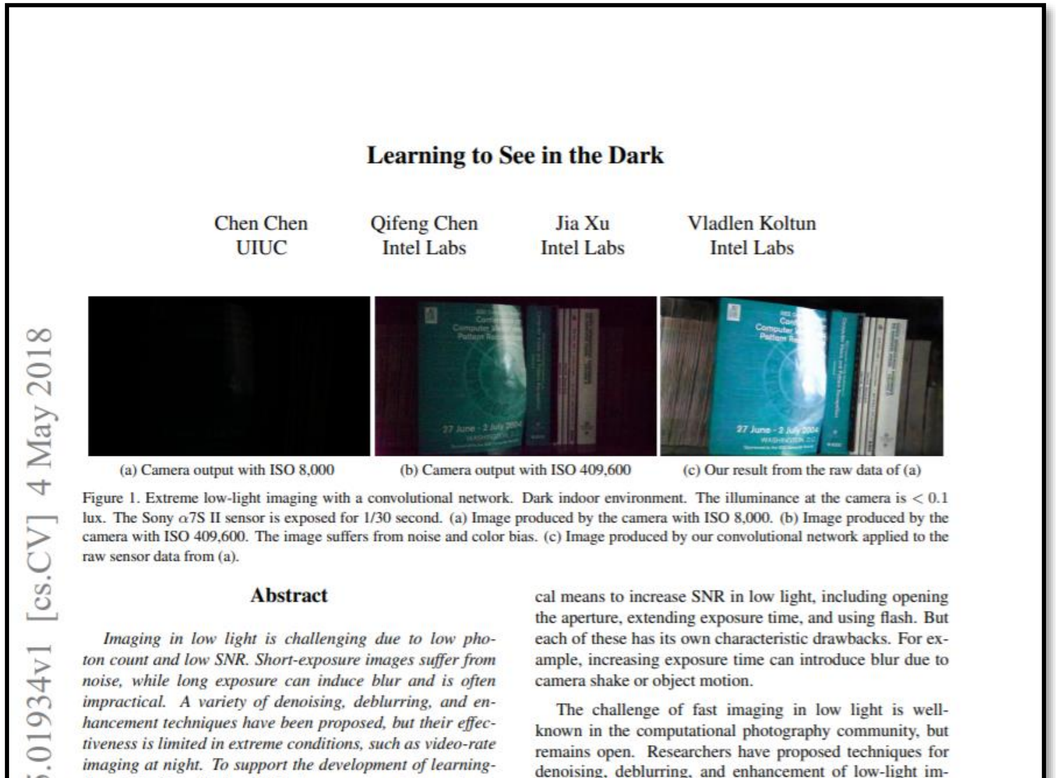
Additional perceptual loss (VGG) is included in training at different U-net scales.

ISP replacement to mimic better camera

"Learning to see in the dark"

This paper is essentially a learned ISP.
However, it learns to process noisy RAW to clean sRGB.

Chen et al CVPR 2018



Learning to See in the Dark

Chen Chen, UIUC
Qifeng Chen, Intel Labs
Jia Xu, Intel Labs
Vladlen Koltun, Intel Labs

(a) Camera output with ISO 8,000 (b) Camera output with ISO 409,600 (c) Our result from the raw data of (a)

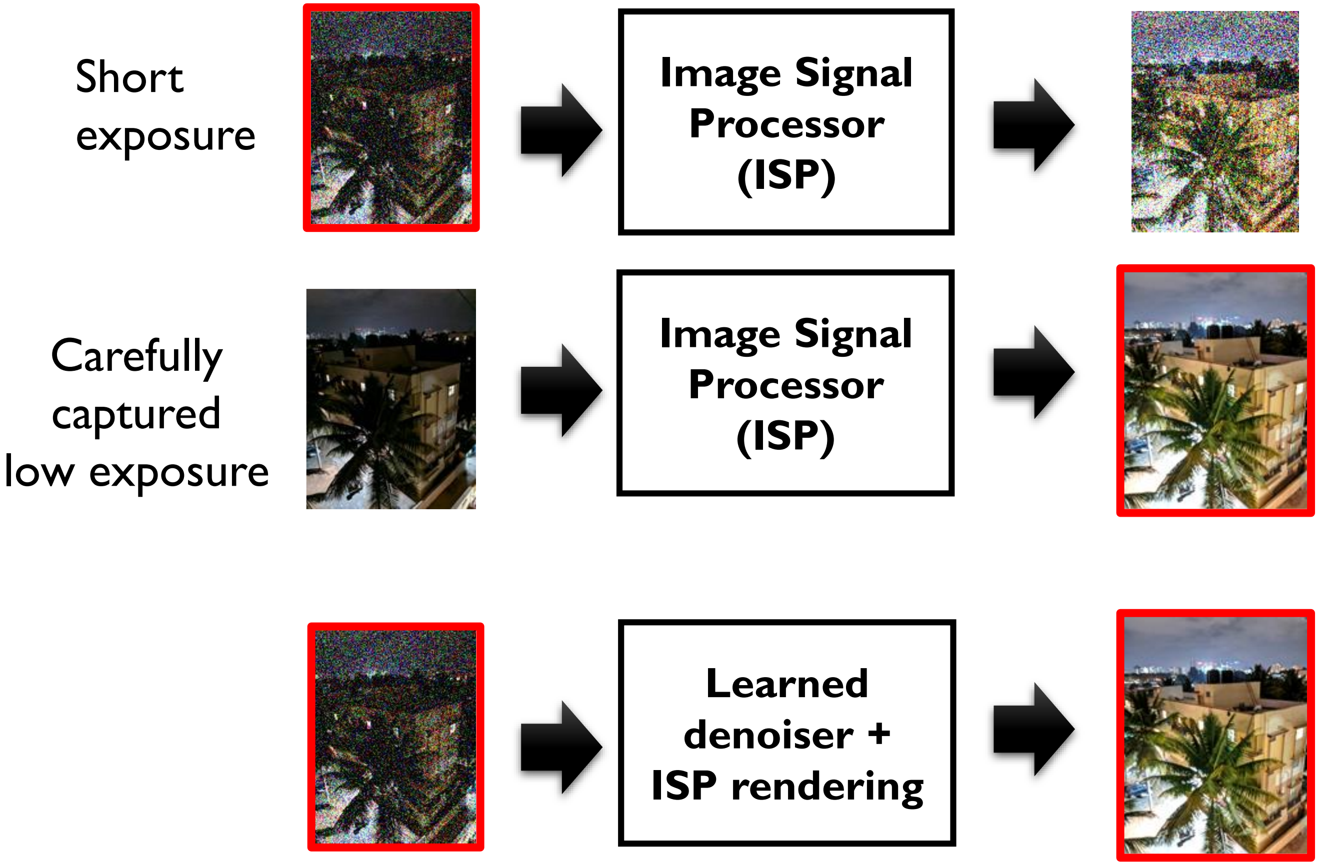
Figure 1. Extreme low-light imaging with a convolutional network. Dark indoor environment. The illuminance at the camera is < 0.1 lux. The Sony α7S II sensor is exposed for 1/30 second. (a) Image produced by the camera with ISO 8,000. (b) Image produced by the camera with ISO 409,600. The image suffers from noise and color bias. (c) Image produced by our convolutional network applied to the raw sensor data from (a).

Abstract

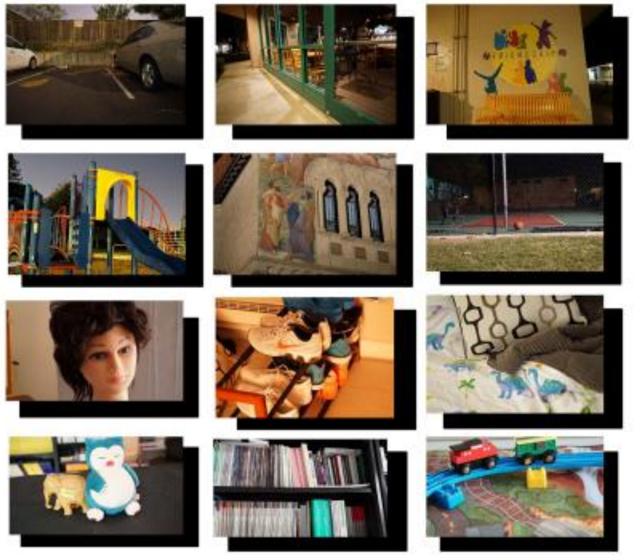
Imaging in low light is challenging due to low photon count and low SNR. Short-exposure images suffer from noise, while long exposure can induce blur and is often impractical. A variety of denoising, deblurring, and enhancement techniques have been proposed, but their effectiveness is limited in extreme conditions, such as video-rate imaging at night. To support the development of learning-

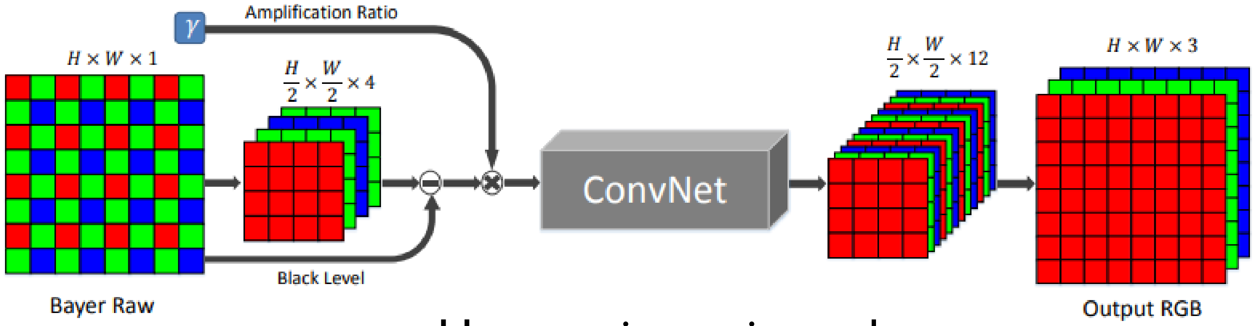
cal means to increase SNR in low light, including opening the aperture, extending exposure time, and using flash. But each of these has its own characteristic drawbacks. For example, increasing exposure time can introduce blur due to camera shake or object motion.

The challenge of fast imaging in low light is well-known in the computational photography community, but remains open. Researchers have proposed techniques for denoising, deblurring, and enhancement of low-light im-



"Learning to see in the dark"

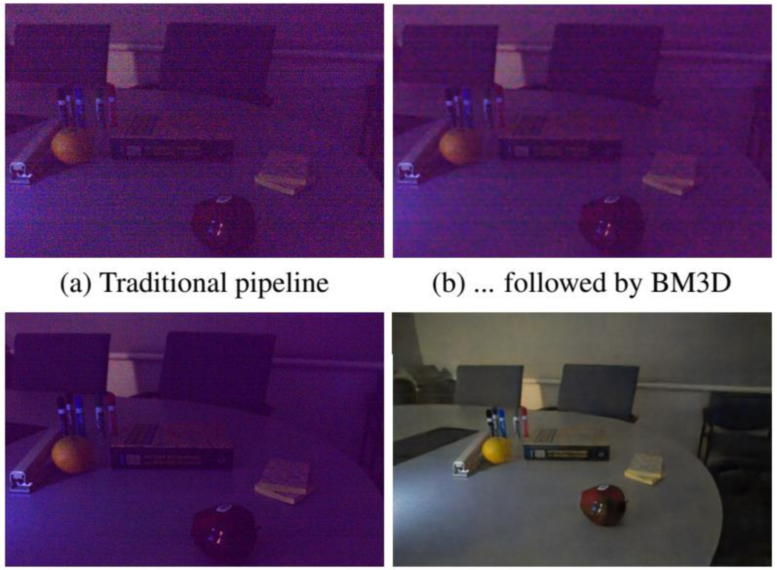




U-net articture is used.

Key to this paper is the careful alignment of data.

Results show for very low-light cases so significant performance.



(a) Traditional pipeline (b) ... followed by BM3D

(c) Burst denoising (d) Our result

CRISPnet (color reproduction ISP)

Souza and Heidrich, Arxv 2021

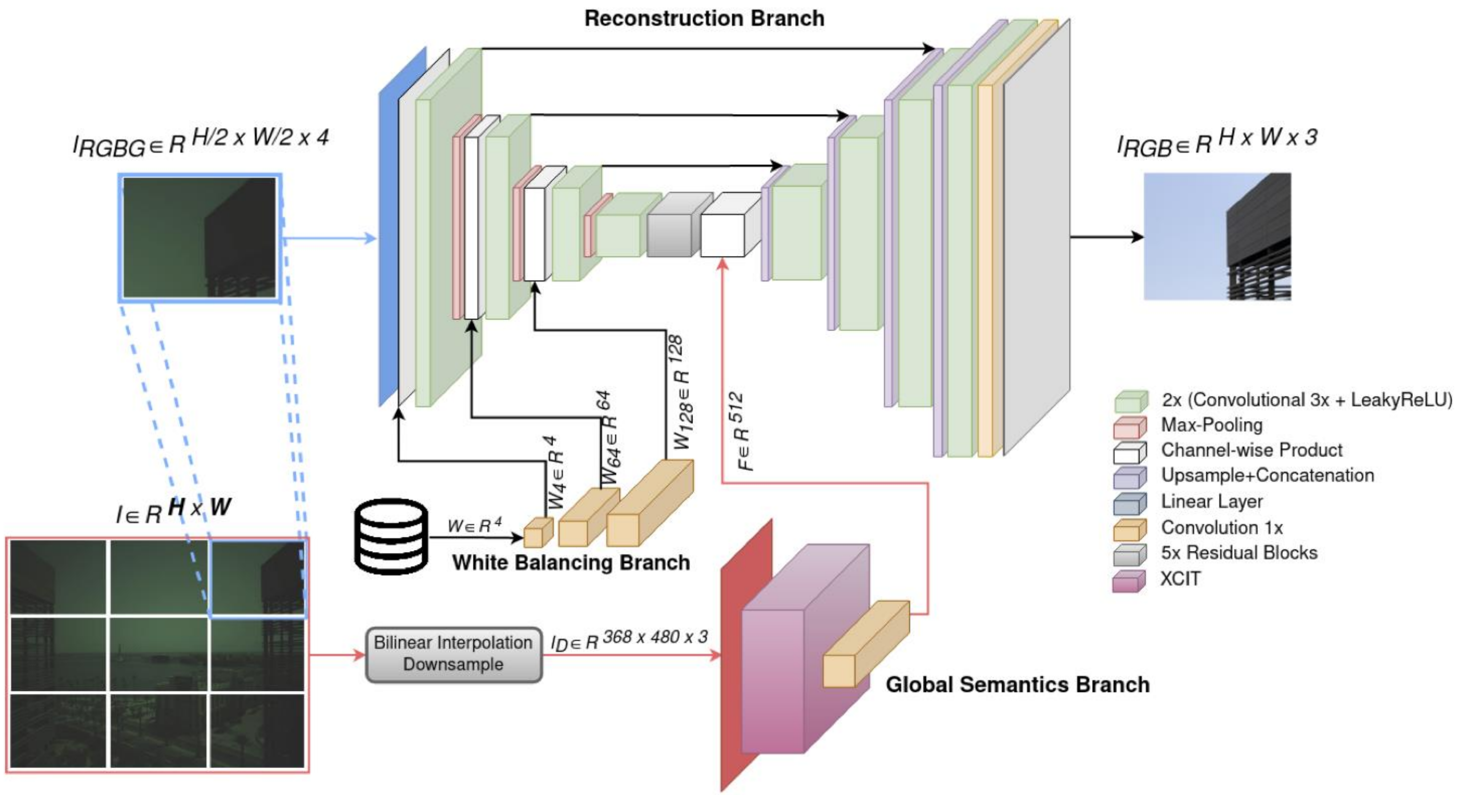
CRISPnet: Color Rendition ISP Net

Matheus Souza and Wolfgang Heidrich

King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
{matheus.medeirosdesouza,wolfgang.heidrich}@kaust.edu.sa

Abstract. Image signal processors (ISPs) are historically grown legacy software systems for reconstructing color images from noisy raw sensor measurements. They are usually composited of many heuristic blocks for denoising, demosaicking, and color restoration. Color reproduction in this context is of particular importance, since the raw colors are often severely distorted, and each smart phone manufacturer has developed their own characteristic heuristics for improving the color rendition, for example of skin tones and other visually important colors.
In recent years there has been strong interest in replacing the historically grown ISP systems with deep learned pipelines. Much progress has been made in approximating legacy ISPs with such learned models. However, so far the focus of these efforts has been on reproducing the structural features of the images, with less attention paid to color rendition.
Here we present CRISPnet, the first learned ISP model to specifically target color rendition accuracy relative to a complex, legacy smart phone ISP. We achieve this by utilizing both image metadata (like a legacy ISP would), as well as by learning simple global semantics based on image classification – similar to what a legacy ISP does to determine the scene type. We also contribute a new ISP image dataset consisting of both high dynamic range monitor data, as well as real-world data, both captured with an actual cell phone ISP pipeline under a variety of lighting conditions, exposure times, and gain settings.

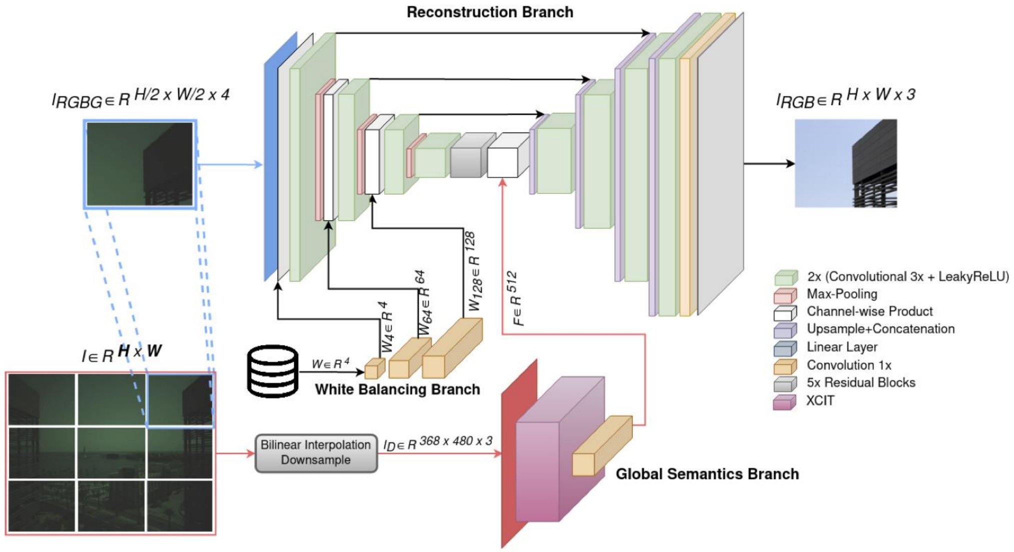
Keywords: image signal processor; image restoration; color rendition.



Monolithic ISP with focus on high-quality color rendering

CRISPNet

Motivation is that one day we will have RAW images from an smartphone, but no hardware to render it, so a DNN will be used instead. The paper refers to this as legacy ISPs.



Some interesting ideas:
(1) WB (from RAW) is injected into network layers.

(2) Global semantics is also incorporated into architecture.

Training data is Apple iPhone images. This DNN essentially learns to "render" RAW like Apple.

.

Results against other DNN ISPs.

A two stage DNN-based ISP

Two stage ISP

Liang, Cao, Zhang – TIP 2021

CameraNet: A Two-Stage Framework for Effective Camera ISP Learning

Zhetong Liang, Jianrui Cai, Zisheng Cao, and Lei Zhang, *Fellow, IEEE*

***Abstract*—Traditional image signal processing (ISP) pipeline consists of a set of cascaded image processing modules onboard a camera to reconstruct a high-quality sRGB image from the sensor raw data. Recently, some methods have been proposed to learn a convolutional neural network (CNN) to improve the performance of traditional ISP. However, in these works usually a CNN is directly trained to accomplish the ISP tasks without considering much the correlation among the different components in an ISP. As a result, the quality of reconstructed images is barely satisfactory in challenging scenarios such as low-light imaging. In this paper, we firstly analyze the correlation among the different tasks in an ISP, and categorize them into two weakly correlated groups: restoration and enhancement. Then we design a two-stage network, called CameraNet, to progressively learn the two groups of ISP tasks. In each stage, a ground truth is specified to supervise the subnetwork learning, and the two subnetworks are jointly fine-tuned to produce the final output. Experiments on three benchmark datasets show that the proposed CameraNet achieves consistently compelling reconstruction quality and outperforms the recently proposed ISP learning methods.**

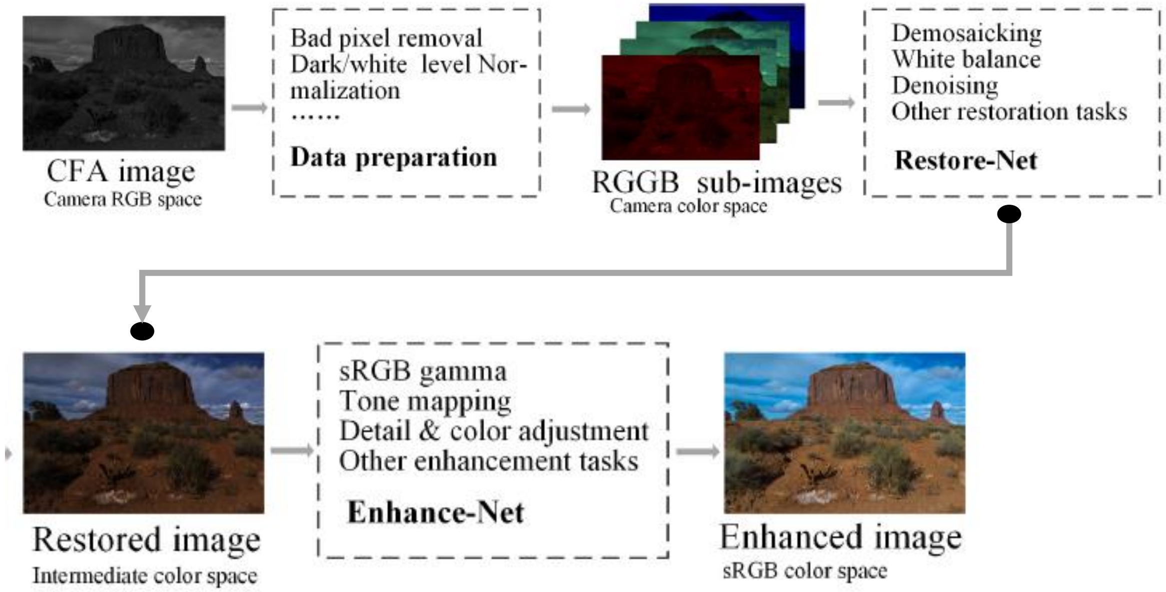
***Index Terms*—Image signal processing, image restoration, image enhancement, convolutional neural networks.**

I. Introduction

THE raw image data captured by camera sensors are typically red, green and blue channel-mosaiced irradiance signals containing noise, less vivid colors and improper tones [1], [2]. To reconstruct a displayable

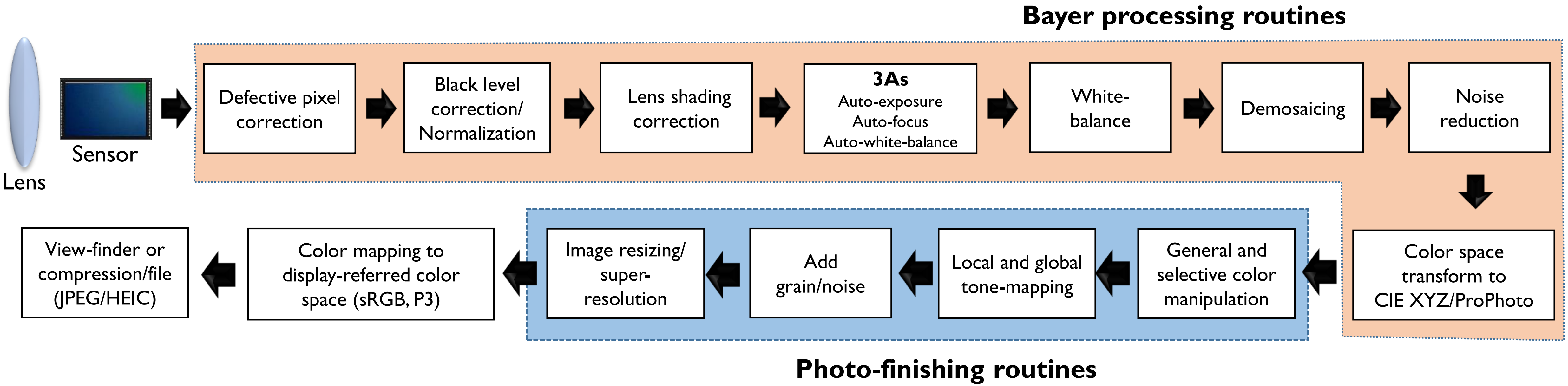
The traditional ISP is usually designed as a set of hand-crafted modules, each of which addresses a specific task [1]. For instance, a 3D lookup table is typically employed for the color enhancement task [2]. In most traditional ISP models, the modules are designed in a divide-and-conquer manner (i.e., splitting the ISP into a set of modules and developing them independently), while little attention has been paid to design them as a whole [3]. Moreover, it is time-consuming to tune each module for high image quality since the best output of one module may not result in the desired quality of the final output. Besides the standard ISP pipeline, there are also some ISP methods designed for burst imaging in the literature [4], [5]. However, these methods are subject to the effectiveness of image alignment techniques [6], which may generate ghost artifacts caused by object motion.

Recently, it has been shown that the performance of some image processing tasks, such as denoising [7], [8], white balance [9], [10], color demosaicking [11], [12], color enhancement [13]–[15], etc, can be significantly improved by deep learning techniques. In these methods, a convolutional neural network (CNN) is trained with a task-specific dataset that contains image pairs for supervised learning. Inspired by these methods, an intuitive idea is that we can train a subnetwork for each subtask of the ISP pipeline, and then chain them together as a whole ISP network. However, this

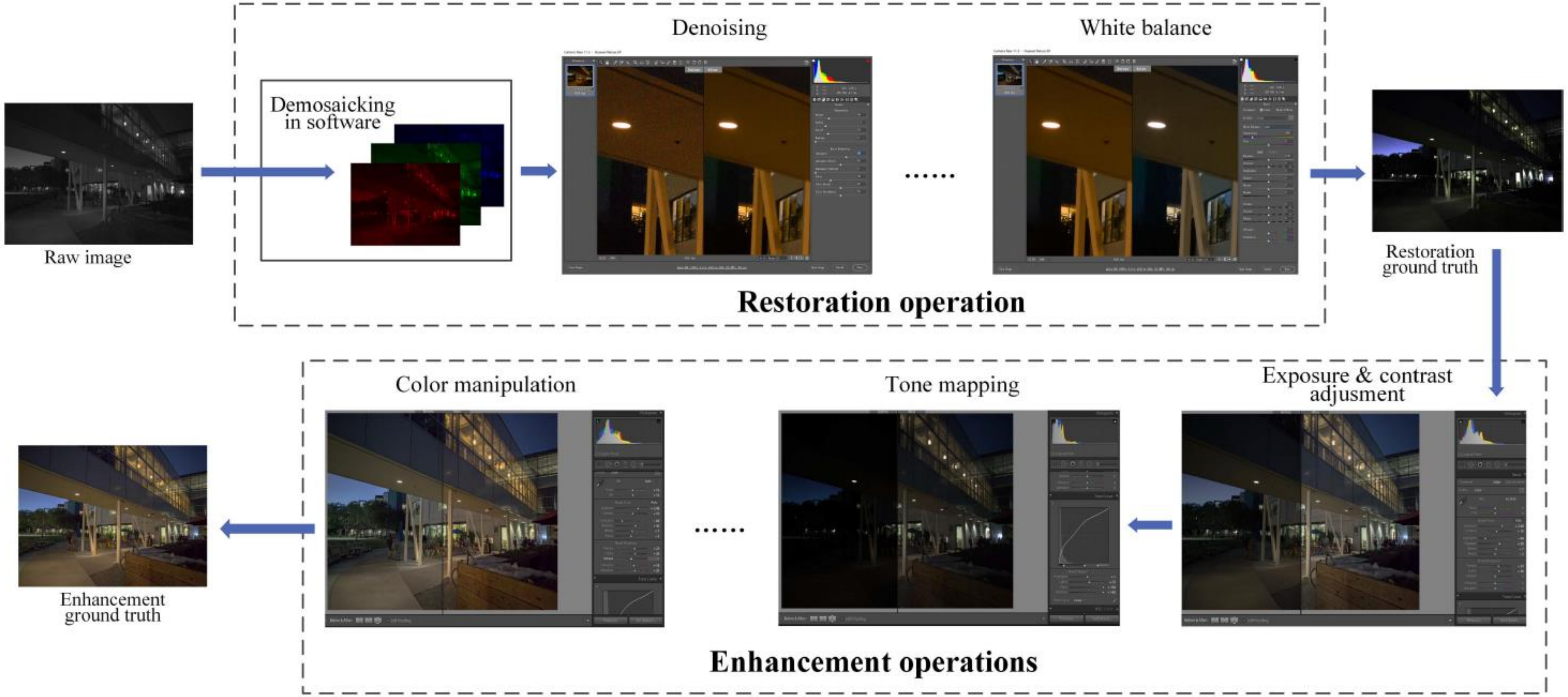


Proposes a "restore-net" and "enhance-net".

CameraNet considers real ISP stages



CameraNet – Training data preparation



Details are not very clear, but Photoshop is claimed to be used to process RAW to denoised RAW. Lightroom is used to generate enhanced images. Assumed trained per sensor type.

CameraNet

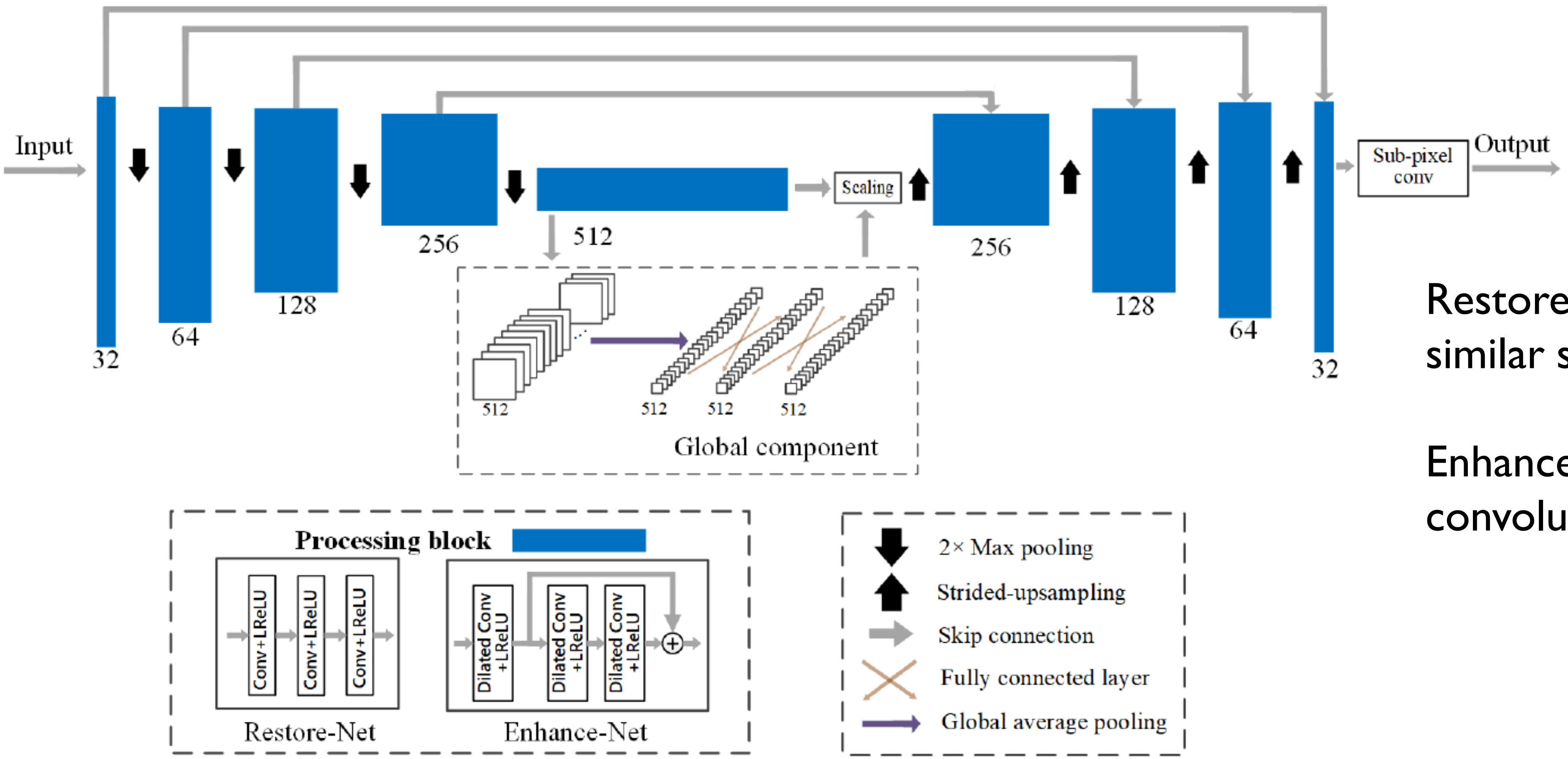
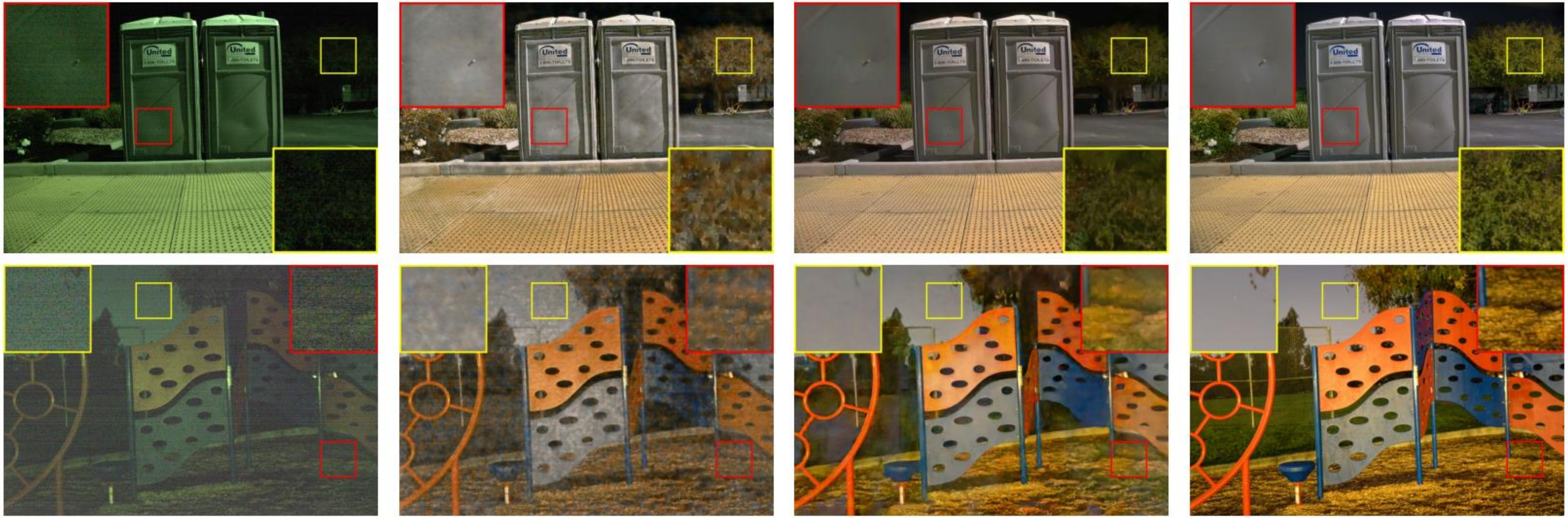


Fig. 4. The structure of UNet-like Restore-Net and Enhance-Net modules in the proposed CameraNet system.

RestoreNet/EnhanceNet share similar structure.

EnhanceNet uses a dilated convolution.

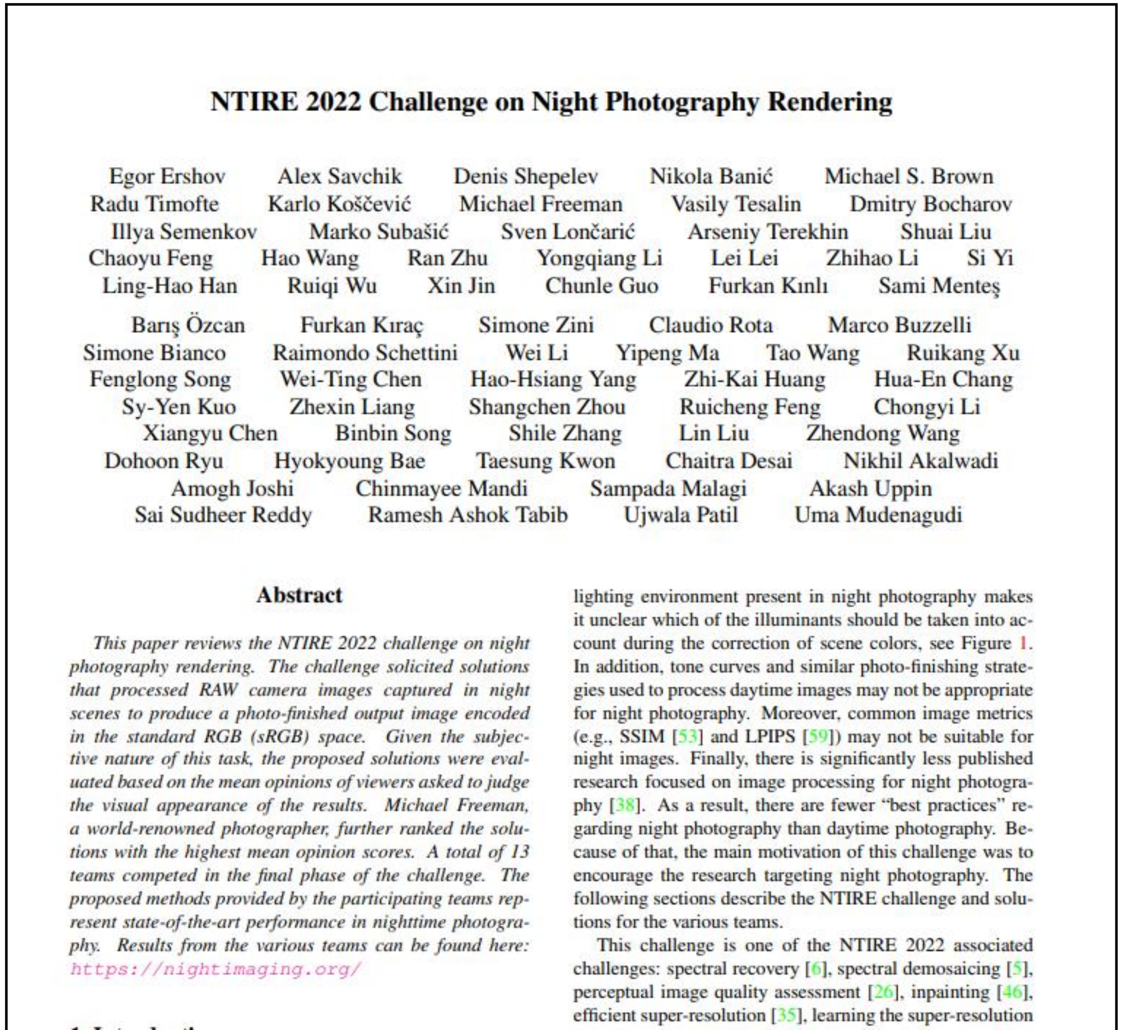
Results



(a) Raw image (b) Result by one-stage setting (c) Result by two-stage setting (d) Groundtruth

Three stage DNN-based ISP

Winner of the night photography challenge (2022)



NTIRE 2022 Challenge on Night Photography Rendering

Egor Ershov, Alex Savchik, Denis Shepelev, Nikola Banić, Michael S. Brown, Radu Timofte, Karlo Koščević, Michael Freeman, Vasily Tesalin, Dmitry Bocharov, Illya Semenkov, Marko Subašić, Sven Lončarić, Arseniy Terekhin, Shuai Liu, Chaoyu Feng, Hao Wang, Ran Zhu, Yongqiang Li, Lei Lei, Zhihao Li, Si Yi, Ling-Hao Han, Ruiqi Wu, Xin Jin, Chunle Guo, Furkan Kınlı, Sami Menteş, Barış Özcan, Furkan Kıraç, Simone Zini, Claudio Rota, Marco Buzzelli, Simone Bianco, Raimondo Schettini, Wei Li, Yipeng Ma, Tao Wang, Ruikang Xu, Fenglong Song, Wei-Ting Chen, Hao-Hsiang Yang, Zhi-Kai Huang, Hua-En Chang, Sy-Yen Kuo, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, Chongyi Li, Xiangyu Chen, Binbin Song, Shile Zhang, Lin Liu, Zhendong Wang, Dohoon Ryu, Hyokyoung Bae, Taesung Kwon, Chaitra Desai, Nikhil Akalwadi, Amogh Joshi, Chinmayee Mandi, Sampada Malagi, Akash Uppin, Sai Sudheer Reddy, Ramesh Ashok Tabib, Ujwala Patil, Uma Mudenagudi

Abstract

This paper reviews the NTIRE 2022 challenge on night photography rendering. The challenge solicited solutions that processed RAW camera images captured in night scenes to produce a photo-finished output image encoded in the standard RGB (sRGB) space. Given the subjective nature of this task, the proposed solutions were evaluated based on the mean opinions of viewers asked to judge the visual appearance of the results. Michael Freeman, a world-renowned photographer, further ranked the solutions with the highest mean opinion scores. A total of 13 teams competed in the final phase of the challenge. The proposed methods provided by the participating teams represent state-of-the-art performance in nighttime photography. Results from the various teams can be found here: https://nightimaging.org/

lighting environment present in night photography makes it unclear which of the illuminants should be taken into account during the correction of scene colors, see Figure 1. In addition, tone curves and similar photo-finishing strategies used to process daytime images may not be appropriate for night photography. Moreover, common image metrics (e.g., SSIM [53] and LPIPS [59]) may not be suitable for night images. Finally, there is significantly less published research focused on image processing for night photography [38]. As a result, there are fewer "best practices" regarding night photography than daytime photography. Because of that, the main motivation of this challenge was to encourage the research targeting night photography. The following sections describe the NTIRE challenge and solutions for the various teams.

This challenge is one of the NTIRE 2022 associated challenges: spectral recovery [6], spectral demosaicing [5], perceptual image quality assessment [26], inpainting [46], efficient super-resolution [35], learning the super-resolution

Welcome to the "Night Photography" challenge part of the NTIRE workshop at CVPR 2022.

NEWS AND UPDATES

- Teams were asked to process night RAW images to sRGB
- Toloka was used to evaluate results.
- Professional photographer Michael Freeman also evaluated.
- Winning team was from Xiaomi (net slide)

FlexISP

Liu et al NTIRE'22/CVPRW'22 (Xiaomi)

Deep-FlexISP: A Three-Stage Framework for Night Photography Rendering

Shuai Liu Chaoyu Feng Xiaotao Wang Hao Wang Ran Zhu Yongqiang Li Lei Lei
Xiaomi Inc., China
{liushuai21, fengchaoyu, wangxiaotao, wanghao35, zhuran, liyongqiang, leilei1}@xiaomi.com

Abstract

Night photography rendering is challenging due to images' high noise level, less vivid color, and low dynamic range. In this work, we propose a three-stage cascade framework named Deep-FlexISP, which decomposes the ISP into three weakly correlated sub-tasks: raw image denoising, white balance, and Bayer to sRGB mapping, for the following considerations. First, task decomposition can enhance the learning ability of the framework and make it easier to converge. Second, weak correlation sub-tasks do not influence each other too much, so the framework has a high degree of freedom. Finally, noise, color, and brightness are essential for night photographs. Our framework can flexibly adjust different styles according to personal preferences with the vital learning ability and the degree of freedom. Compared with the other Deep-ISP methods, our proposed Deep-FlexISP shows state-of-the-art performance and achieves first place in people's choice and photographer's choice in NTIRE 2022 Night Photography Render Challenge.

(a) Baseline

1. Introduction

Night photography is a challenging task due to several reasons. First, the low light condition will cause high-level noise in the raw image. Second, it is hard to estimate the

- Winner for Night Photography challenge

- Results were far better than competitors

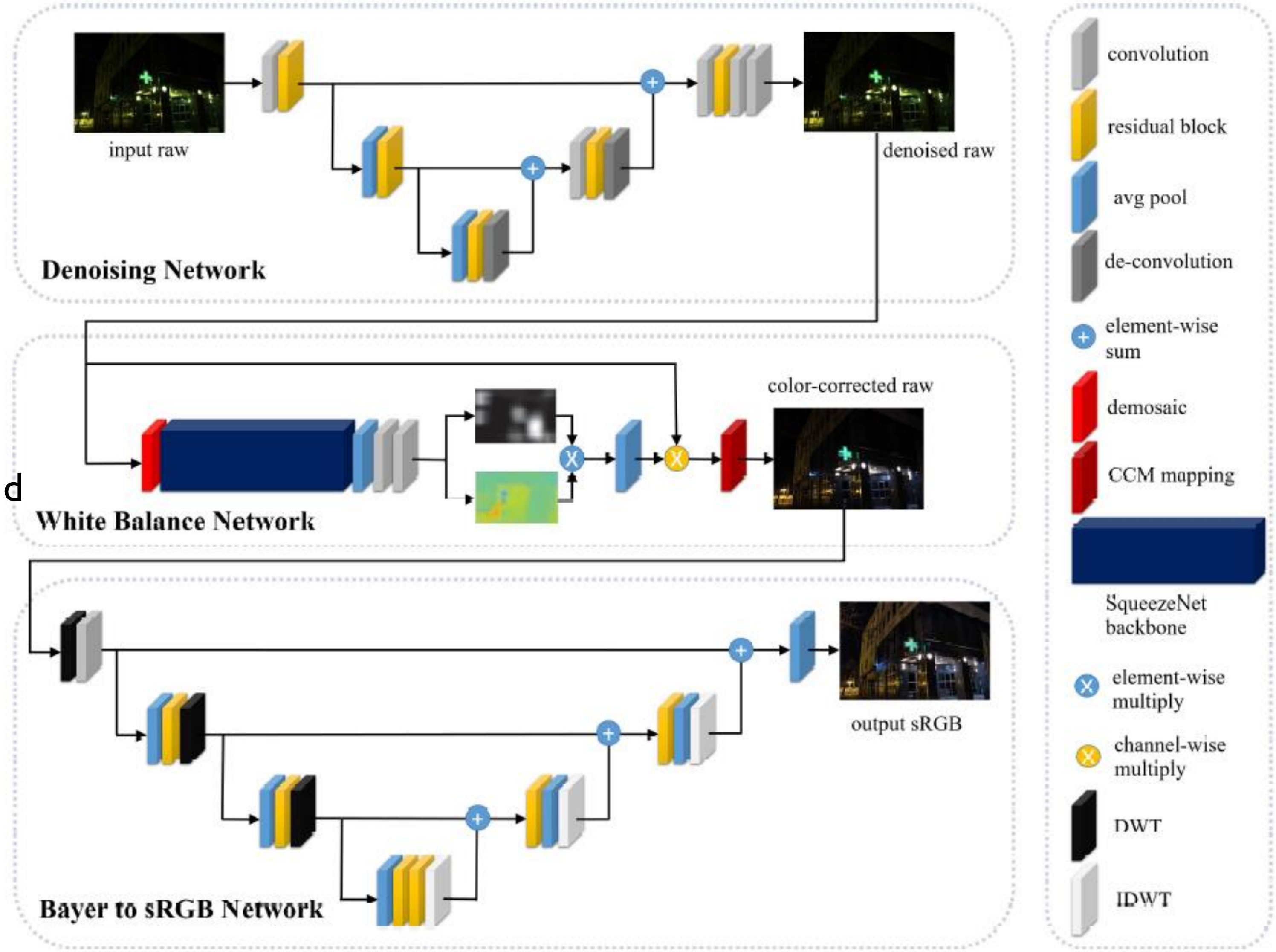
- Introduced a 3-stage ISP

FlexISP

Custom denoiser. Training data unclear (possibly in-house Xiaomi denoiser used to generate ground truth).
Network was conditioned in noise level.
Allowing adjustment.

F4C was used for white-balance.
Two networks were used, each predicting biased results towards warm/cold ground truth. User can "slide" between results.

Images were manually adjusted (lightroom?) at different levels.
Users could "slide" between results.



Baseline was a simple software ISP given to participants.

(a) Baseline

PyNet is single DNN method.

(b) PyNet

HERN (Enhancement network)

(c) HERN

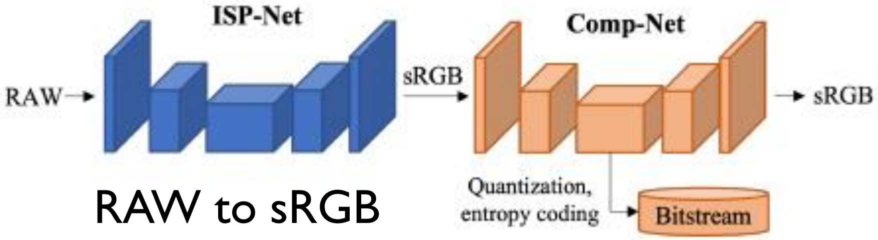
FlexISP.

(d) Ours Deep-FlexISP

Misc: RAW to bits

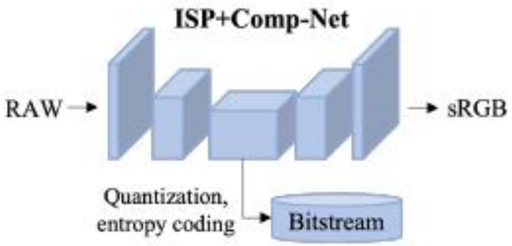
RAW to bit

Encoder/Decoder (sRGB)
[compression]



RAW to sRGB

(a)



Combined RAW to sRGB with encoder.

Jeong and Jung - ECCV'22

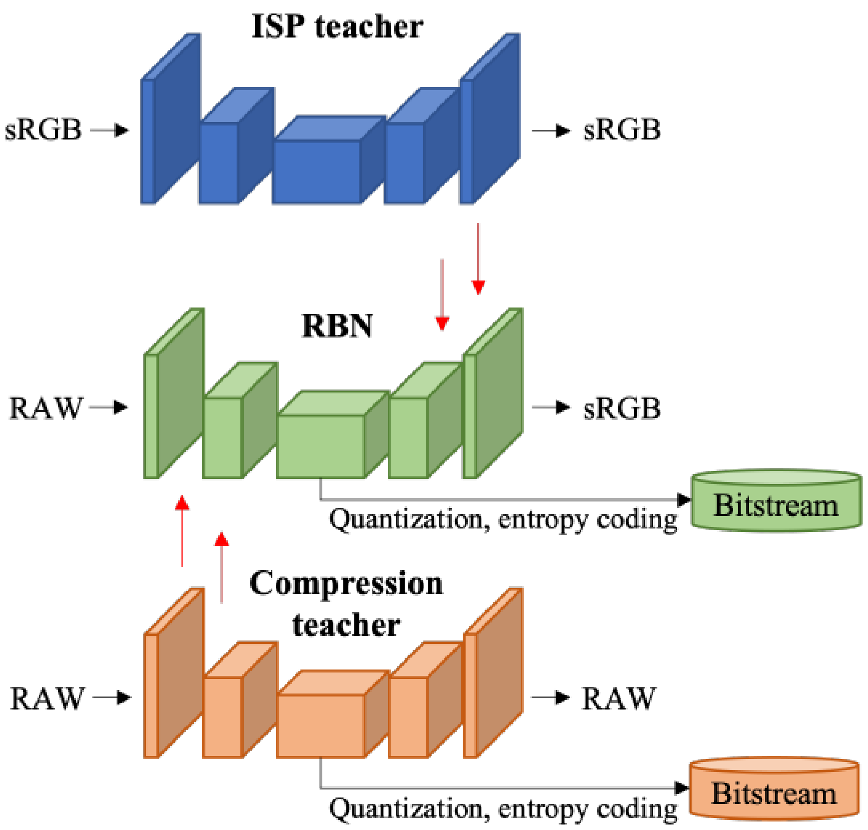
RAWtoBit: A Fully End-to-end Camera ISP Network

Wooseok Jeong and Seung-Won Jung*

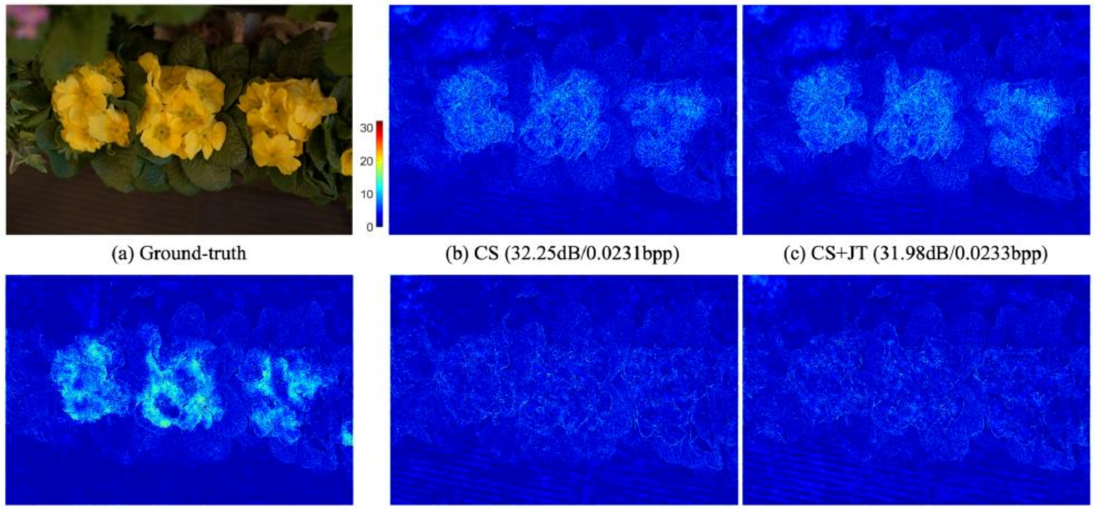
Department of Electrical Engineering, Korea University, Seoul, Korea
{561wesd, swjung83}@korea.ac.kr

Abstract. Image compression is an essential and last processing unit in the camera image signal processing (ISP) pipeline. While many studies have been made to replace the conventional ISP pipeline with a single end-to-end optimized deep learning model, image compression is barely considered as a part of the model. In this paper, we investigate the designing of a fully end-to-end optimized camera ISP incorporating image compression. To this end, we propose RAWtoBit network (RBN) that can effectively perform both tasks simultaneously. RBN is further improved with a novel knowledge distillation scheme by introducing two teacher networks specialized in each task. Extensive experiments demonstrate that our proposed method significantly outperforms alternative approaches in terms of rate-distortion trade-off.

RAW to bit



Paper shows that a knowledge distillation strategy is best to learn the RAW to sRGB with bit encoder.



(a) Ground-truth (b) CS (32.25dB/0.0231bpp) (c) CS+JT (31.98dB/0.0233bpp)

(d) US (32.84dB/0.0310bpp) (e) RBN (33.75dB/0.0216bpp) (f) RBN+KD (34.59dB/0.0223bpp)

DNN-based ISP considerations and challenges

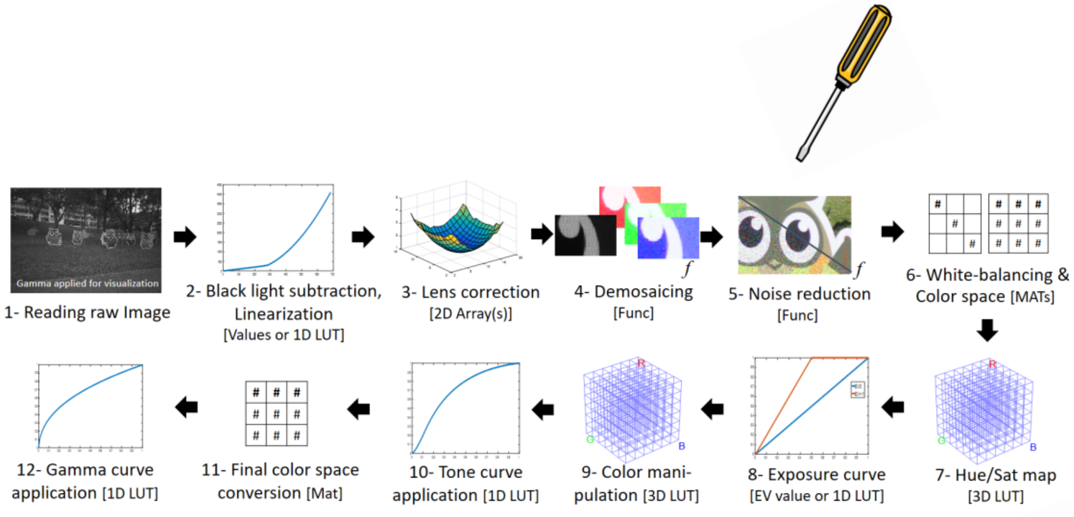
Training data

- It is important to remember that RAW images are sensor-specific
  - This means we often need to train ISPs (and ISP modules) per sensor
  - Modern smartphones can have 3-4 different sensors
  - Capturing training data can be overwhelming for camera engineers
- Care is required when capturing training data
  - Many of the low-light/HDR papers, the real contribution is the carefully captured training data
  - Again, this needs to be captured "per" sensors

- Single stage ISPs have limited "tune-ability"
  - Conventional IPS are designed to be tunable
  - DNNs are often tuned by changing the training data

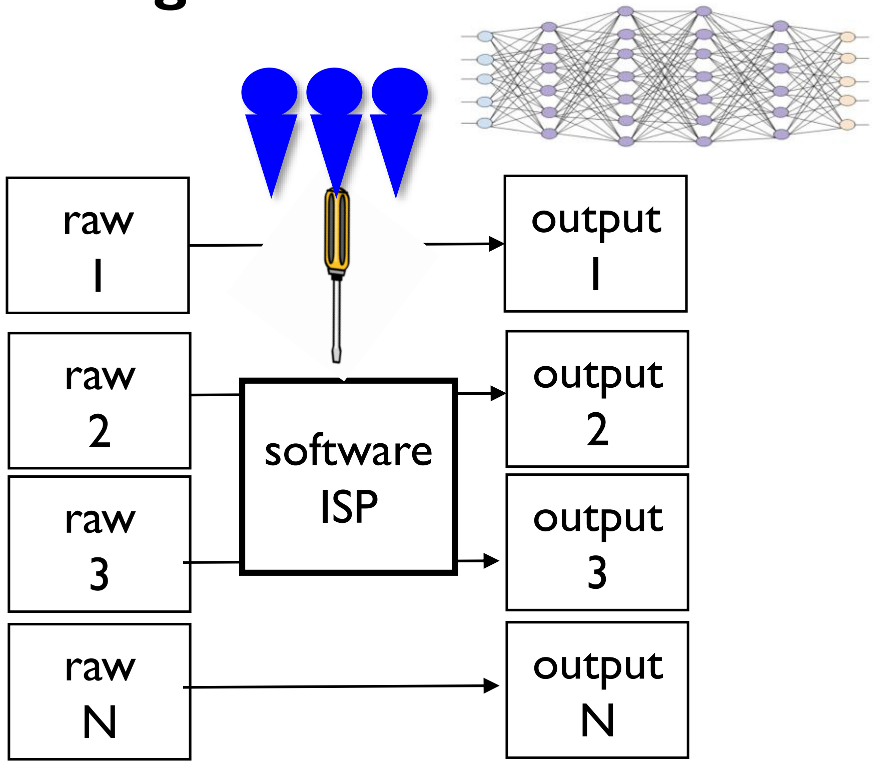
Consideration for DNN-based ISP

- **A conventional ISP is still required to produce training data**
- **Can we beat conventional ISPs?**



Current ISP workflow

Team of Image Quality Engineers tune ISP parameters to produce desired images.



Neural-ISP workflow

Team of Image Quality Engineers process thousand of RAW images with a "software ISP" to produce training data?

Tutorial summary

- Background on color and color spaces
  - This topic mixes many disciplines
  - Color constancy and terminology for illumination (e.g. color temperature)
- Overview of basic steps on camera pipeline
- Discussion of more modern multi-frame methods
- Discussion of some recent AI-based methods

Last slide (almost)

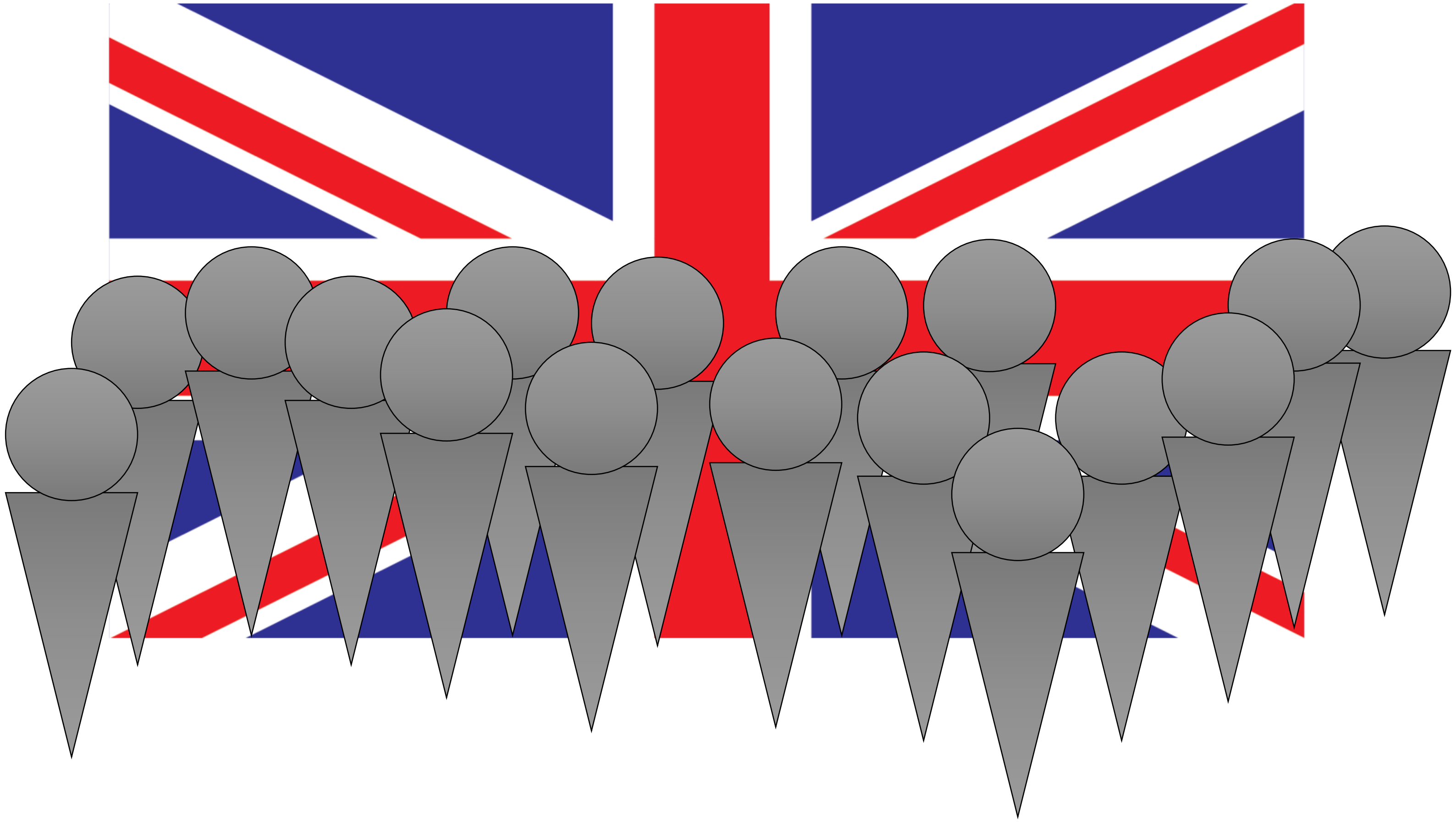
- I hope you have learned more about color and the in-camera rendering pipeline.
- I encourage you to state your assumptions about your image's color space in your research papers:

  For example, replace this: "Our input is an RGB image …"
  to: "Our input is an RGB image encoded in standard RGB…"

  - Such a small clarification in your paper will greatly help other researchers.

Acknowledgements

And of course…



"The Standard Observers"

Funding acknowledgement

Adobe

Microsoft Research

SAMSUNG

CRSNG
NSERC

LASSONDE
SCHOOL OF ENGINEERING

YORK
UNIVERSITÉ
UNIVERSITY

Chaires de recherche du Canada

Canada Research Chairs

Canada

Canada First
Research Excellence Fund

Vision Science to Applications (VISTA)

SINGAPORE

Agency for Science, Technology and Research