Humans and Compression

EE 274, Fall 23

11/29/23

Announcements

- Next Lecture (Mon, 12/4)
  - Guest talk from **Kedar Tatwawadi** on **Video Compression**
  - Course summary and wrap-up
- Reminder: no lecture on Wed, 12/6; instead we will have final project presentations. We will release more information by this weekend.
  - Main Idea: present to peers what you learnt and results so-far
  - 5 mins presentation, 5 mins QnA
  - Mandatory attendance

So far in Lossy Compression

- Quantization as a core mechanism of introducing loss
- Theoretical underpinnings of rate-distortion trade-off
- Optimal solutions in case of Gaussian sources and MSE distortion
- Image Compression and JPEG
- Learnt Image Compression

This class

All multimedia is eventually consumed by humans:

Role of *human sensory perception* in designing lossy multimedia compressors

- Why some of the design decisions were made in the image compressors we saw
- How can we further improve image/video compression accounting for human perception
- Going beyond MSE as a distortion metric

Resources

- Foundations of Vision, Brian A. Wandell
    - Check out PSYCH 221: Image Systems Engineering
- **Papers**:
  Perceptual Video Compression: A Survey, SSIM, MS-SSIM, Image Quality Metric Comparison, LPIPS, HiFiC, RDP Tradeoff, DISTS
- **Blogs**:
  VMAF, The ultimate guide to JPEG including JPEG Compression and Encoding, Optical Illusions
- **Videos**:
  Color Space, Opponent Color Theory
- Images obtained by doing a simple google search

Disclaimer
some of the material presented in this class will have hand-wavy coverage from neuro-scientific and psycho-visual literature

Teaser 1

audio compression

Sampling rate of 44.1 kHz is very common in encoded audio. Can you guess why?

Why do we have 2 channels (stereo) in encoded audio?

```
~/Downloads  mediainfo pokemon_theme.mp3
General
Complete name                            : pokemon_theme.mp3
Format                                   : MPEG Audio
File size                                : 3.03 MiB
Duration                                 : 3 min 18 s
Overall bit rate mode                    : Constant
Overall bit rate                         : 128 kb/s
Album                                    : Pokemon X: Ten Years of Pokémon
Album/Performer                          : Pokemon
Track name                               : Pokemon Theme
Performer                                : Pokemon
Genre                                    : Soundtrack

Audio
Format                                   : MPEG Audio
Format version                           : Version 1
Format profile                           : Layer 3
Format settings                          : Joint stereo / Intensity Stereo + MS Stereo
Duration                                 : 3 min 18 s
Bit rate mode                            : Constant
Bit rate                                 : 128 kb/s
Channel(s)                               : 2 channels
Sampling rate                            : 44.1 kHz
Frame rate                               : 38.281 FPS (1152 SPF)
Compression mode                         : Lossy
Stream size                              : 3.02 MiB (100%)
```

Teaser 2

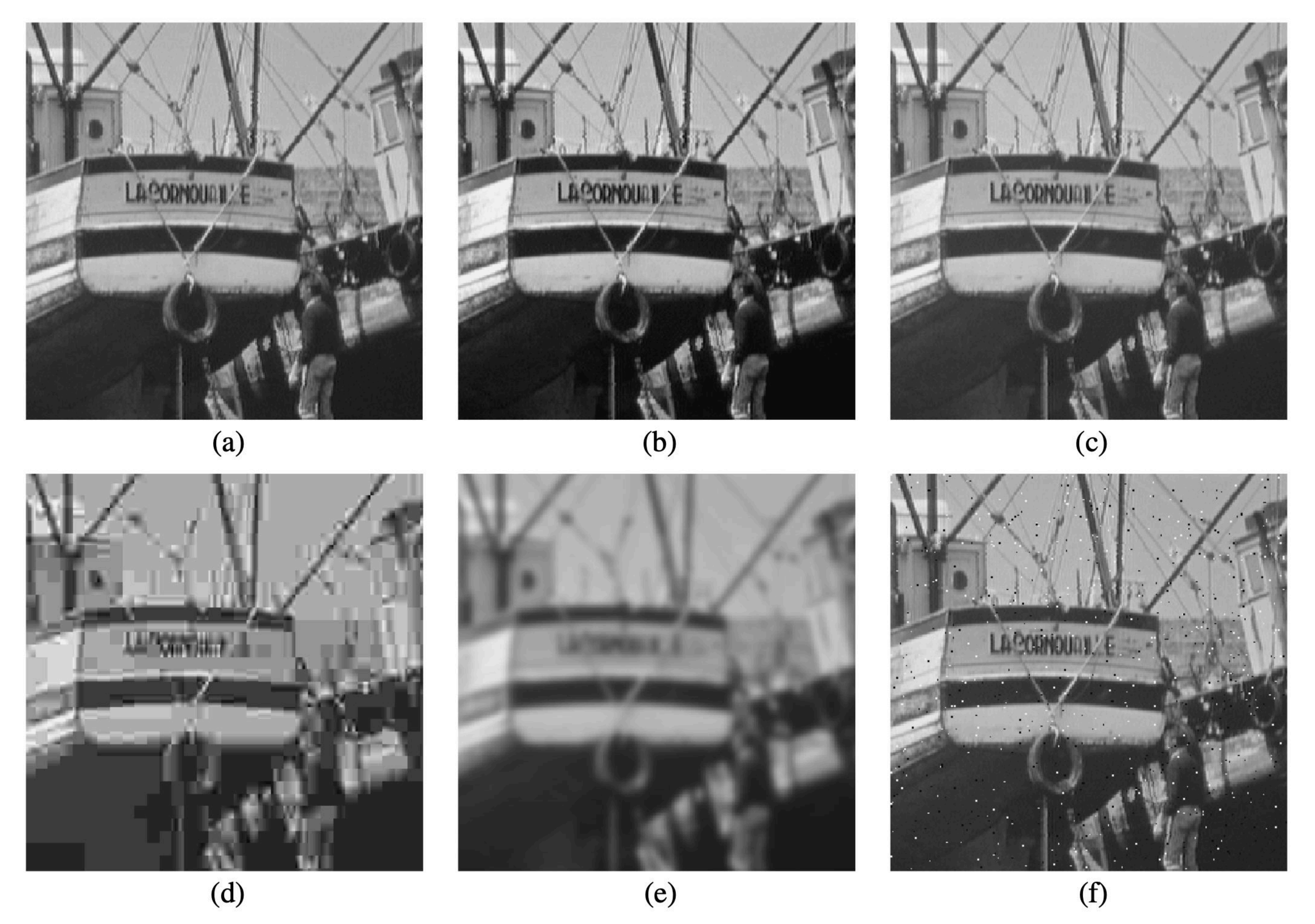
image compression

Given source image (a) which of the following images do you prefer visually?

(b), (c), (d), (e), (f)

Given source image (a) which of the following images does a compressor with MSE distortion prefer?

(b), (c), (d), (e), (f)

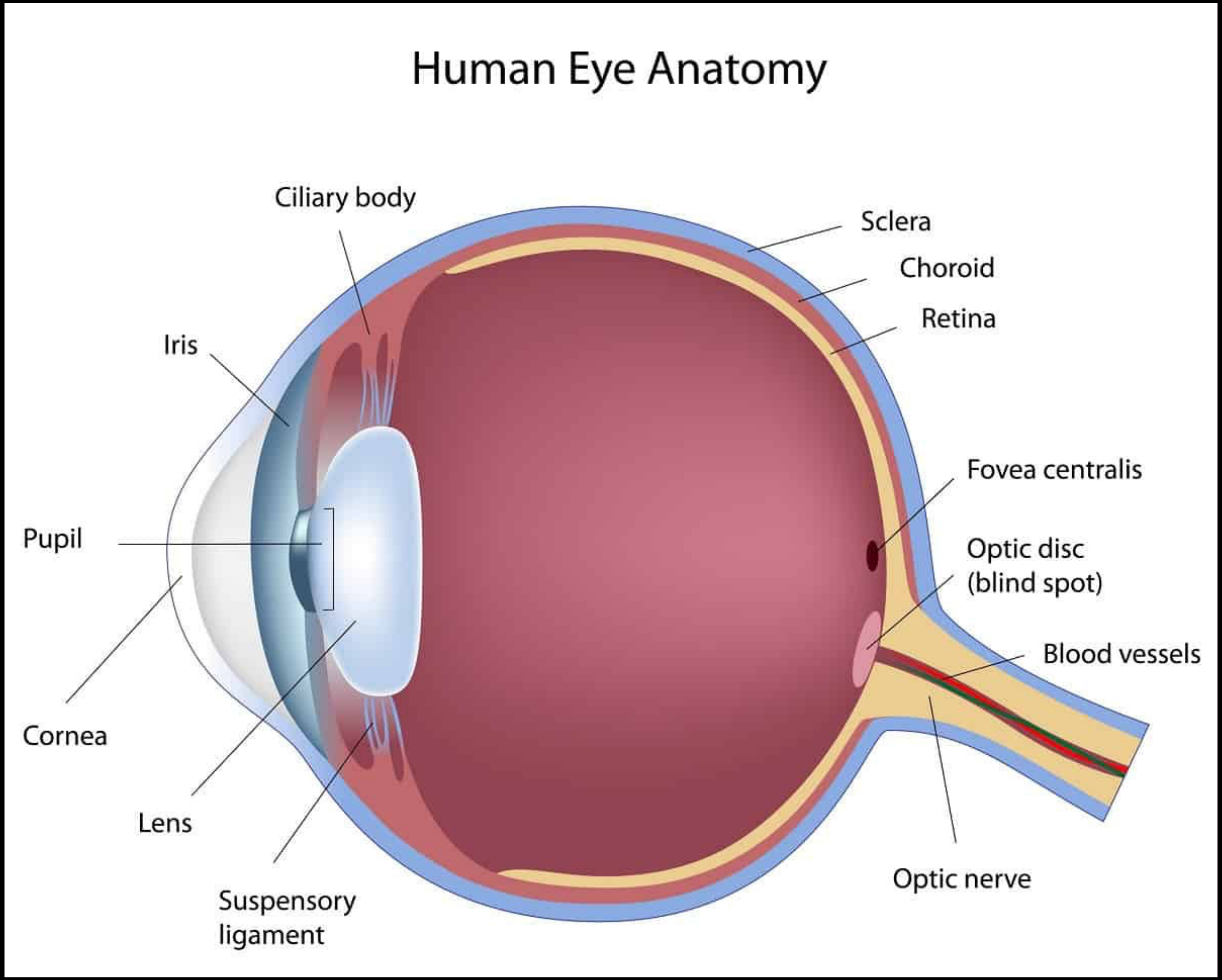


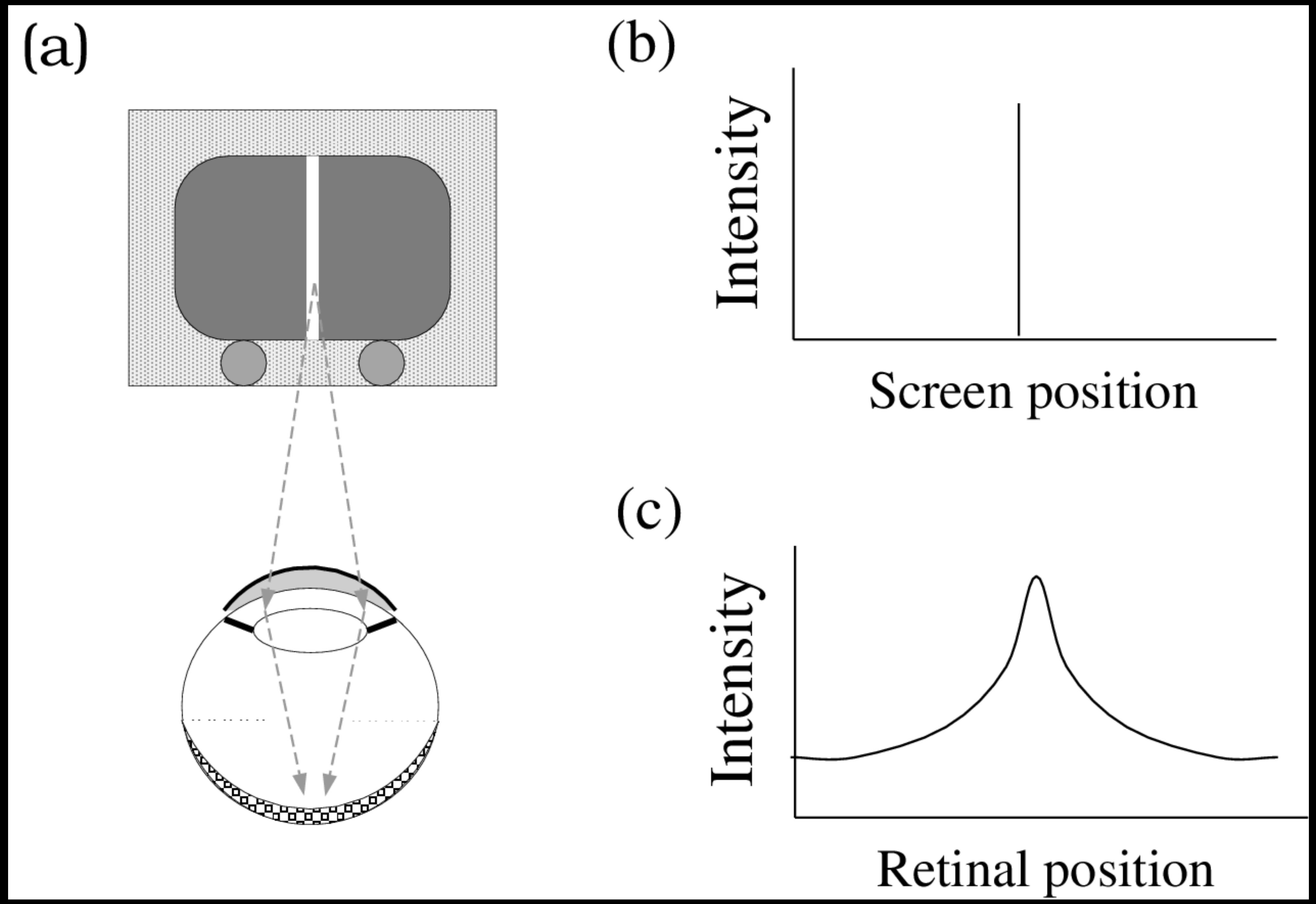
(a) (b) (c) (d) (e) (f)

Rest of the lecture

- Will get some preliminary understanding of how human vision works
- See it’s role in design of traditional compressors such as JPEG
- Learn more about perceptual metrics
- How to take into account perceptual properties in the RD framework

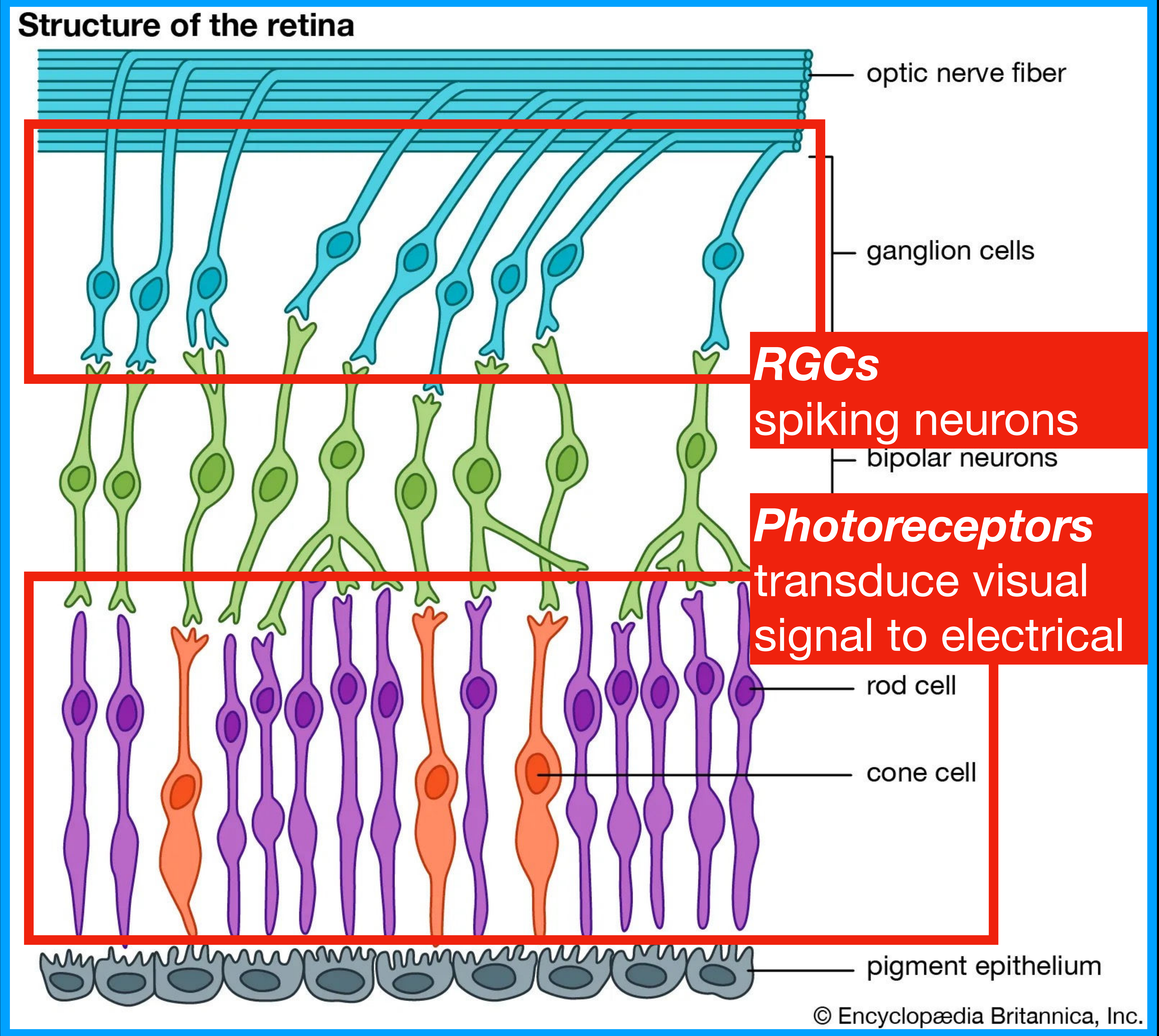
Part 1:
Human Vision and it's implications on image encoding

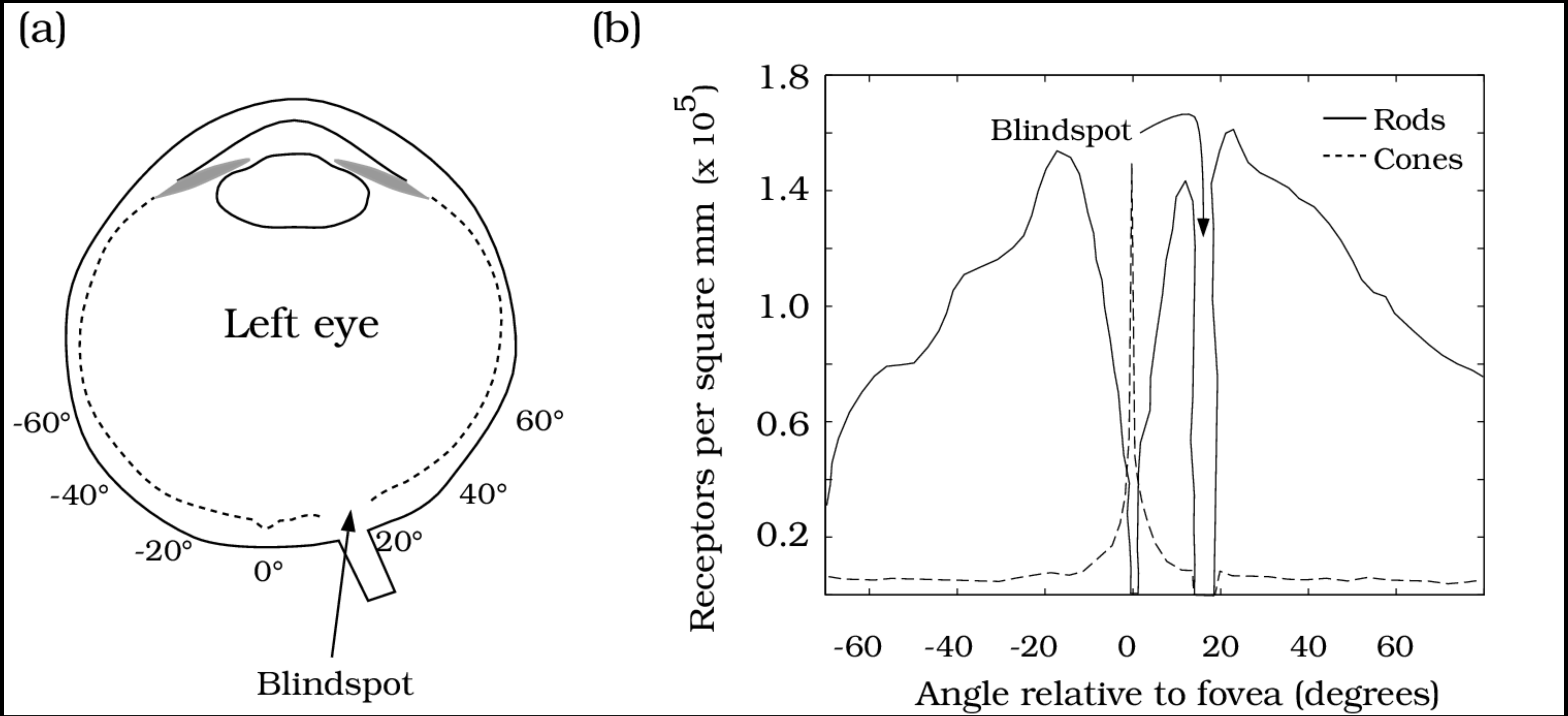




Structure of the retina
optic nerve fiber
ganglion cells
bipolar neurons
rod cell
cone cell
pigment epithelium
© Encyclopædia Britannica, Inc.

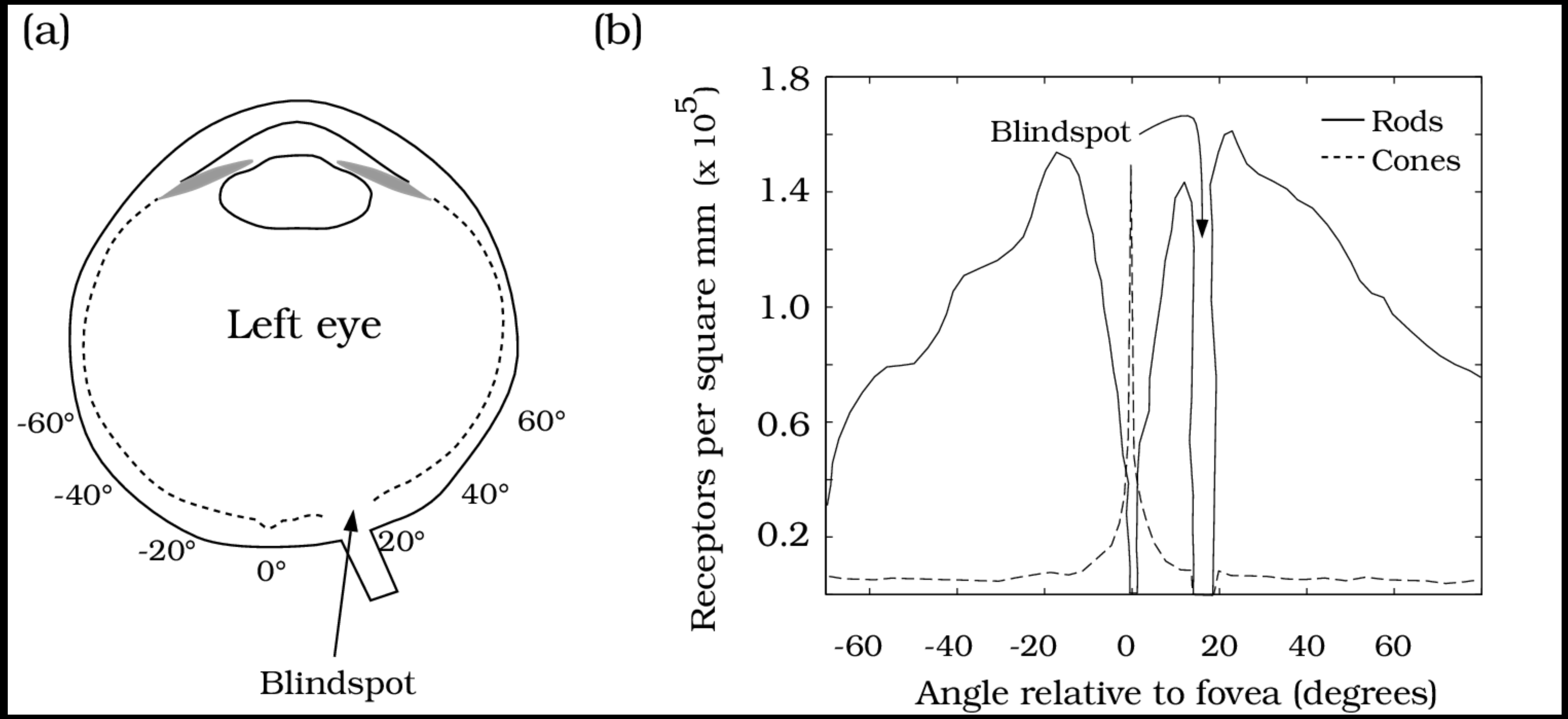
visual cortex





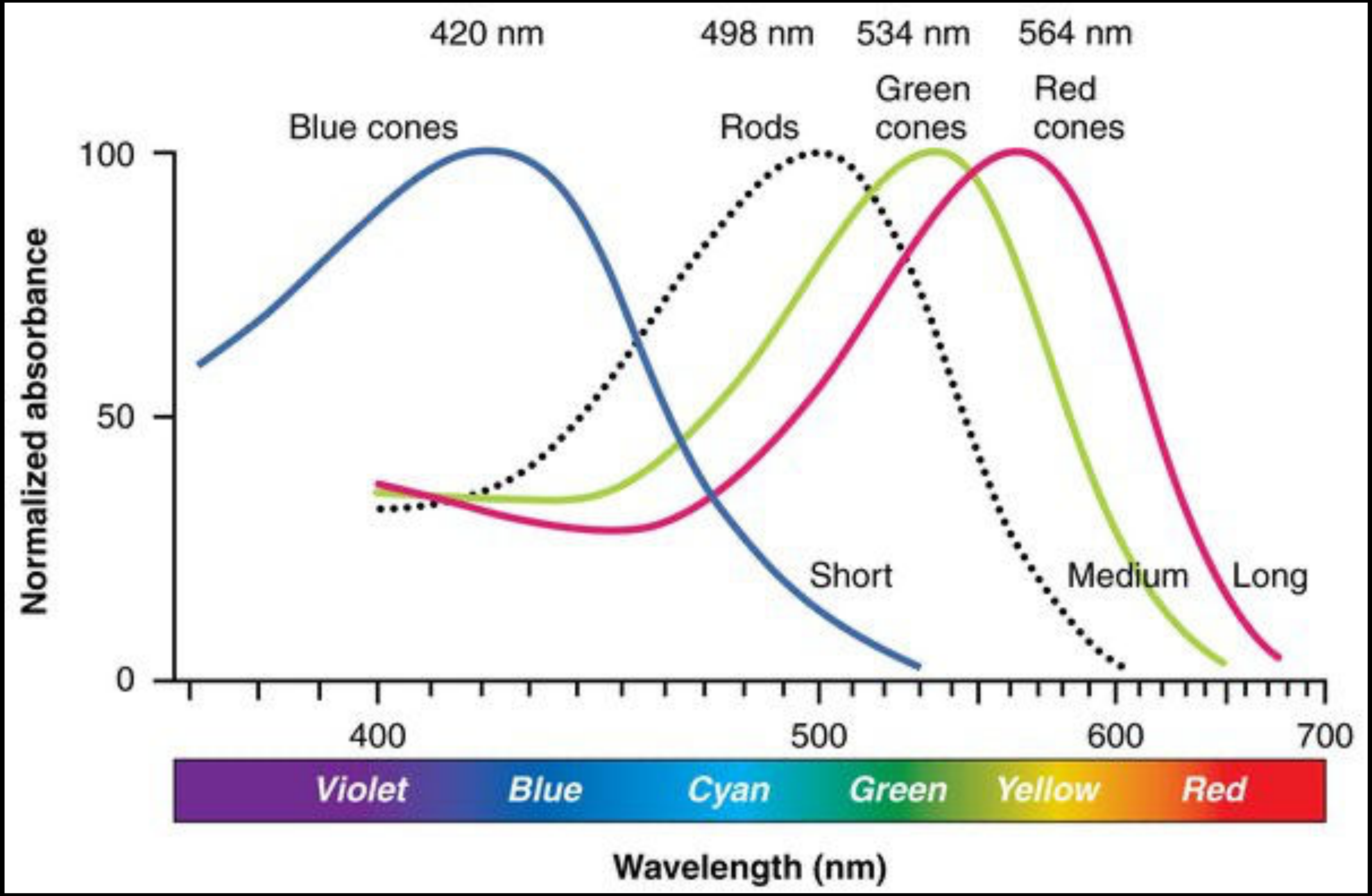
- ***Rods*** responsible for encoding intensity, ~100×10^6, absent from **fovea**
- ***Cones*** responsible for encoding colors, ~5×10^6, concentrated only in **fovea**

Some Fun (and *useful*) Observations



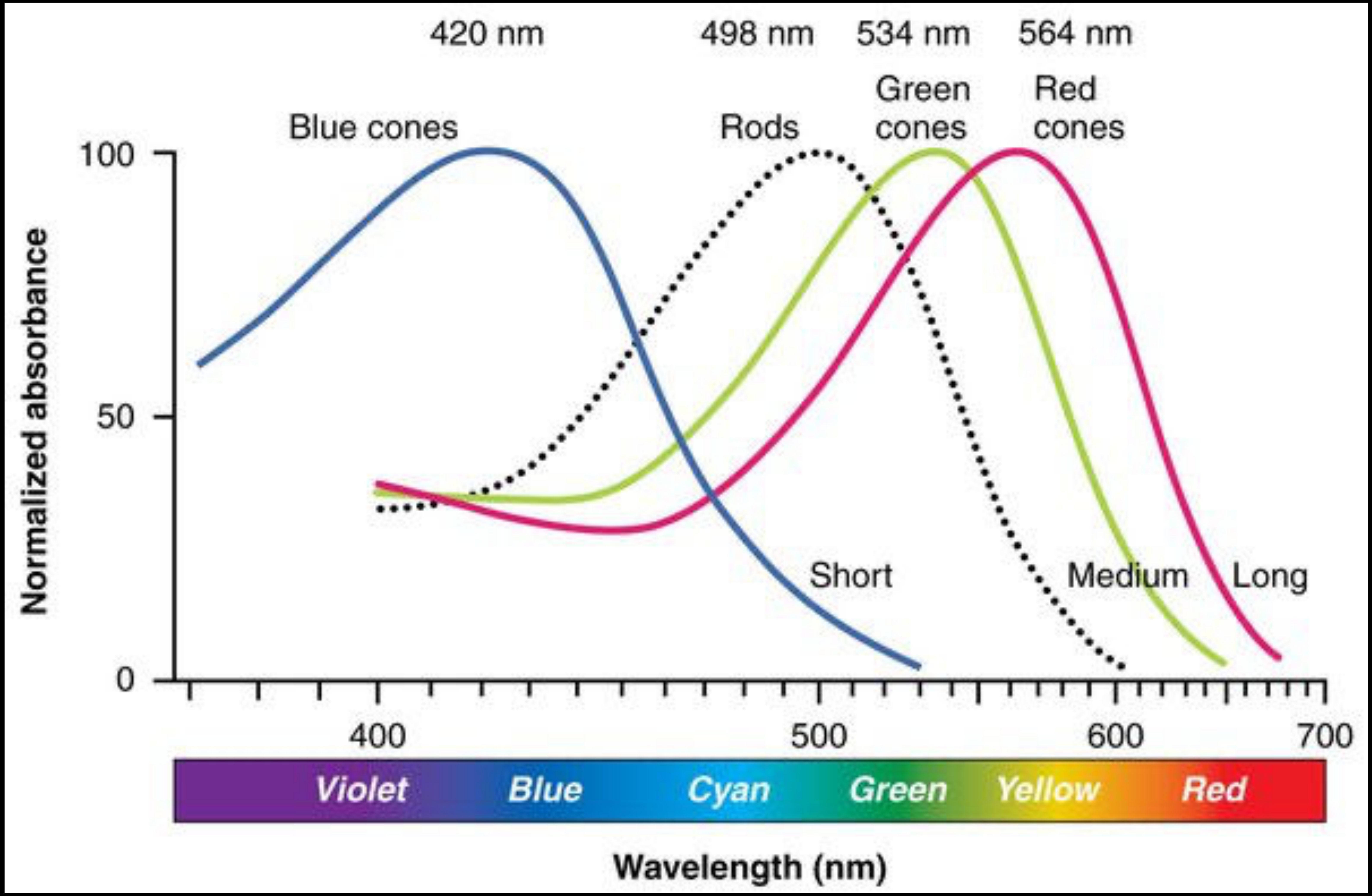
- ***Rods*** are extremely sensitive to intensity, respond even to single photon! Responsible for adaptation to wide-range of luminance in natural world. The amount of just noticeable-luminance is proportional to luminance, called ***Weber's Law***

- ***Fovea*** is region of high visual acuity and responsible for high spatial resolution. We are constantly sampling a visual scene to align it with our fovea, called ***Foveation***



- ***Cones*** come in three varieties: L, M, S cones

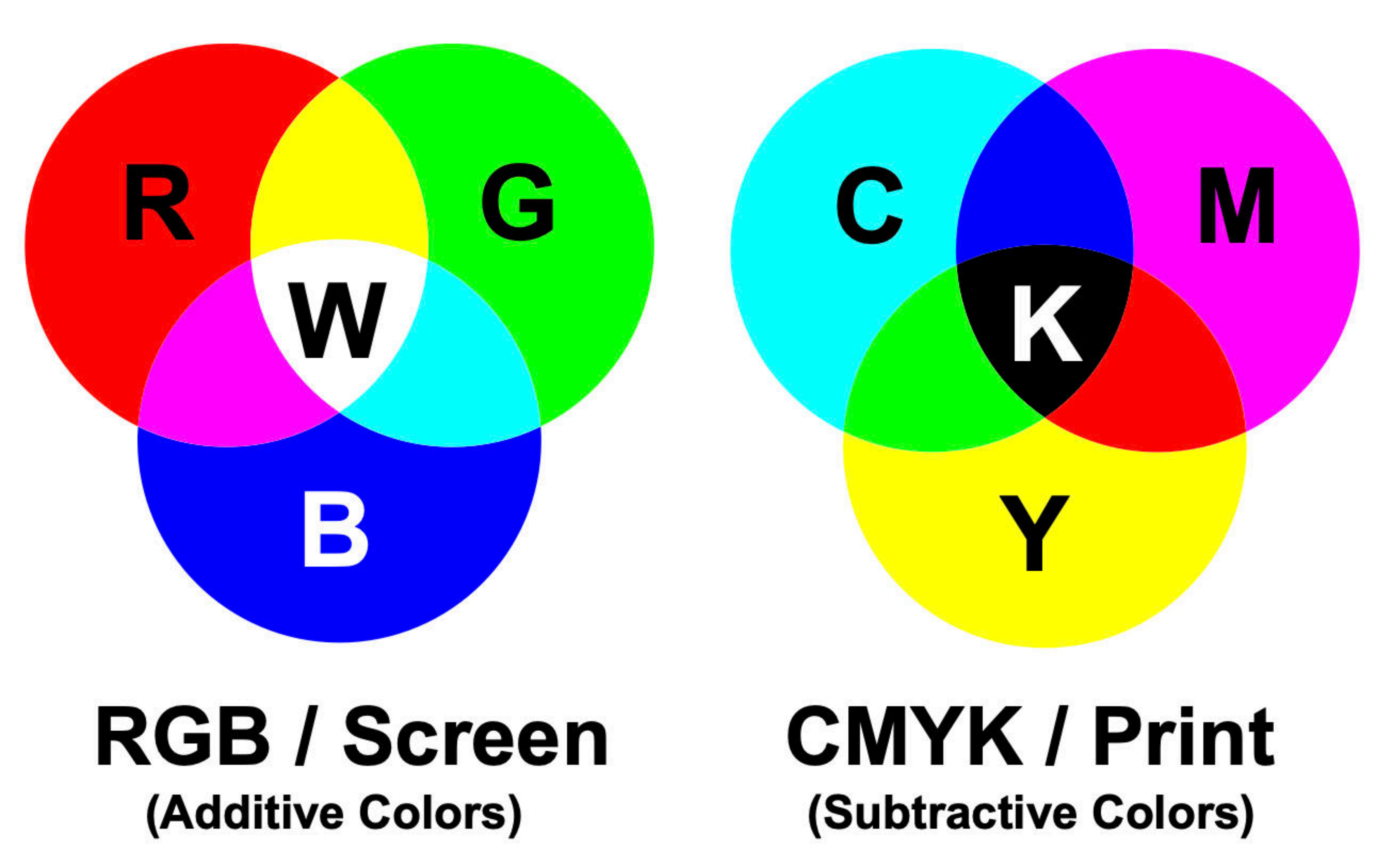
Q. What is the significance of the L, M, S cones?

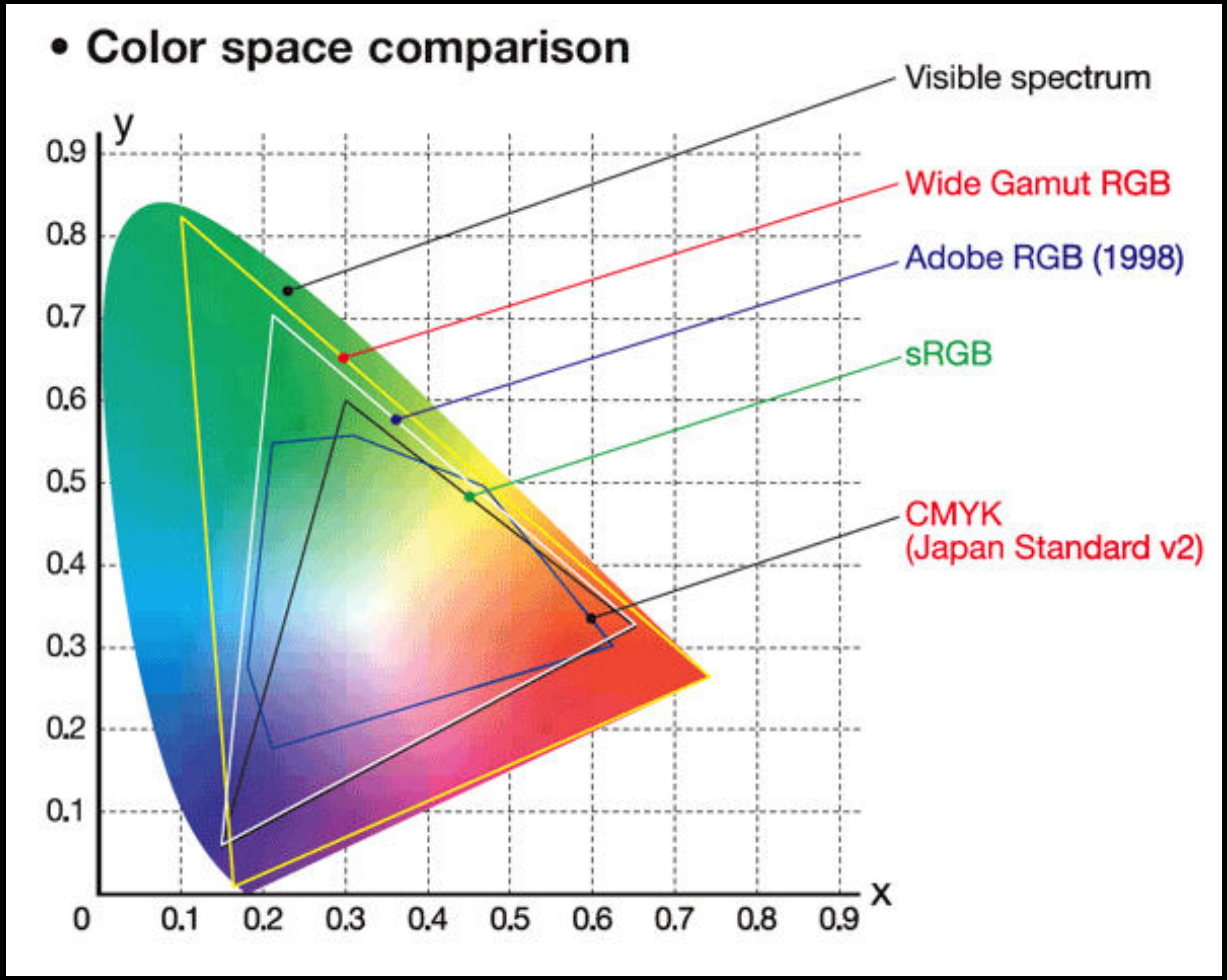


- ***Cones*** come in three varieties: L, M, S cones

Origin of RGB color model! Called trichromatic theory of color vision

But what is a color model? Is RGB special?





CIE 1931 xy chromaticity diagram

Illusion Time

Lilac Chaser

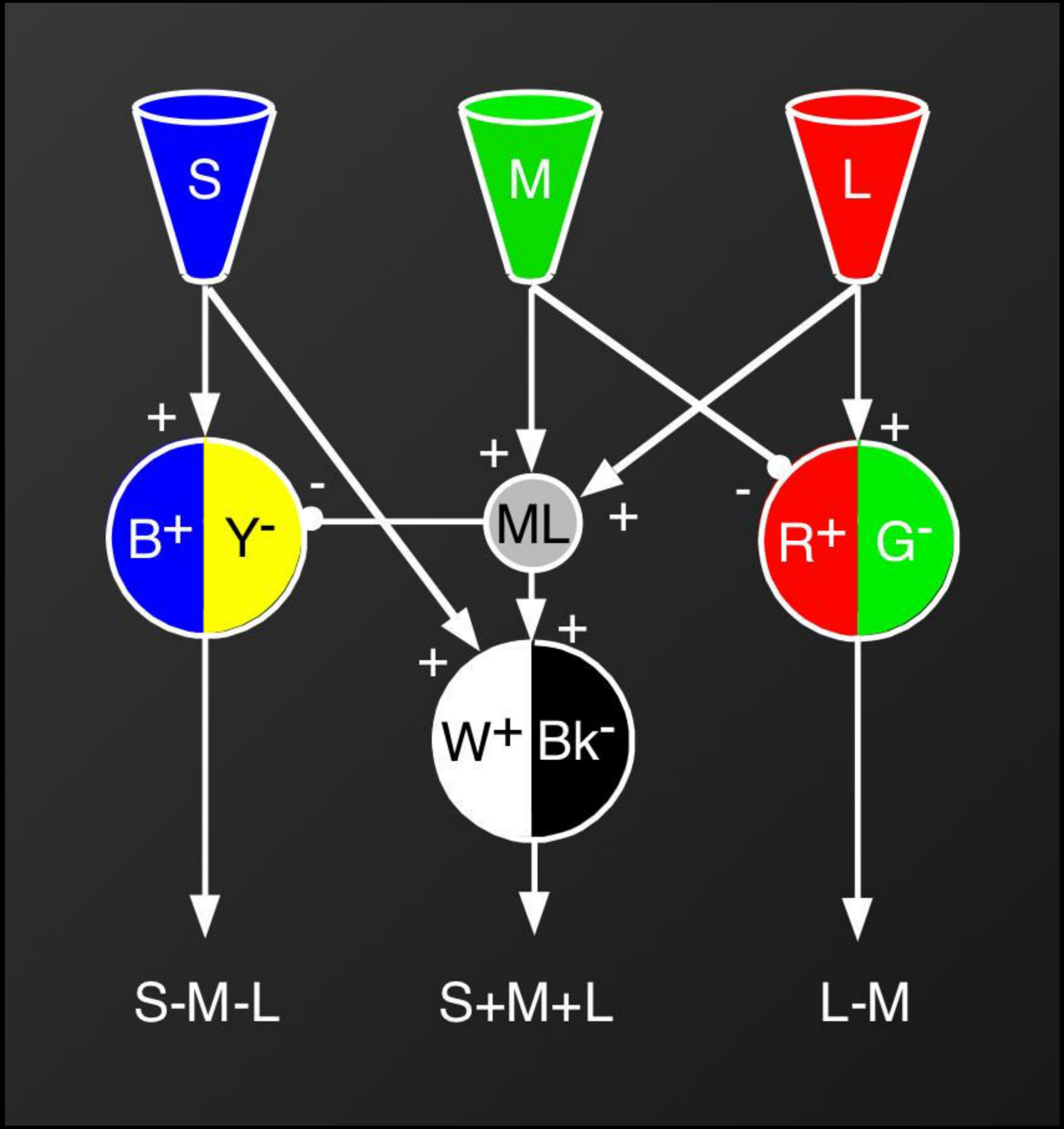
Opponent-process theory of color vision

Instead of RGB we perceive colors through three opponent channels:

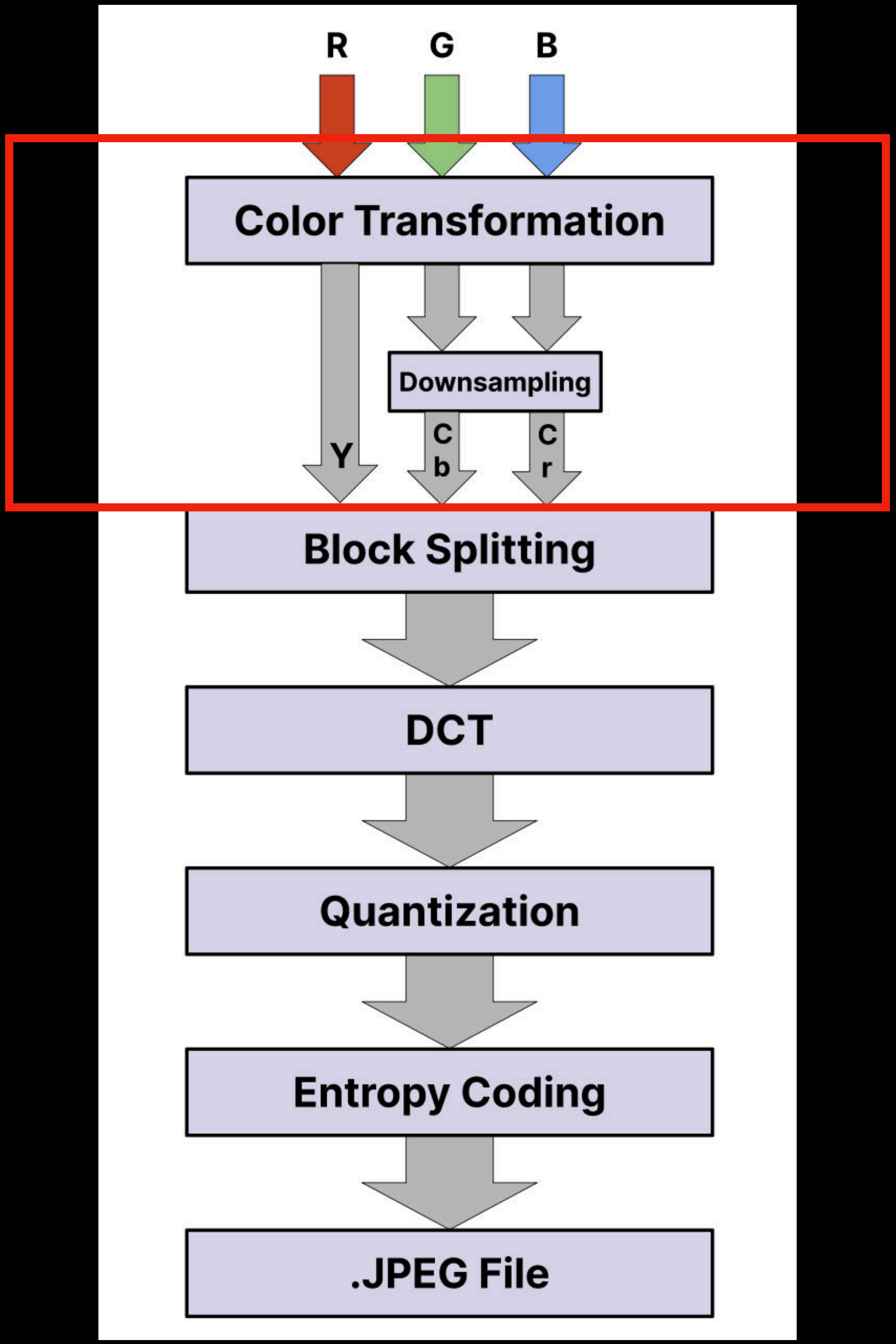
White-Black

Blue-Yellow

Red-Green

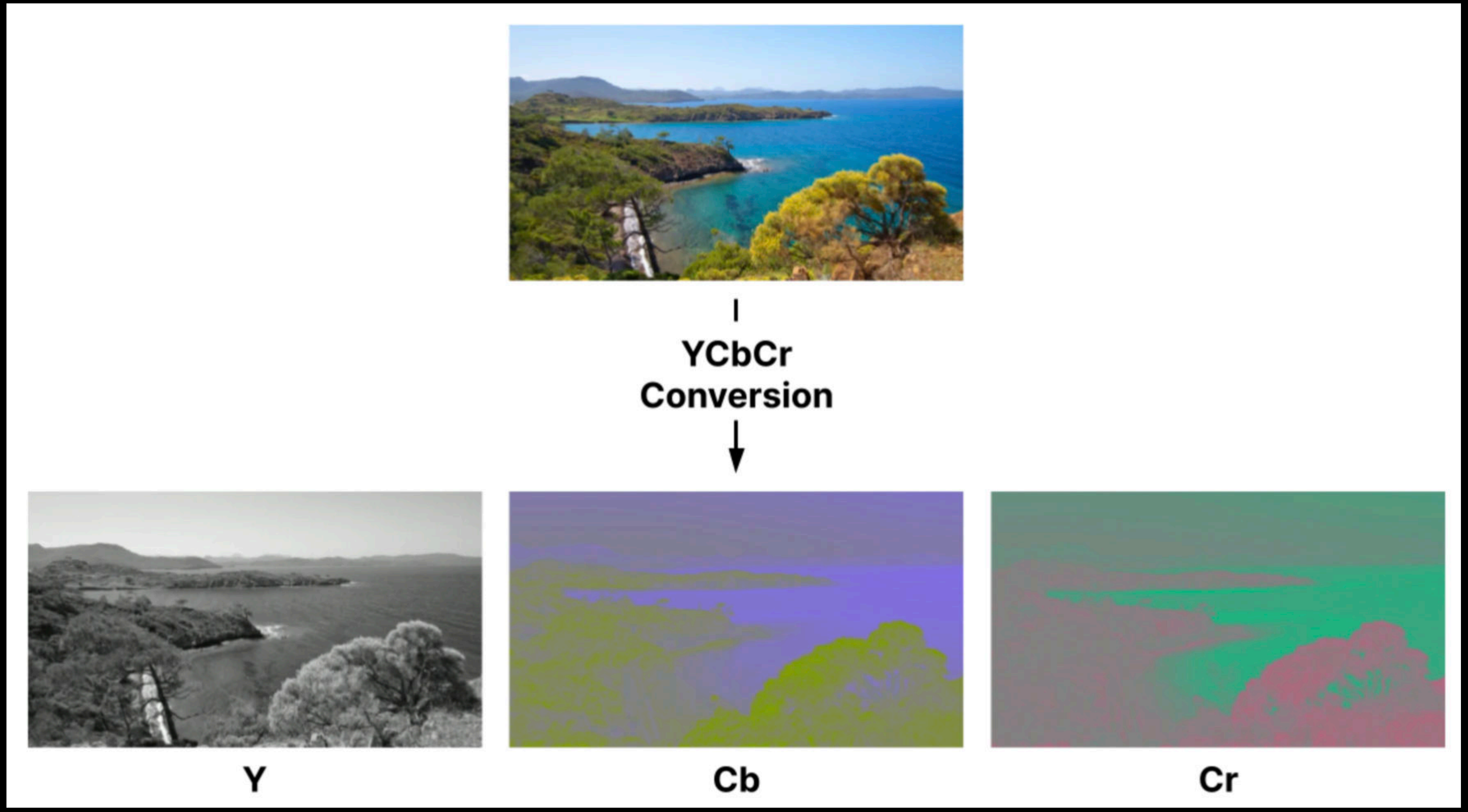


Recall: JPEG Image Compression



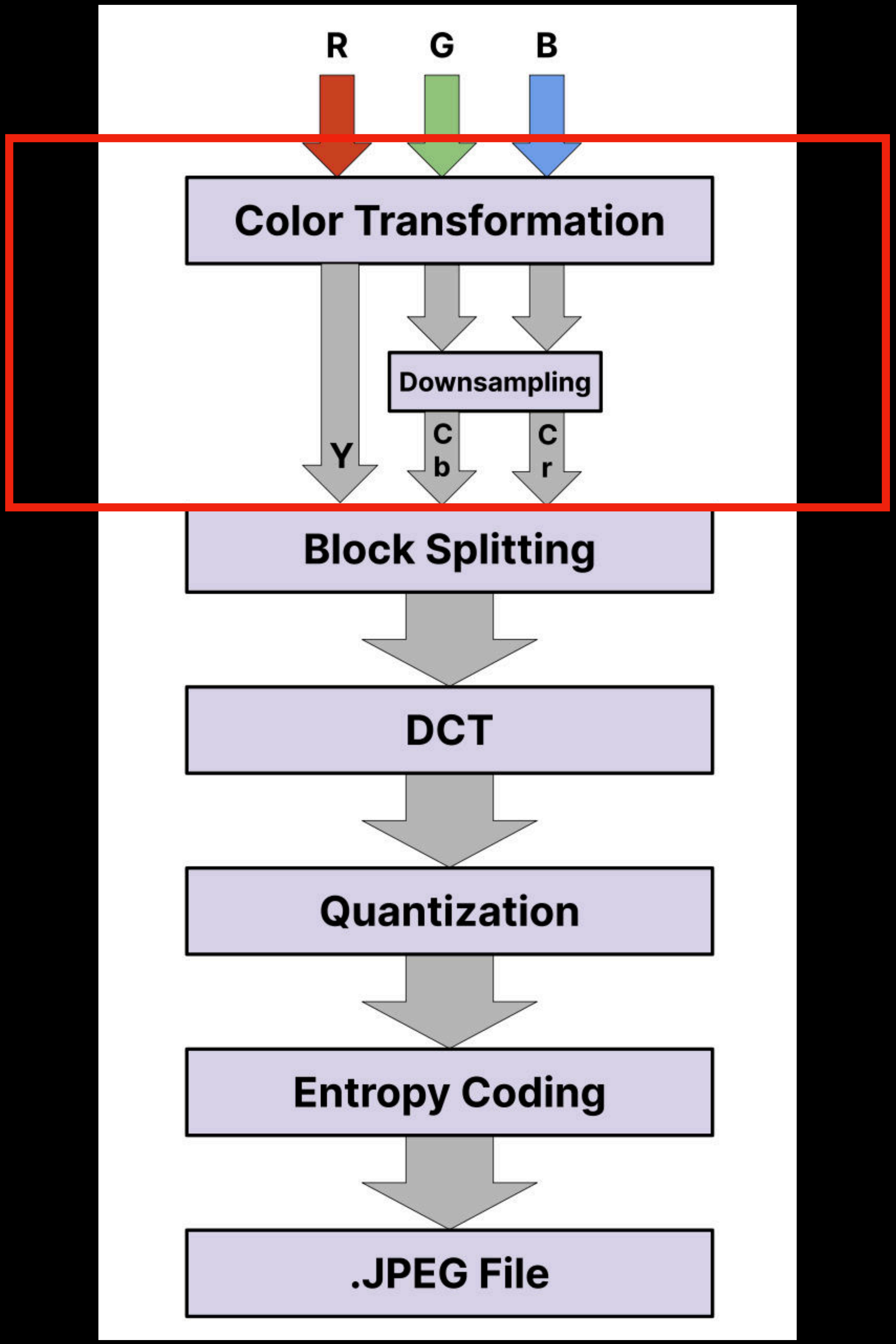
First step is Color Transformation from ***RGB*** to ***YCbCr****

But why?



Might hear* ***YUV, Y'UV, YCbCr, Y'CbCr*
Consider them all same for today.

Recall: JPEG Image Compression

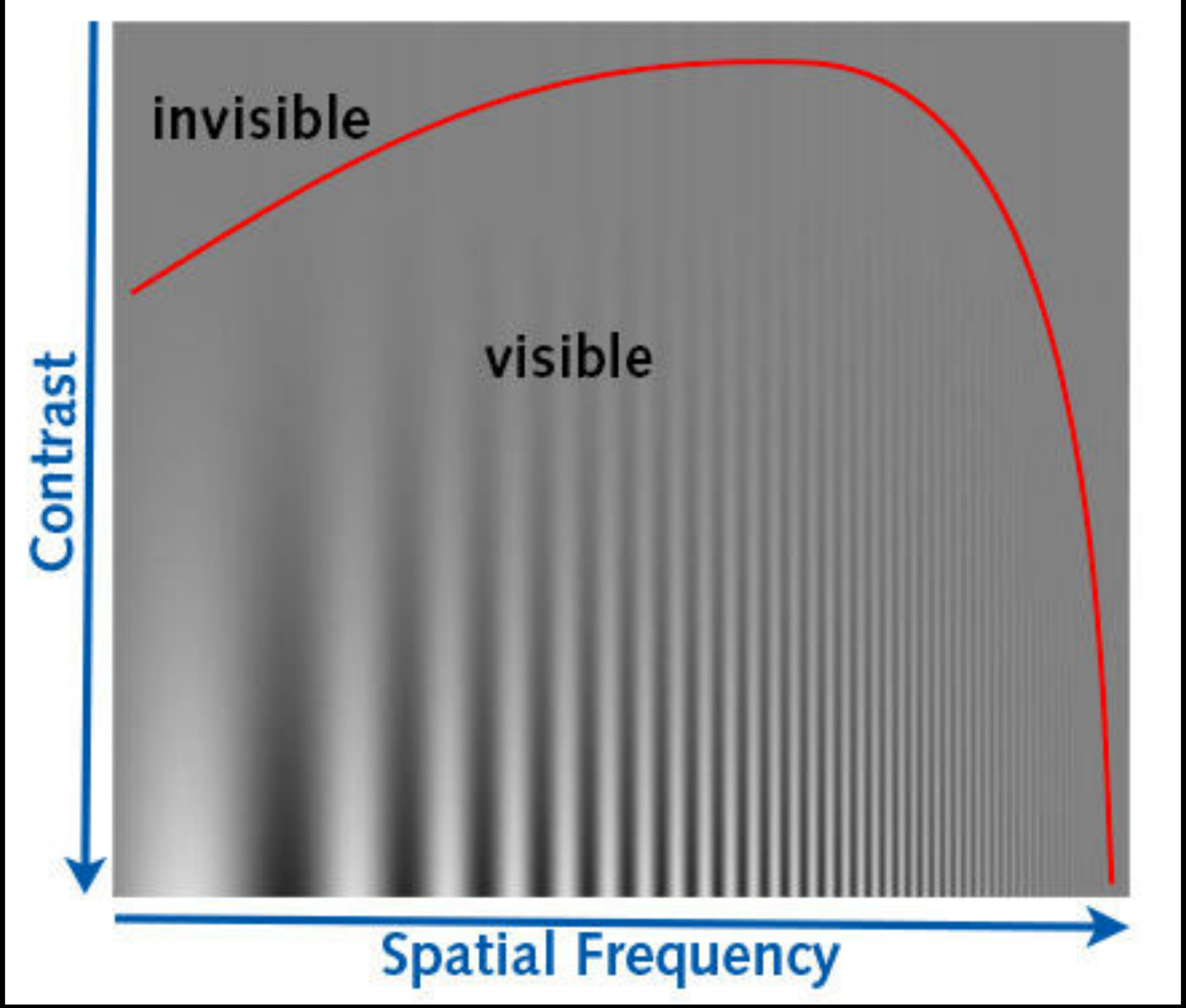


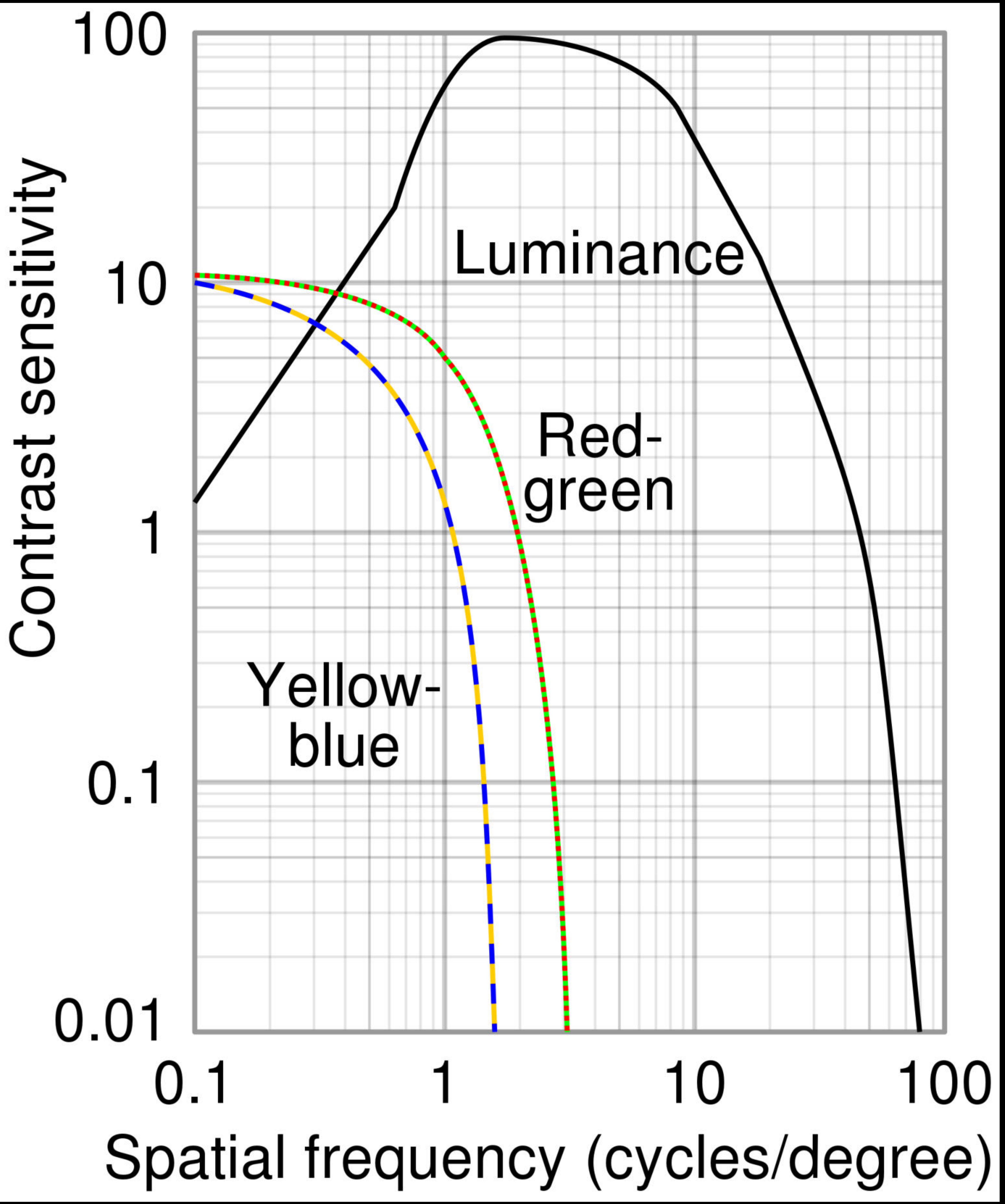
First step is Color Transformation from ***RGB*** to ***YCbCr***

But why?

- Reason 1:
  perceptual color space based on opponent process theory of color vision
- Reason 2:
  different ***contrast sensitivity*** of Y, Cb, Cr channels

Contrast Sensitivity





Contrast Sensitivity

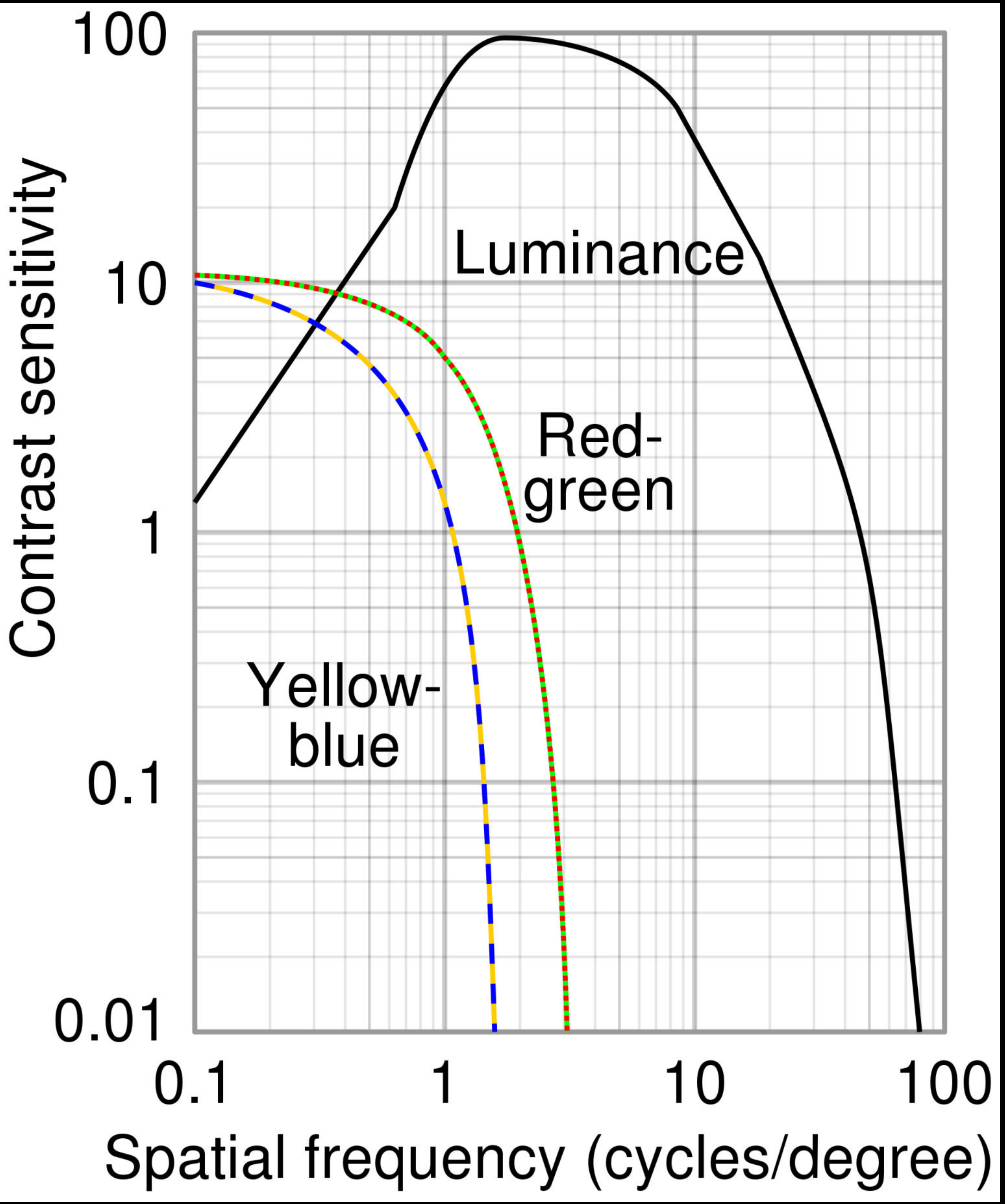
The contrast sensitivity curve has lots of implications for us:

- Higher quantization of high-frequency DCT components

- Chroma Subsampling

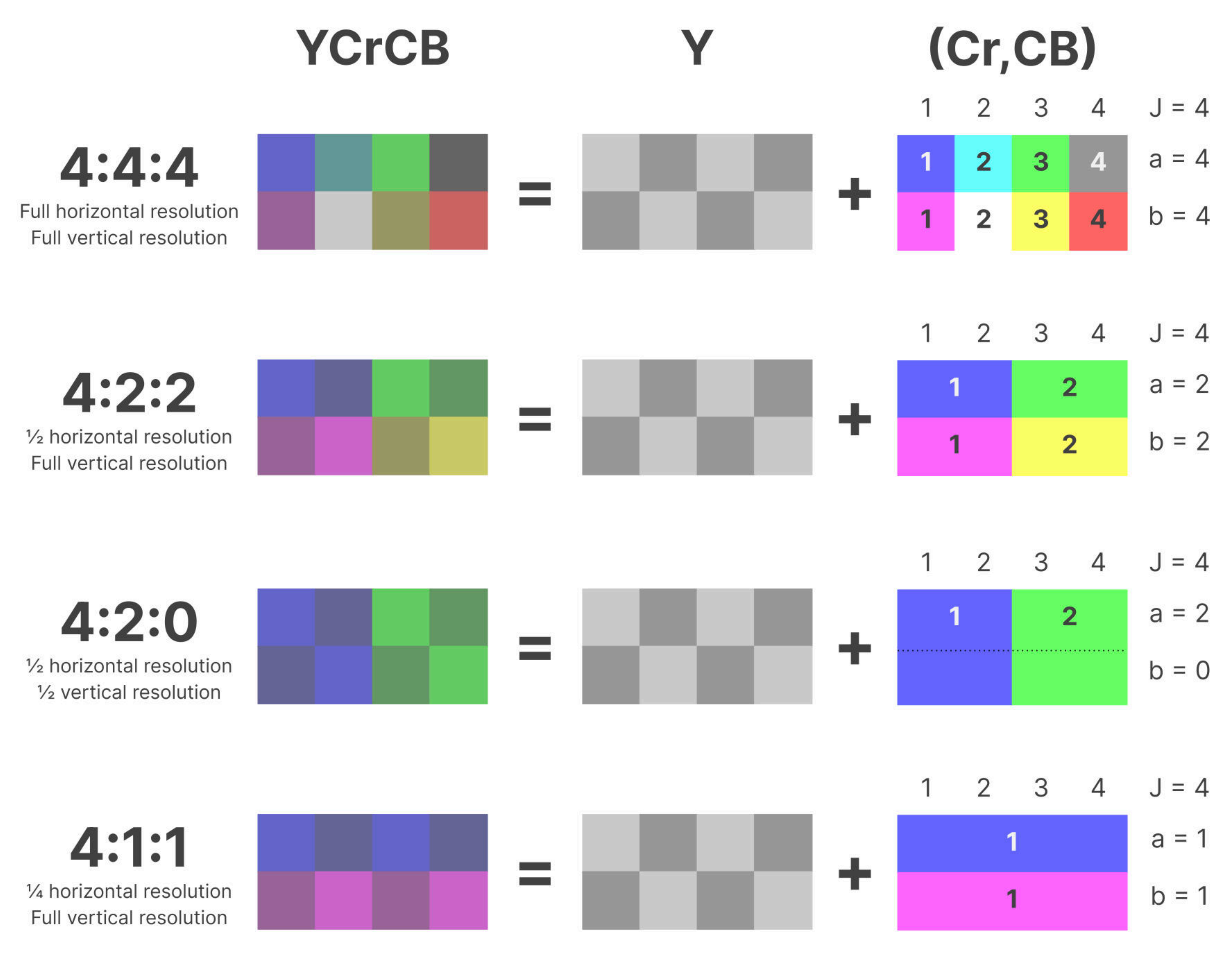
- Separate Quantization Matrices for Luma and Chroma Components

All of these are used in JPEG



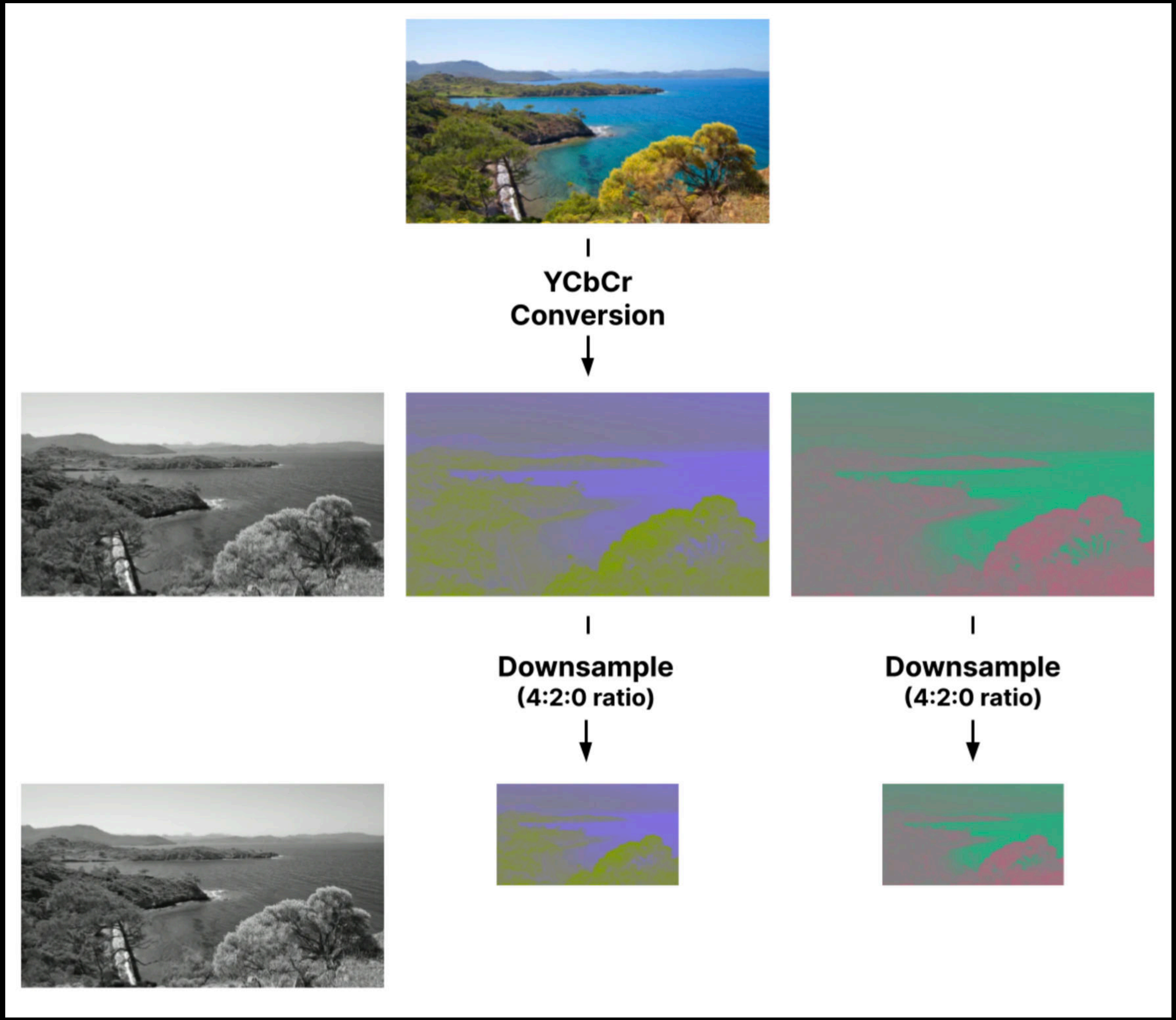
Chroma Subsampling

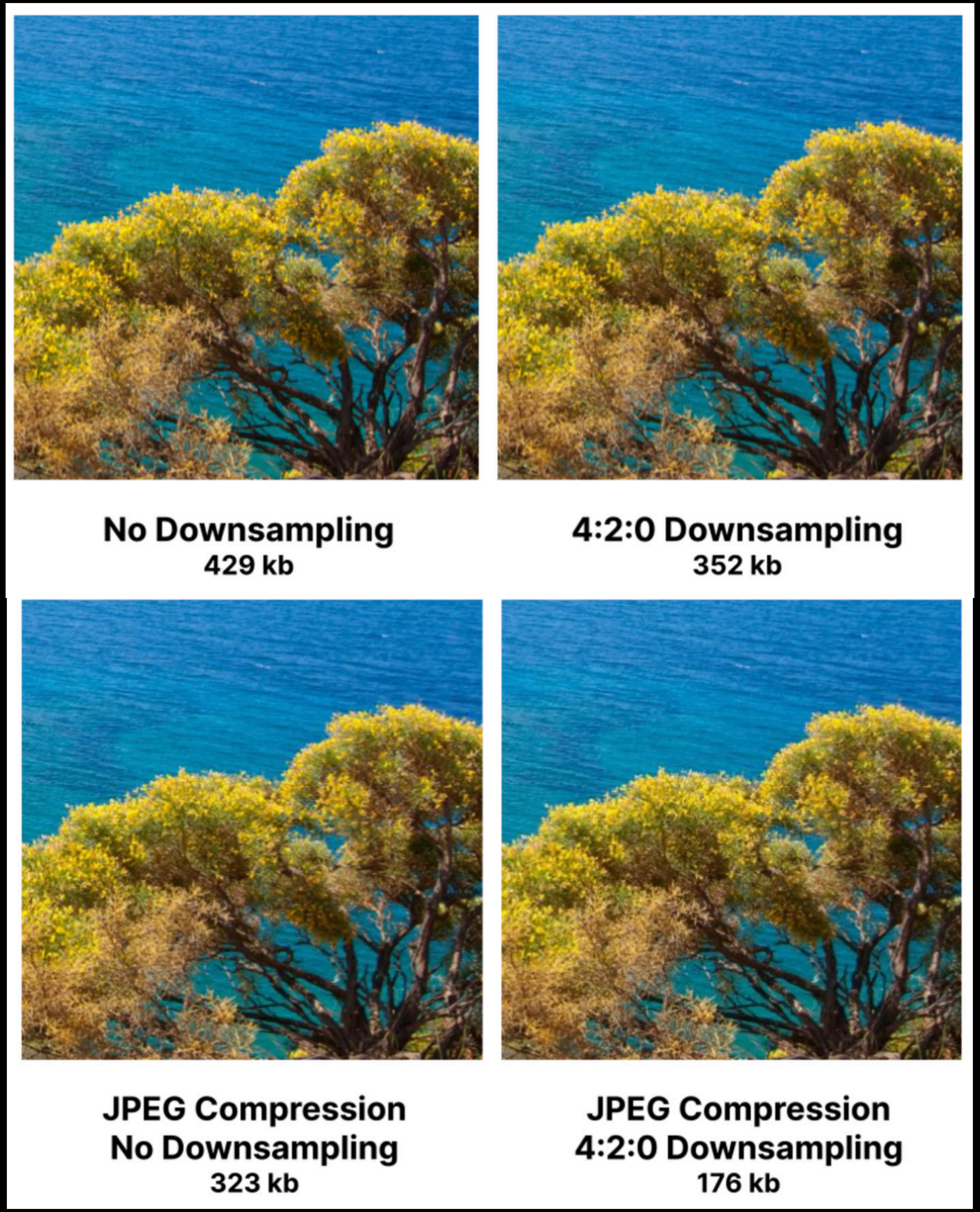
downsample color information in Cb and Cr channels



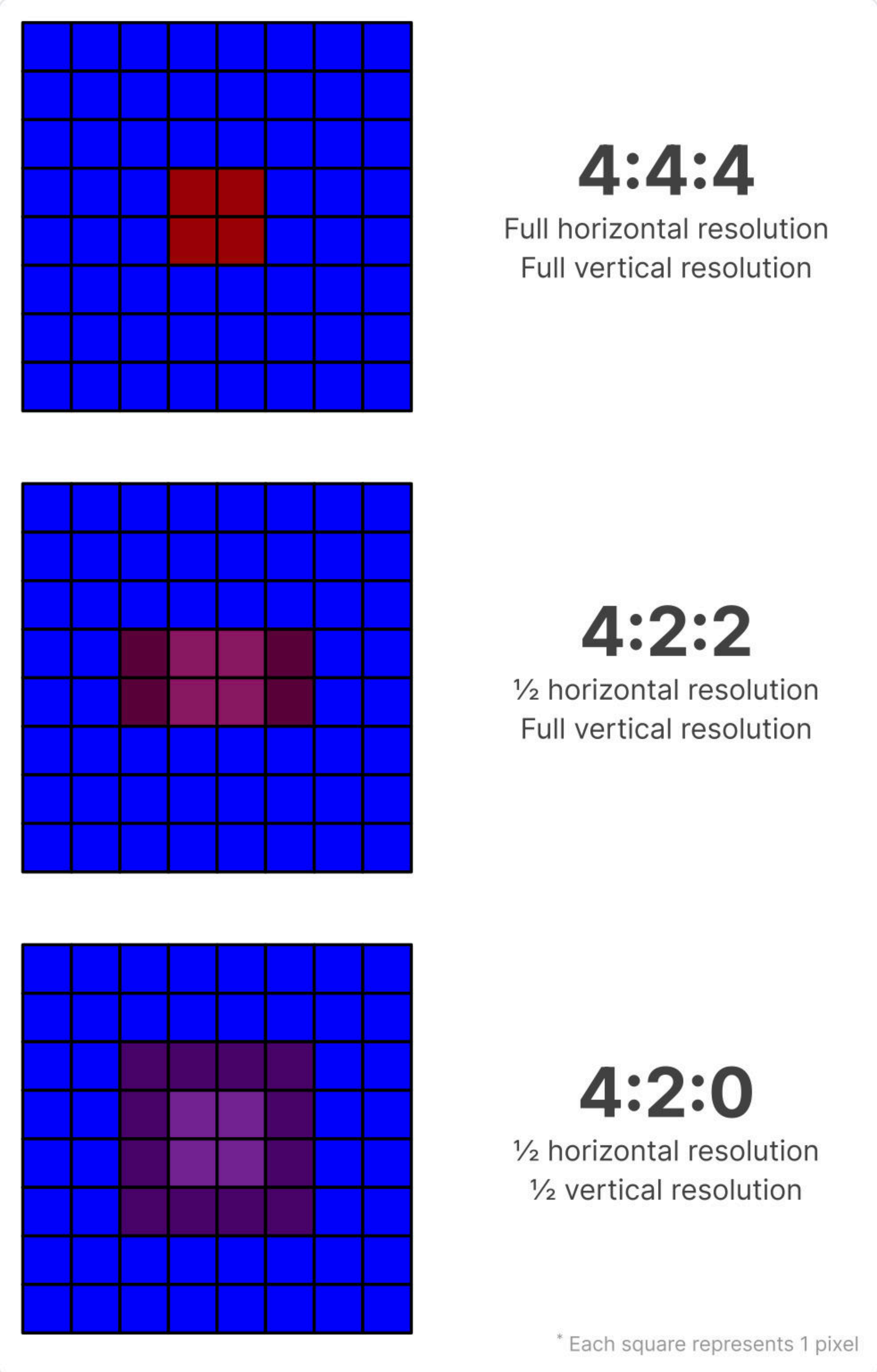
Chroma Subsampling

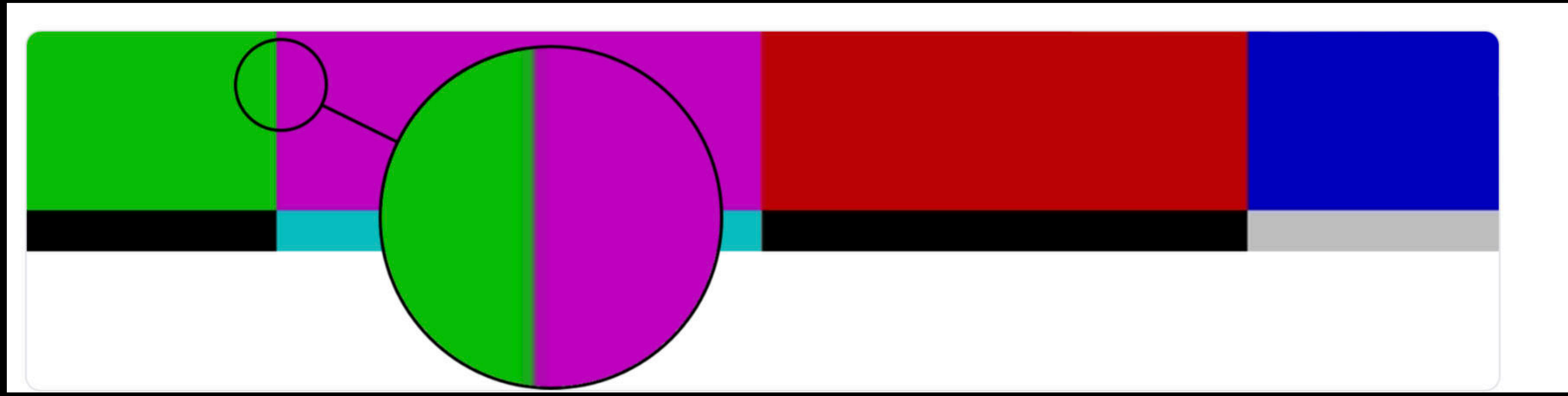
downsample color information in Cb and Cr channels; **Demo**





Chroma Subsampling Artifacts





Ok, but what exactly is RGB ➔ YCbCr transform?

Answer not as simple as it sounds because it depends on color-spaces!

$$
\begin{aligned}
Y' &= 16 + \frac{65.481 \cdot R'_D}{255} + \frac{128.553 \cdot G'_D}{255} + \frac{24.966 \cdot B'_D}{255} \\
C_B &= 128 - \frac{37.797 \cdot R'_D}{255} - \frac{74.203 \cdot G'_D}{255} + \frac{112.0 \cdot B'_D}{255} \\
C_R &= 128 + \frac{112.0 \cdot R'_D}{255} - \frac{93.786 \cdot G'_D}{255} - \frac{18.214 \cdot B'_D}{255}
\end{aligned}
$$

ITU-R BT.601 conversion
used with RGB
(~SDTV) colorspace

$$
\begin{bmatrix} Y' \\ C_B \\ C_R \end{bmatrix} = \begin{bmatrix} 0.2126 & 0.7152 & 0.0722 \\ -0.1146 & -0.3854 & 0.5 \\ 0.5 & -0.4542 & -0.0458 \end{bmatrix} \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix}
$$

$$
\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1.5748 \\ 1 & -0.1873 & -0.4681 \\ 1 & 1.8556 & 0 \end{bmatrix} \begin{bmatrix} Y' \\ C_B \\ C_R \end{bmatrix}
$$

ITU-R BT.709 conversion
used with sRGB
(~HDTV) colorspace

Different Quantization Matrix for Luma and Chroma Components

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

Luma base quantization matrix (quality level 50)

17	18	24	47	99	99	99	99
18	21	26	66	99	99	99	99
24	26	56	99	99	99	99	99
47	66	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99

Chroma base quantization matrix (quality level 50)

Recall: 8 x 8 2D-DCT transforms; lower right represents higher frequency component

Let’s look at a JPEG encoded image

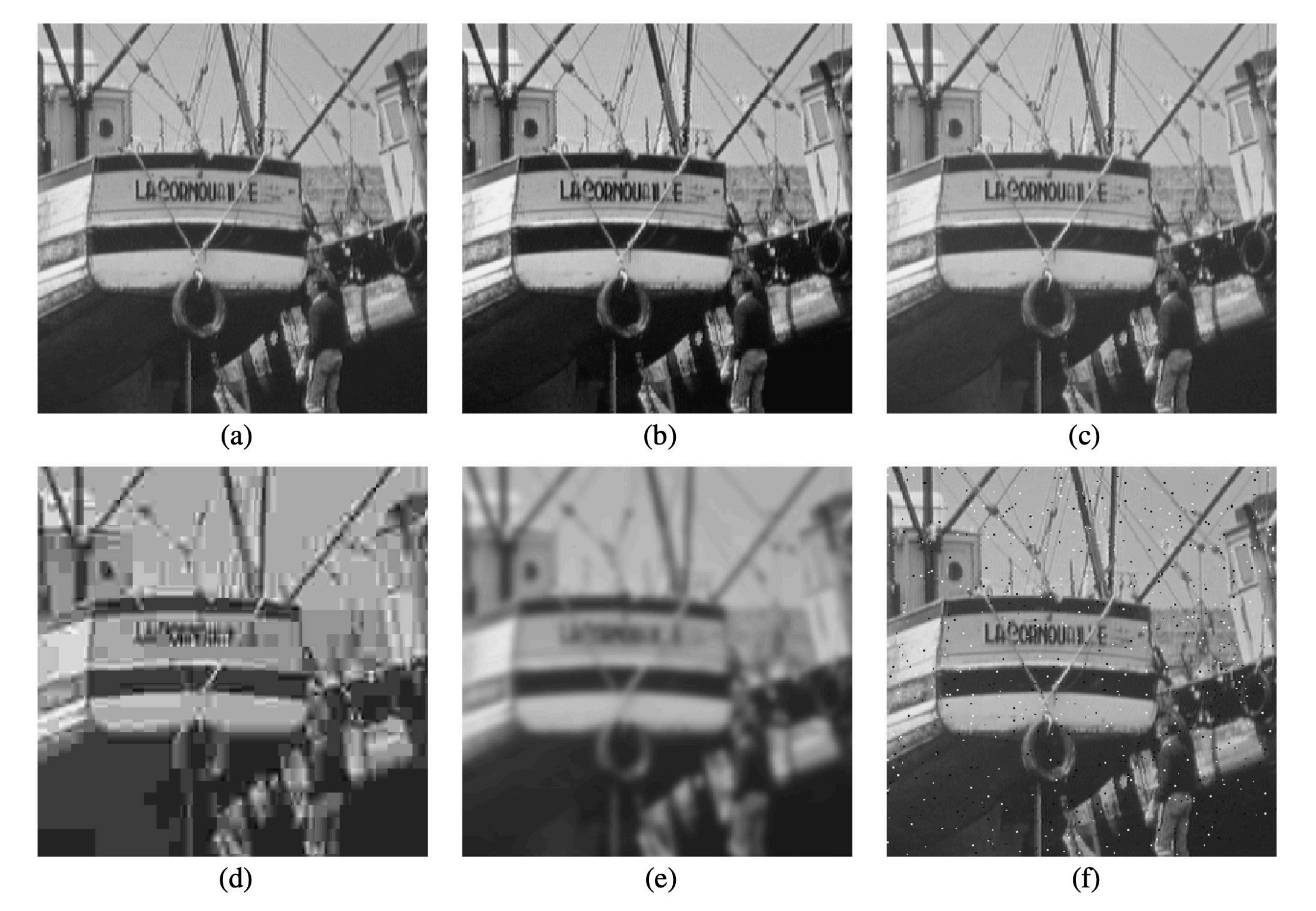
```
~/Downloads  exiftool stanford_logo.jpg
ExifTool Version Number         : 12.51
File Name                       : stanford_logo.jpg
Directory                       : .
File Size                       : 26 kB
File Modification Date/Time     : 2022:11:29 01:02:43-08:00
File Access Date/Time           : 2022:11:29 01:06:41-08:00
File Inode Change Date/Time     : 2022:11:29 01:03:09-08:00
File Permissions                : -rw-r--r--
File Type                       : JPEG
File Type Extension             : jpg
MIME Type                       : image/jpeg
JFIF Version                    : 1.01
Resolution Unit                 : None
X Resolution                    : 72
Y Resolution                    : 72
Exif Byte Order                 : Big-endian (Motorola, MM)
Orientation                     : Horizontal (normal)
Color Space                     : sRGB
Exif Image Width                : 400
Exif Image Height               : 400
Current IPTC Digest             : d41d8cd98f00b204e9800998ecf8427e
IPTC Digest                     : d41d8cd98f00b204e9800998ecf8427e
Image Width                     : 400
Image Height                    : 400
Encoding Process                : Baseline DCT, Huffman coding
Bits Per Sample                 : 8
Color Components                : 3
Y Cb Cr Sub Sampling            : YCbCr4:2:0 (2 2)
Image Size                      : 400x400
Megapixels                      : 0.160
```

Part 2:
Human Vision and it's implications on *Distortion Metric*

MSE as a distortion metric is inadequate

We will look into three class of ***perceptual metrics*** as distortion:

- modeling low-level human vision features e.g. SSIM, MS-SSIM, VIF
- learnt ML models as perceptual metrics e.g. LPIPS
- combining metrics using supervised human subjective data e.g. VMAF, DISTS

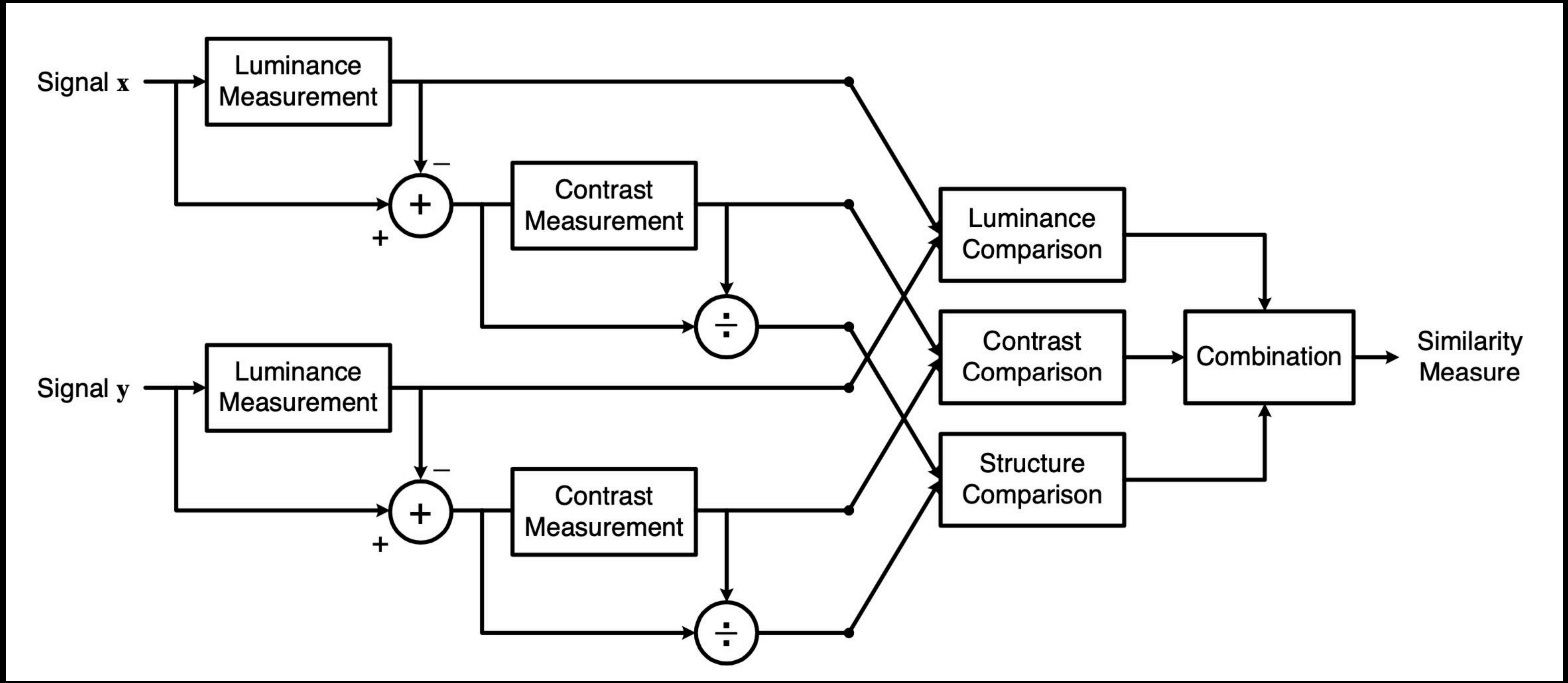


SSIM

structural similarity
modeling low-level human vision features

Uses 3 key features to compare two images:

luminance, contrast, structure



$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}.$$

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^\alpha \cdot [c(\mathbf{x}, \mathbf{y})]^\beta \cdot [s(\mathbf{x}, \mathbf{y})]^\gamma$$

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}.$$

SSIM

structural similarity

modeling low-level human vision features

But apply them over local patches instead of globally

- spatial non-stationarity of image features

- spatial non-stationarity of image distortions

- foveation

How:

calculate locally and then take mean

uses gaussian kernel for smoothening to model ***foveation***!

$$l(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

$$c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

$$s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}.$$

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^\alpha \cdot [c(\mathbf{x},\mathbf{y})]^\beta \cdot [s(\mathbf{x},\mathbf{y})]^\gamma$$

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}.$$

$$\mu_x = \sum_{i=1}^{N} w_i x_i$$

$$\sigma_x = \left(\sum_{i=1}^{N} w_i (x_i - \mu_x)^2\right)^{\frac{1}{2}}$$

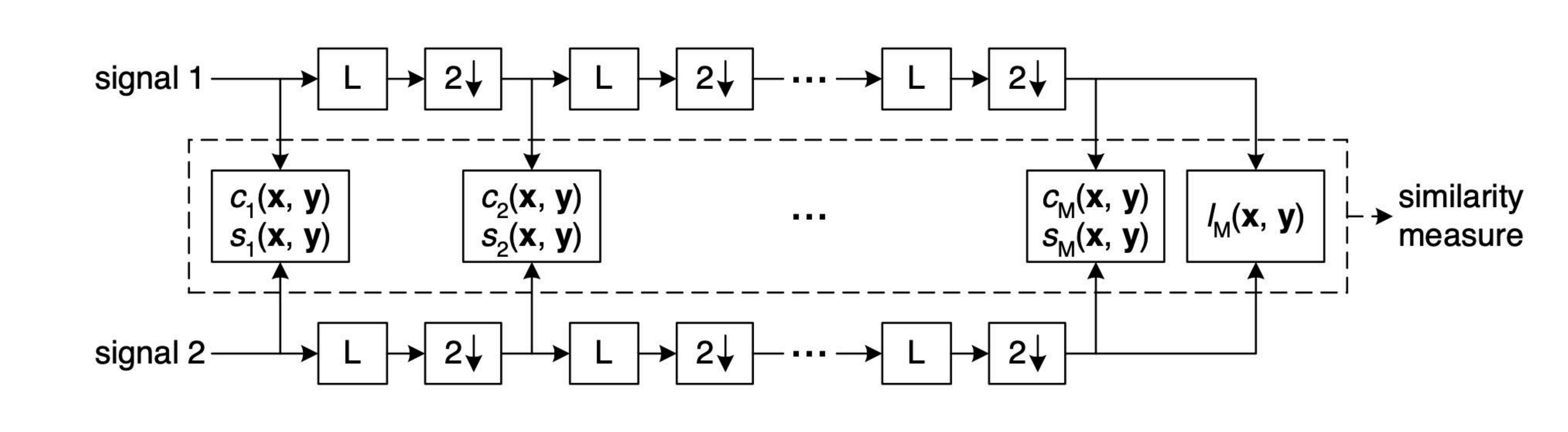
$$\sigma_{xy} = \sum_{i=1}^{N} w_i (x_i - \mu_x)(y_i - \mu_y).$$

$$\mathrm{MSSIM}(\mathbf{X},\mathbf{Y}) = \frac{1}{M}\sum_{j=1}^{M} \mathrm{SSIM}(\mathbf{x}_j,\mathbf{y}_j)$$

MS-SSIM

multi-scale structural similarity

SSIM + better accounting for spatial frequency

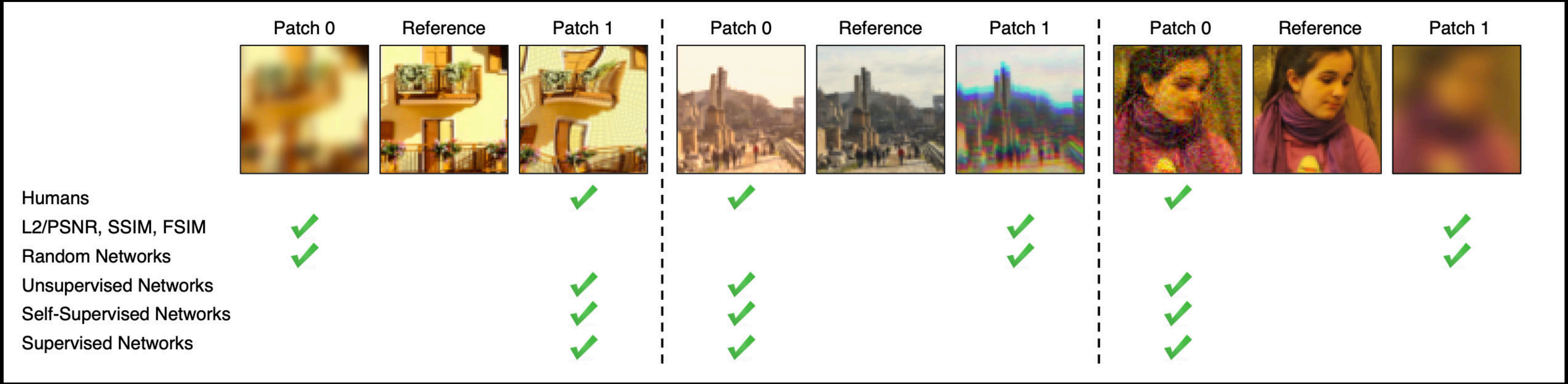


$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^{\alpha} \cdot [c(\mathbf{x},\mathbf{y})]^{\beta} \cdot [s(\mathbf{x},\mathbf{y})]^{\gamma} \longrightarrow \mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l_M(\mathbf{x},\mathbf{y})]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(\mathbf{x},\mathbf{y})]^{\beta_j} [s_j(\mathbf{x},\mathbf{y})]^{\gamma_j}$$

LPIPS

Learned Perceptual Image Patch Similarity

The Unreasonable Effectiveness of Deep Features as a Perceptual Metric



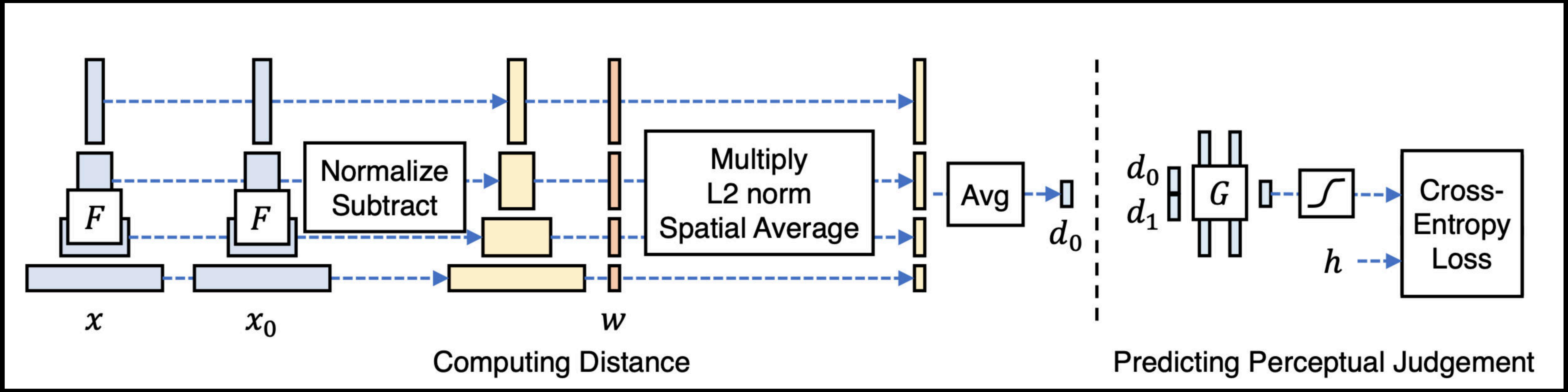
"Our results suggest that perceptual similarity is an emergent property shared across deep visual representations."

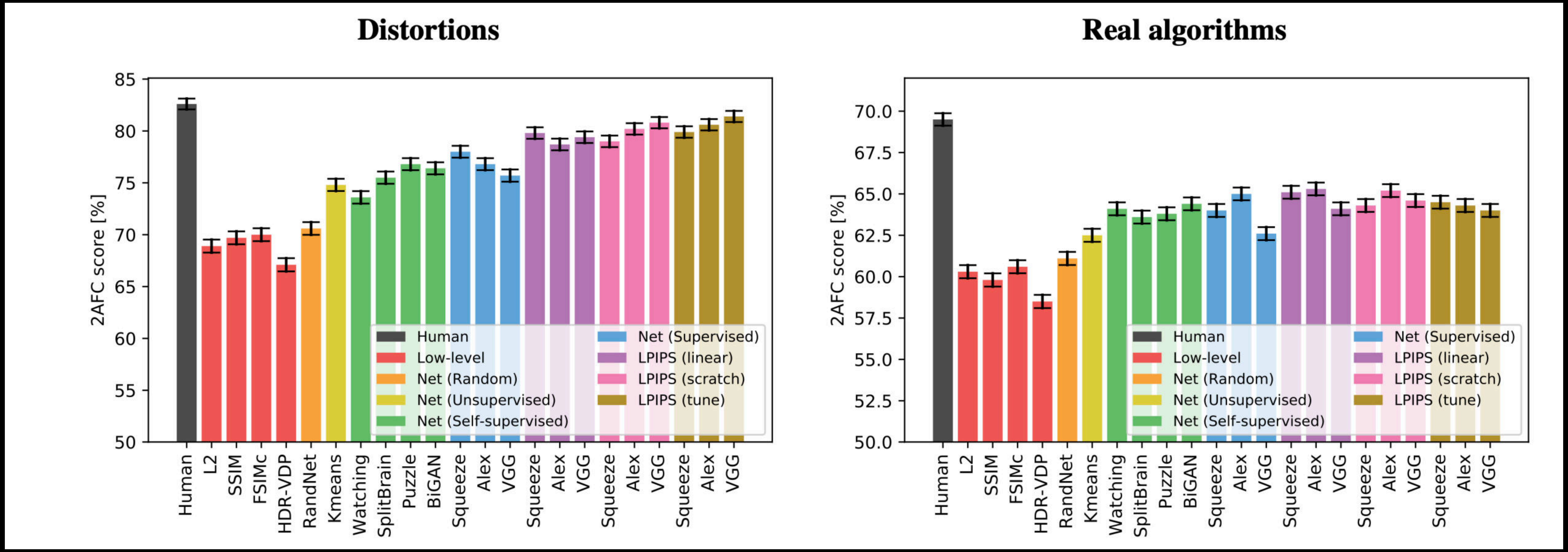
Use ***deep embeddings as a feature space*** for learning perceptual metric.

LPIPS

Learned Perceptual Image Patch Similarity

The Unreasonable Effectiveness of Deep Features as a Perceptual Metric

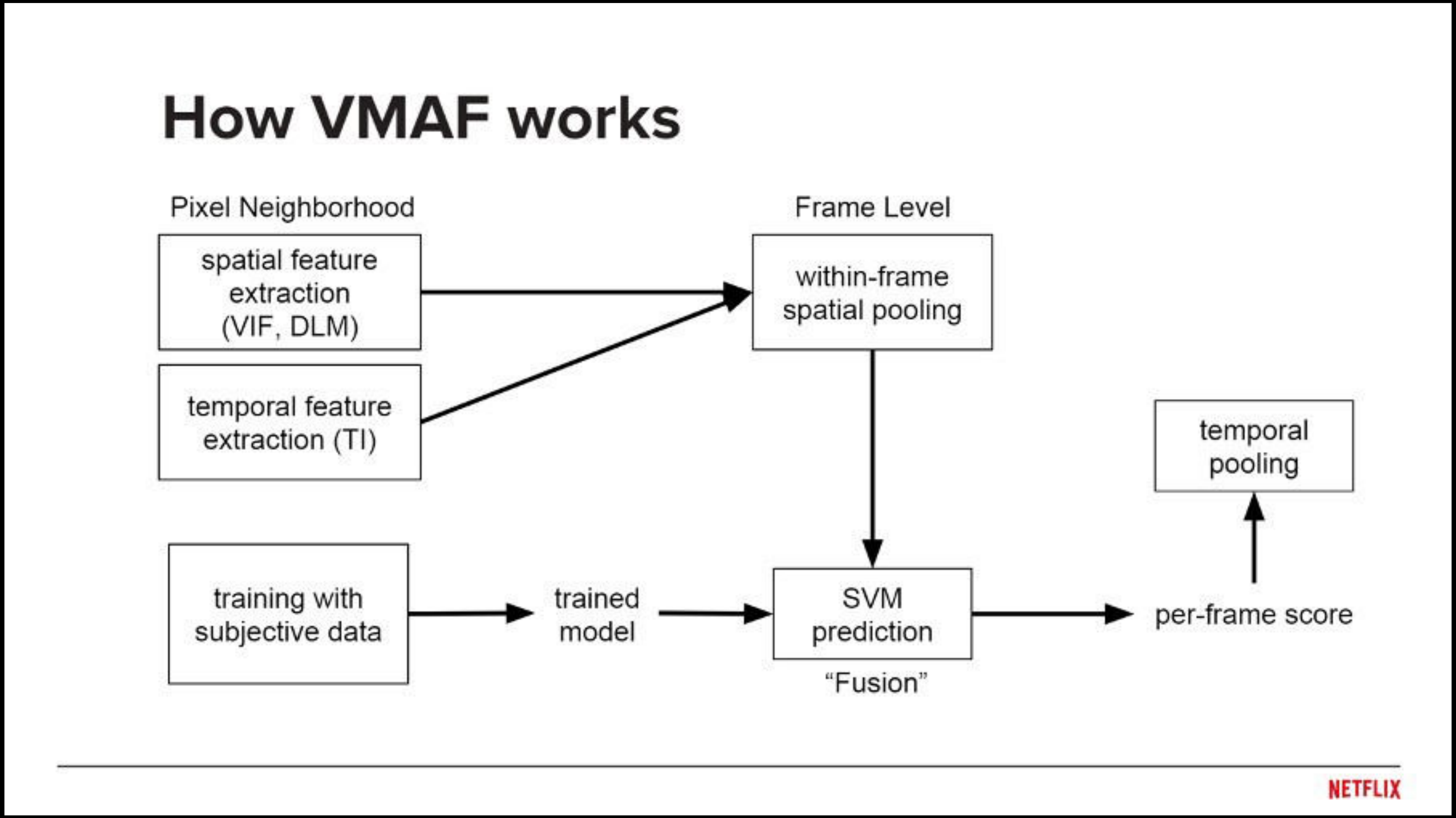




VMAF

Video Multimethod Assessment Fusion

supervised learning over existing perceptual metrics



Part 3: Rate-Distortion-Perception tradeoff

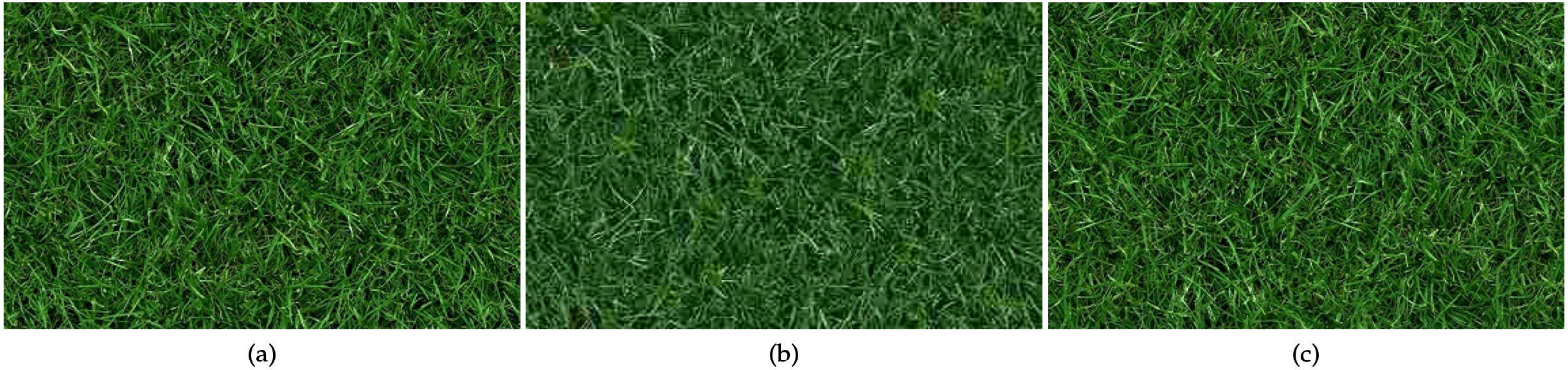
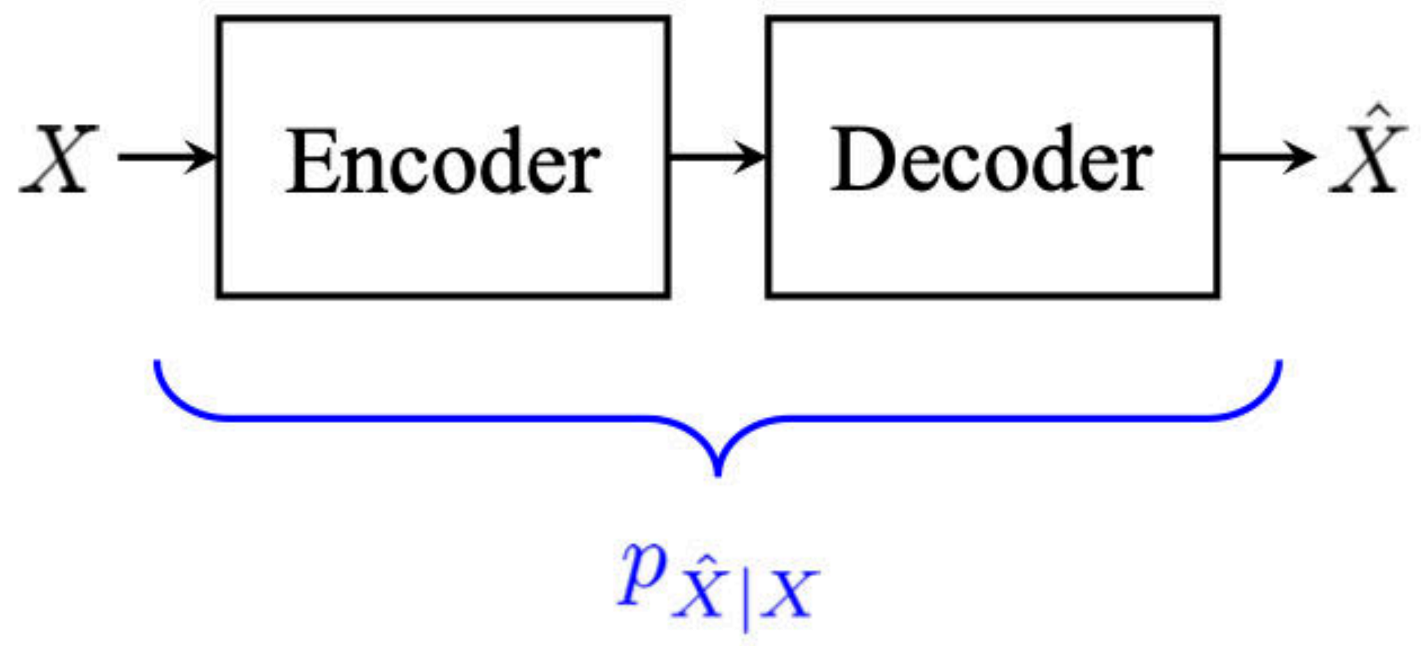


Fig. 1. Existing full-reference IQA models are overly sensitive to point-by-point deviations between images of the same texture. **(a)** A grass image and **(b)** the same image, distorted by JPEG compression. **(c)** Resampling of the same grass as in (a). Popular IQA measures, including PSNR, SSIM [3], FSIM [11], VIF [4], GMSD [12], DeepIQA [13], PieAPP [8], and LPIPS [7], predict that image (b) has a better perceived quality than image (c), which is in disagreement with human rating. In contrast, the proposed DISTS model makes the correct prediction. (Zoom in to improve visibility of details).

Theory

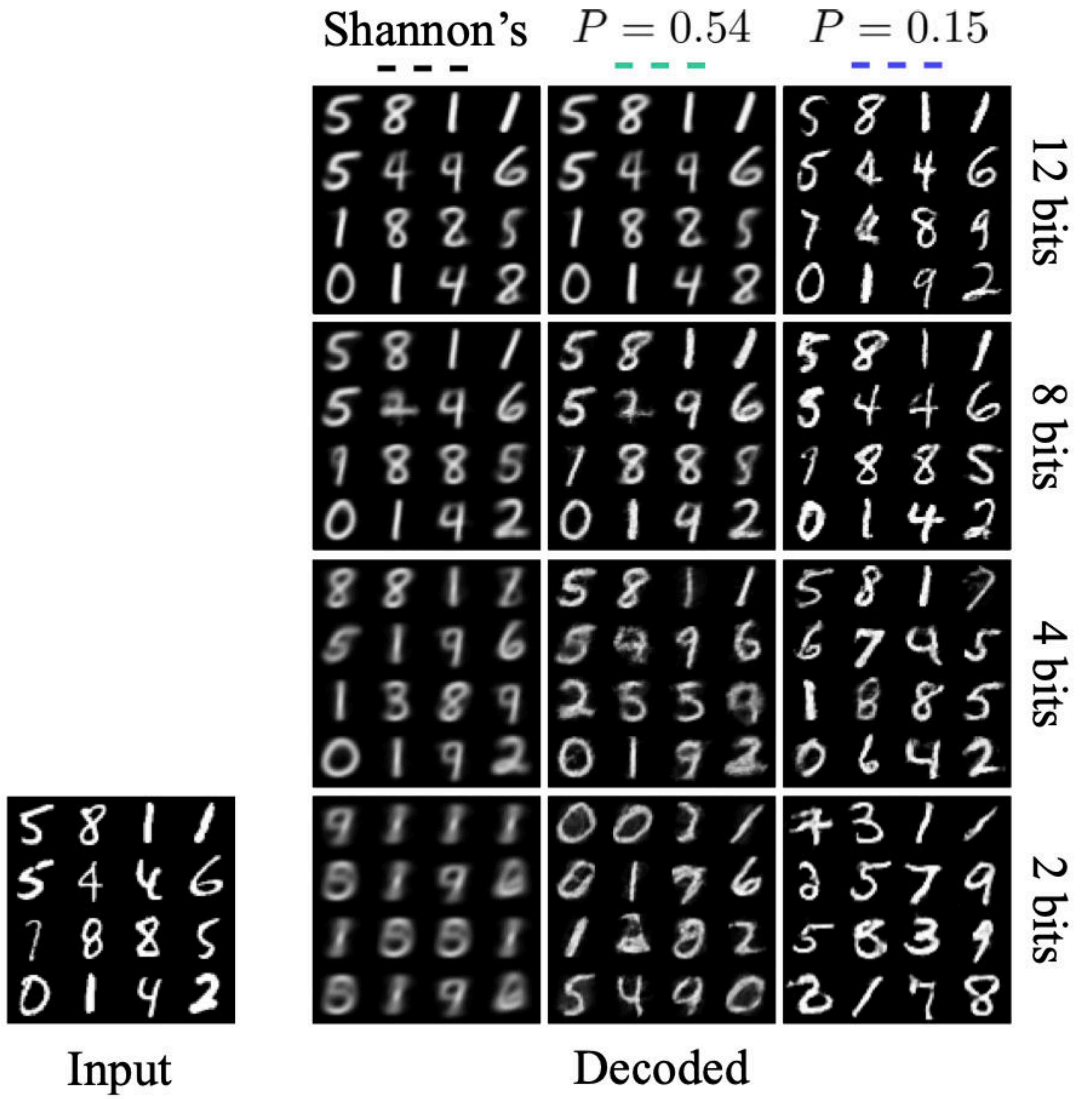


Rate:
$I(X, \hat{X})$

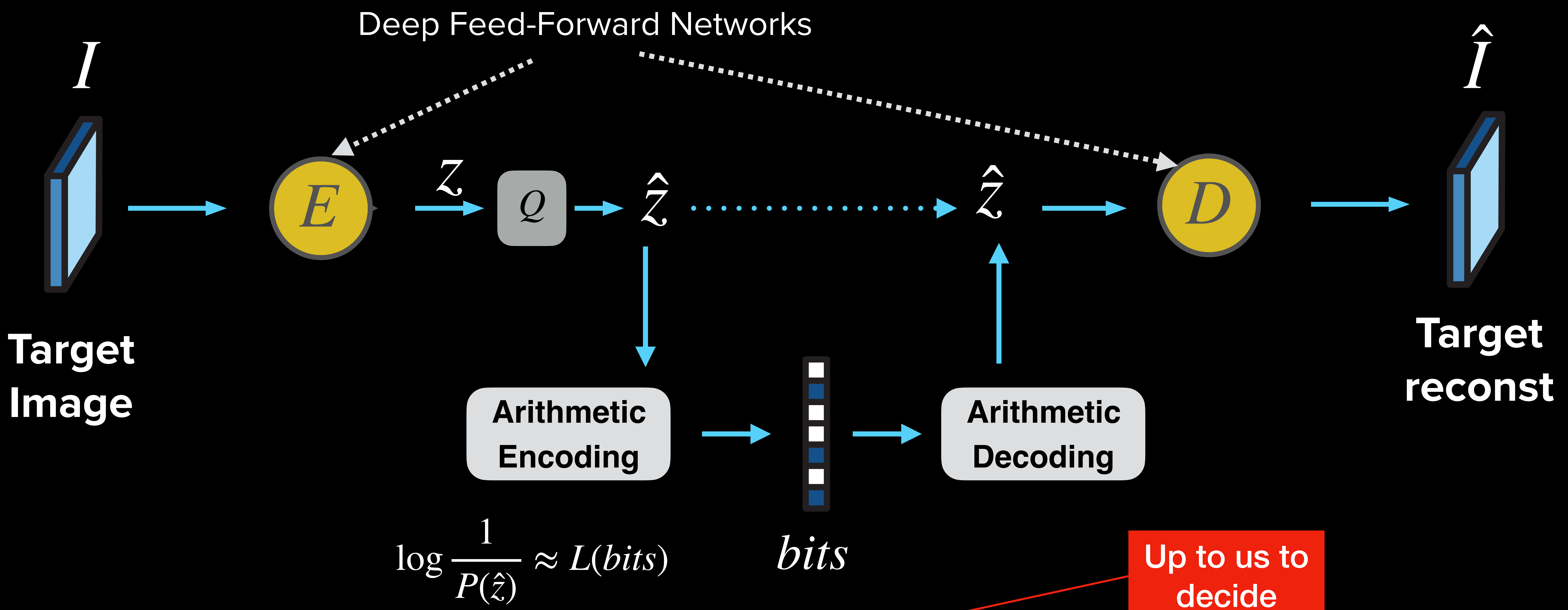
Distortion:
$\mathbb{E}[\Delta(X, \hat{X})]$

Perception:
$d(p_X, p_{\hat{X}})$

Figure 2. **Lossy compression.** A source signal $X \sim p_X$ is mapped into a coded sequence by an encoder and back into an estimated signal $\hat{X}$ by the decoder. Three desired properties are: (i) the coded sequence be compact (low bit rate); (ii) the reconstruction $\hat{X}$ be similar to the source X on average (low distortion); (iii) the distribution $p_{\hat{X}}$ be similar to p_X, so that decoded signals are perceived as genuine source signals (good perceptual quality).



RDP in practice
Learnt Image Compression



Up to us to decide

Loss Function = $L(bits) + \lambda \boxed{d(I, \hat{I})} \approx \log \frac{1}{P(\hat{z})} + \lambda d(I, \hat{I})$

RDP in practice
Learnt Image Compression

Idea 1:
use a perceptual metric such as SSIM as distortion

$$\textbf{Loss Function} = L(bits) + \lambda \boxed{d(I, \hat{I})}$$

$$\approx L(bits) + \lambda \; SSIM(I, \hat{I})$$

RDP in practice
Learnt Image Compression

Idea 2:
add a weighted MSE loss with some perceptual loss like LPIPS

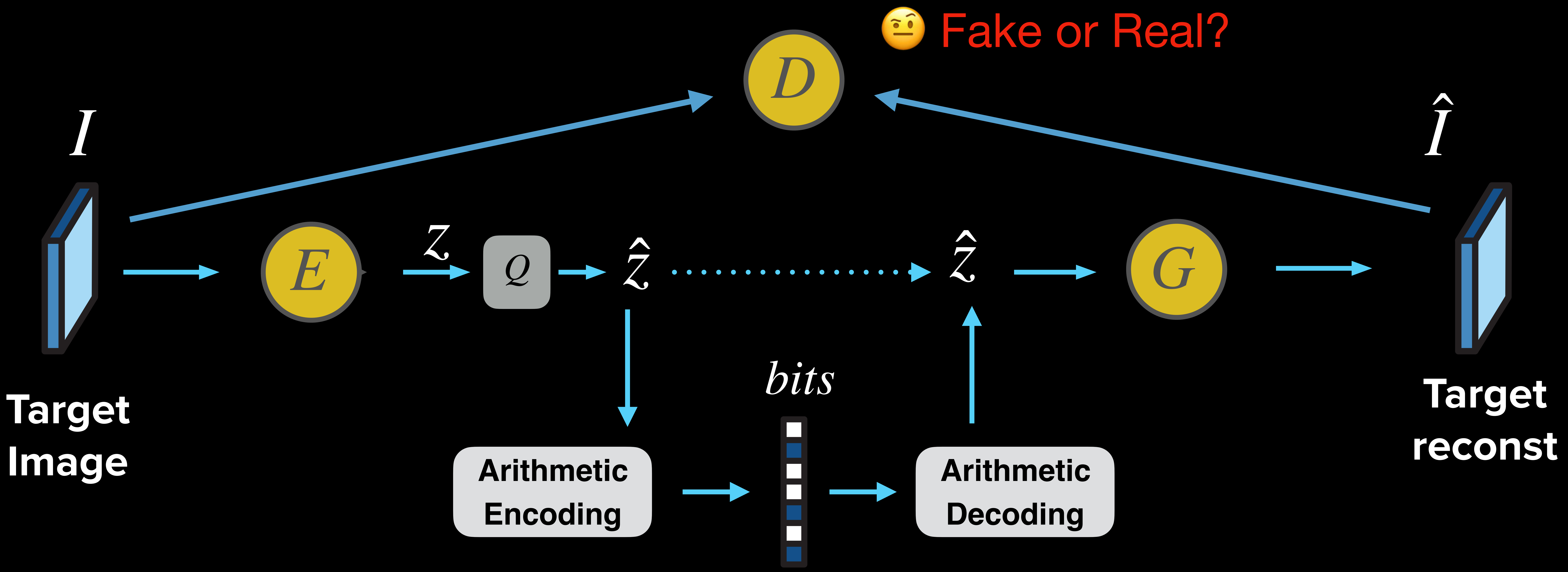
$$\textbf{Loss Function} = L(bits) + \lambda \boxed{d(I, \hat{I})}$$

$$= L(bits) + \lambda_1 \, \mathbb{E}\left(\Delta(I, \hat{I})\right) + \lambda_2 \, d\left(p_I, p_{\hat{I}}\right)$$

$$\approx L(bits) + \lambda_1 \, MAE(I, \hat{I}) + \lambda_2 \, LPIPS(I, \hat{I})$$

RDP in practice
Learnt Image Compression

Idea 3:
Generative Adversarial Network like framework to ensure reconstruction and source have similar distribution



HiFiC

High-Fidelity Generative Image Compression (used conditional GANs + LPIPS in loss)

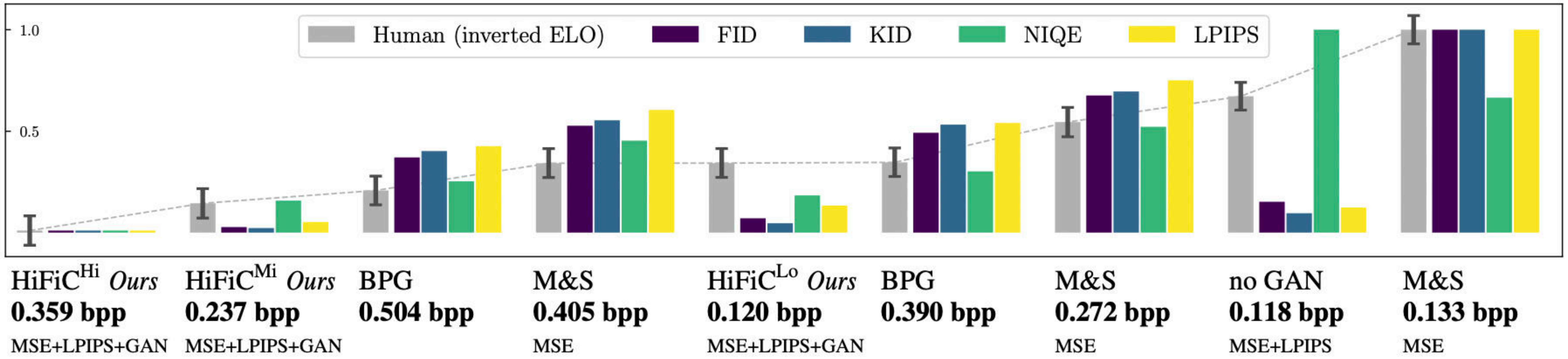
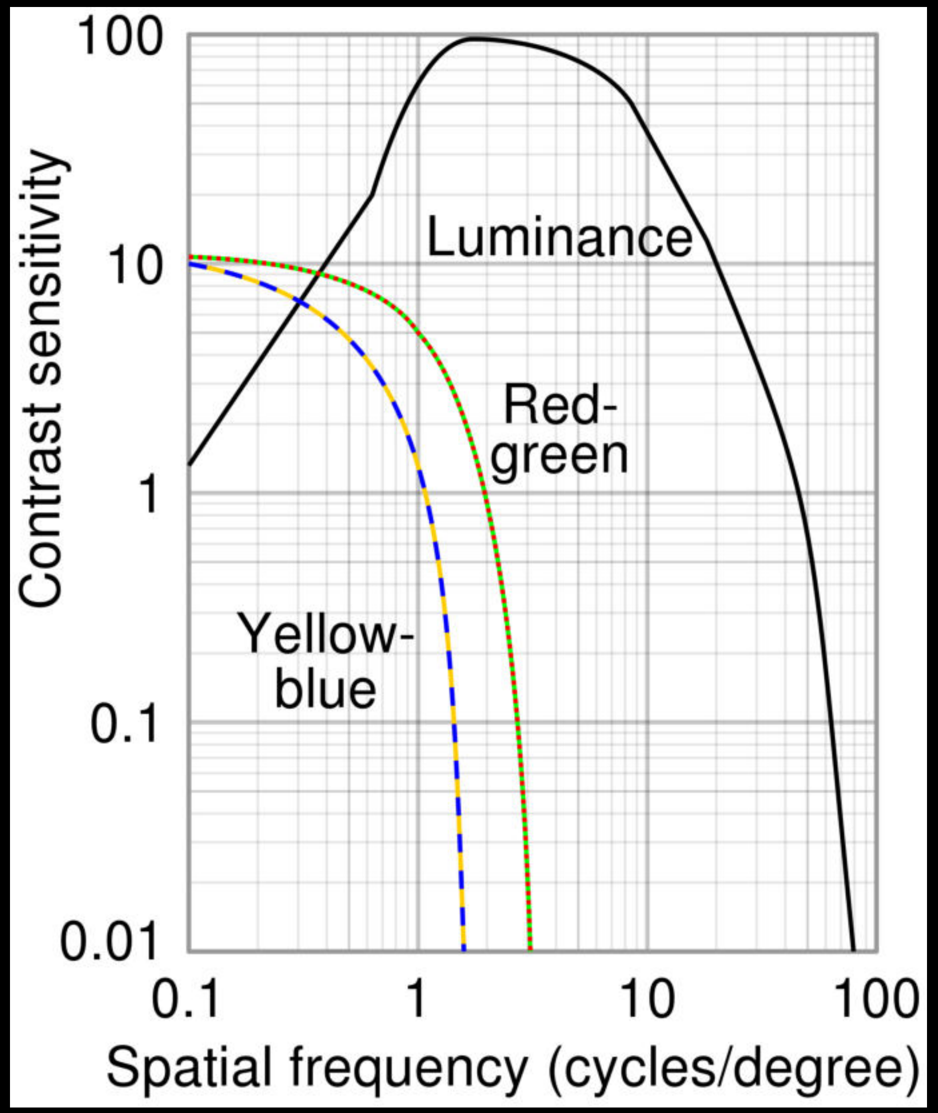
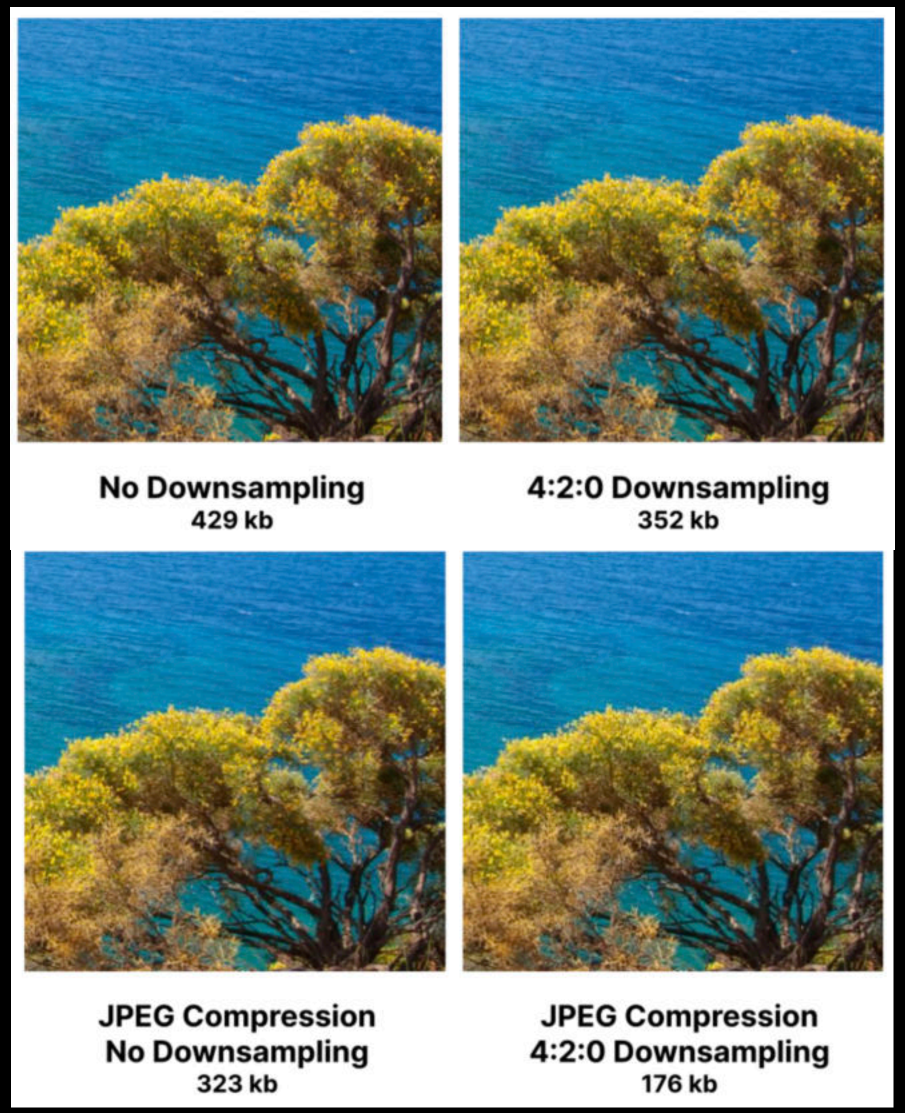
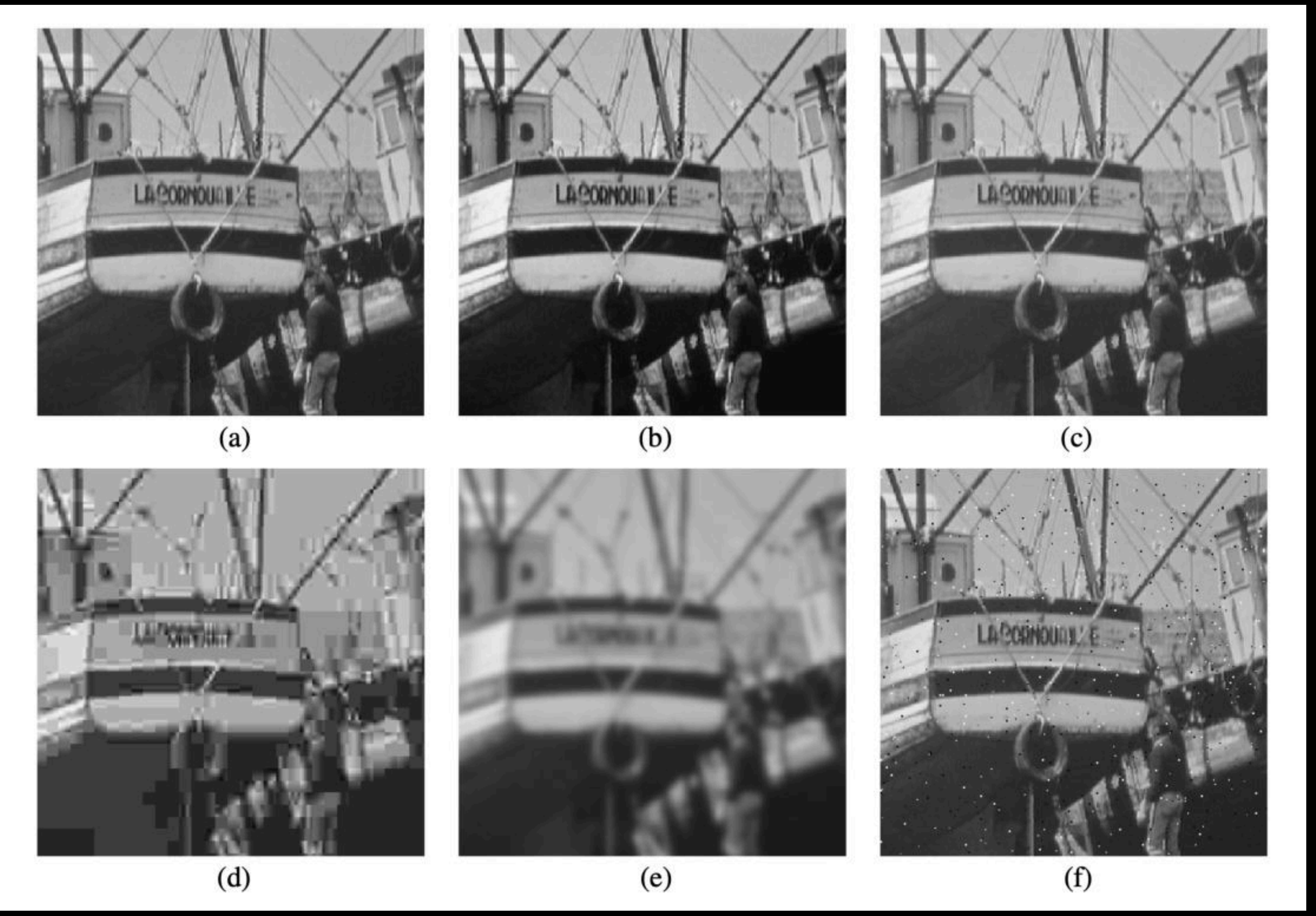


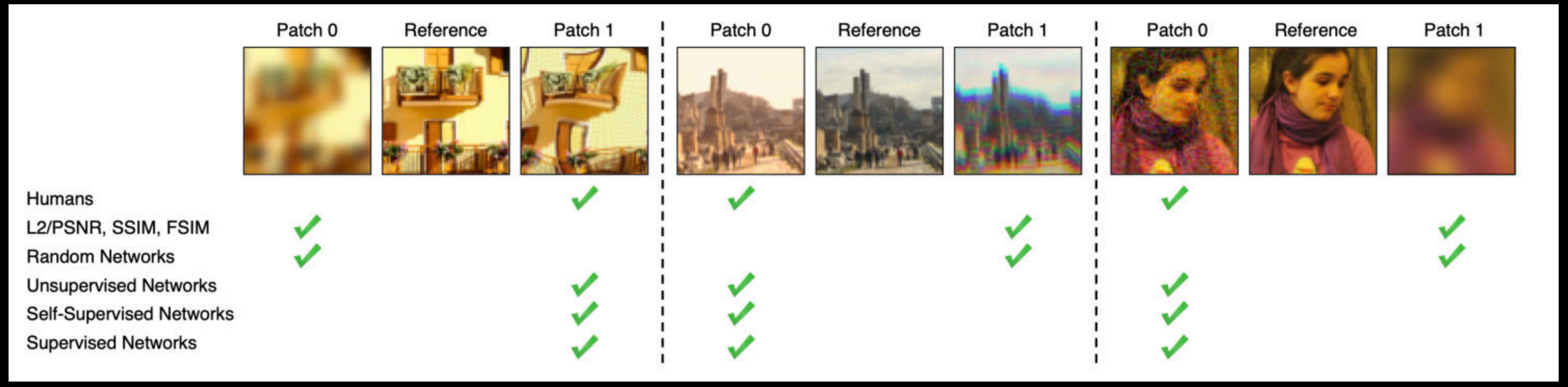
Figure 3: Normalized scores for the user study, compared to perceptual metrics. We invert human scores such that **lower is better** for all. Below each method, we show *average* bpp, and for learned methods we show the loss components. "no GAN" is our baseline, using the same architecture and distortion d as *HiFiC (Ours)*, but no GAN. "M&S" is the *Mean & Scale Hyperprior* MSE-optimized baseline. The study shows that training with a GAN yields reconstructions that outperform BPG at practical bitrates, for high-resolution images. Our model at 0.237bpp is preferred to BPG even if BPG uses 2.1× the bitrate, and to MSE optimized models even if they use 1.7× the bitrate.





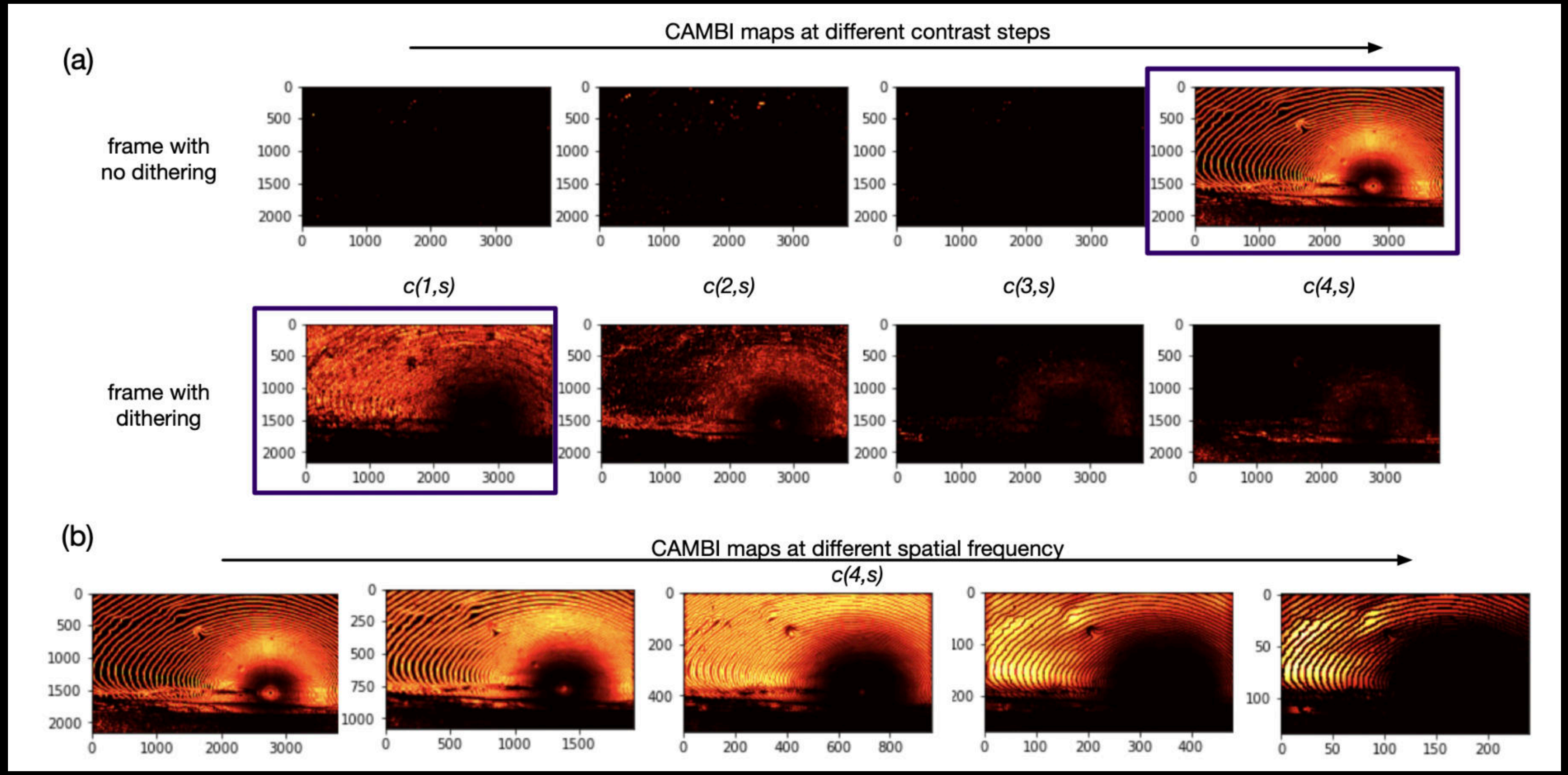


Human Perceptual Properties play a key role in multimedia compression



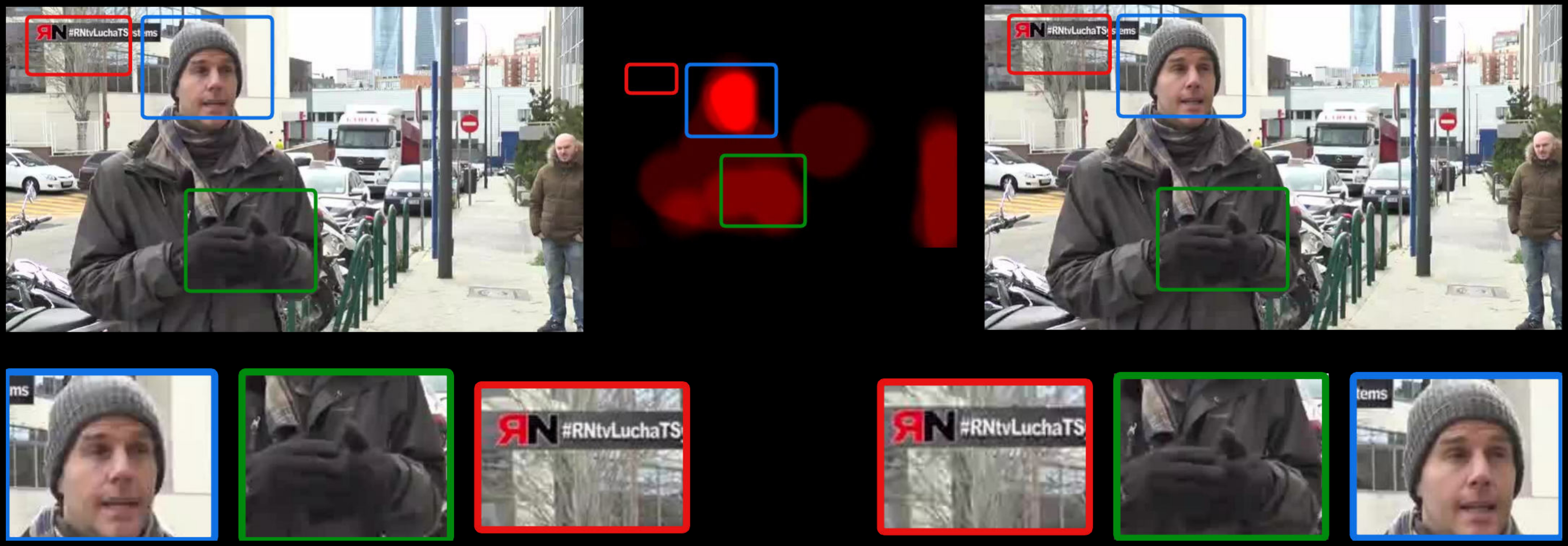
Lots of Open-Research Problems!

1. Detection and avoidance of visual artifacts using first-principles



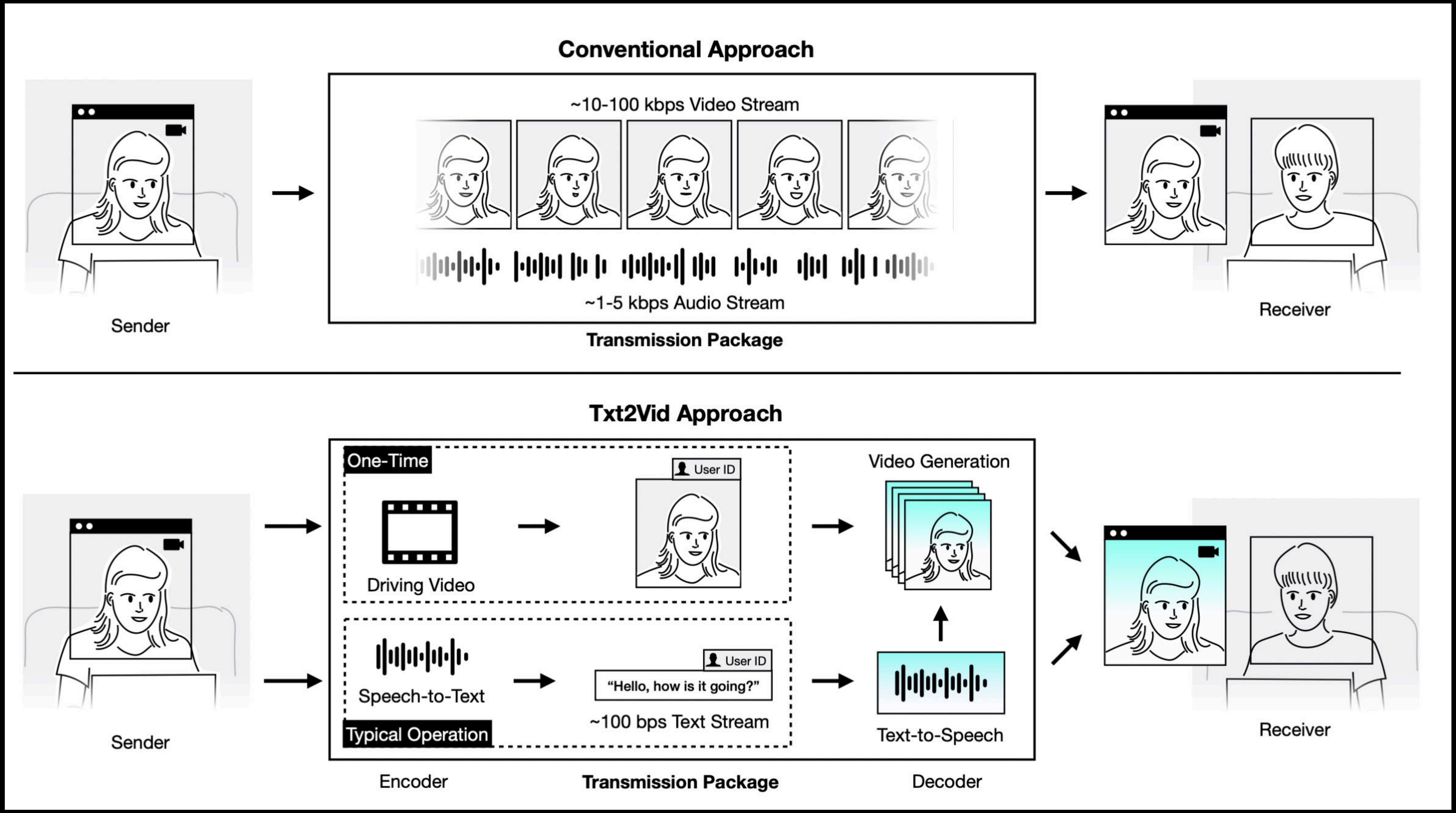
Lots of Open-Research Problems!

2. Automatic identification of visually-salient regions



Lots of Open-Research Problems!

3. Utilizing human-priors for ultra-low bitrate compression



Lots of Open-Research Problems!

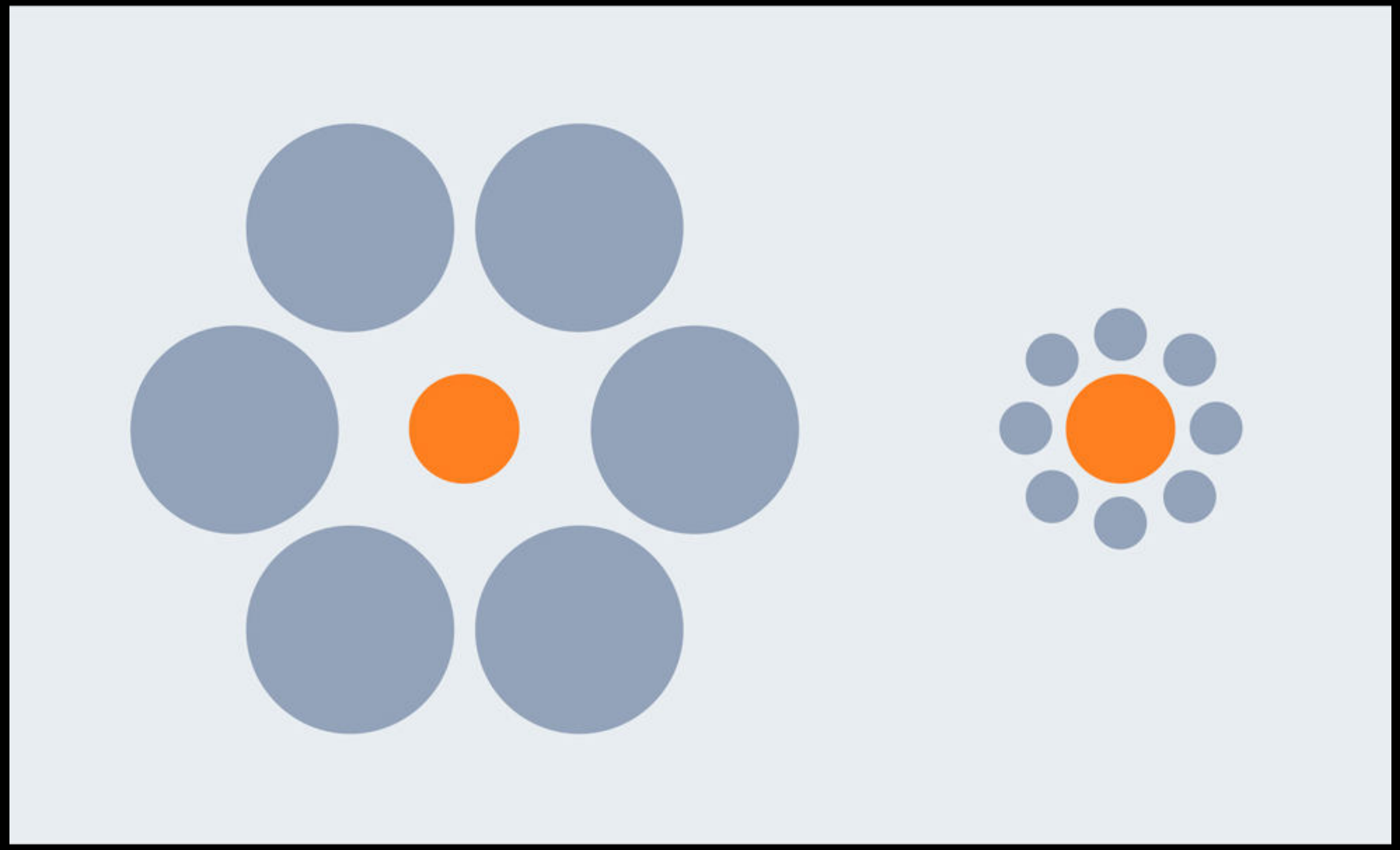
4. Reproducible perceptual study design

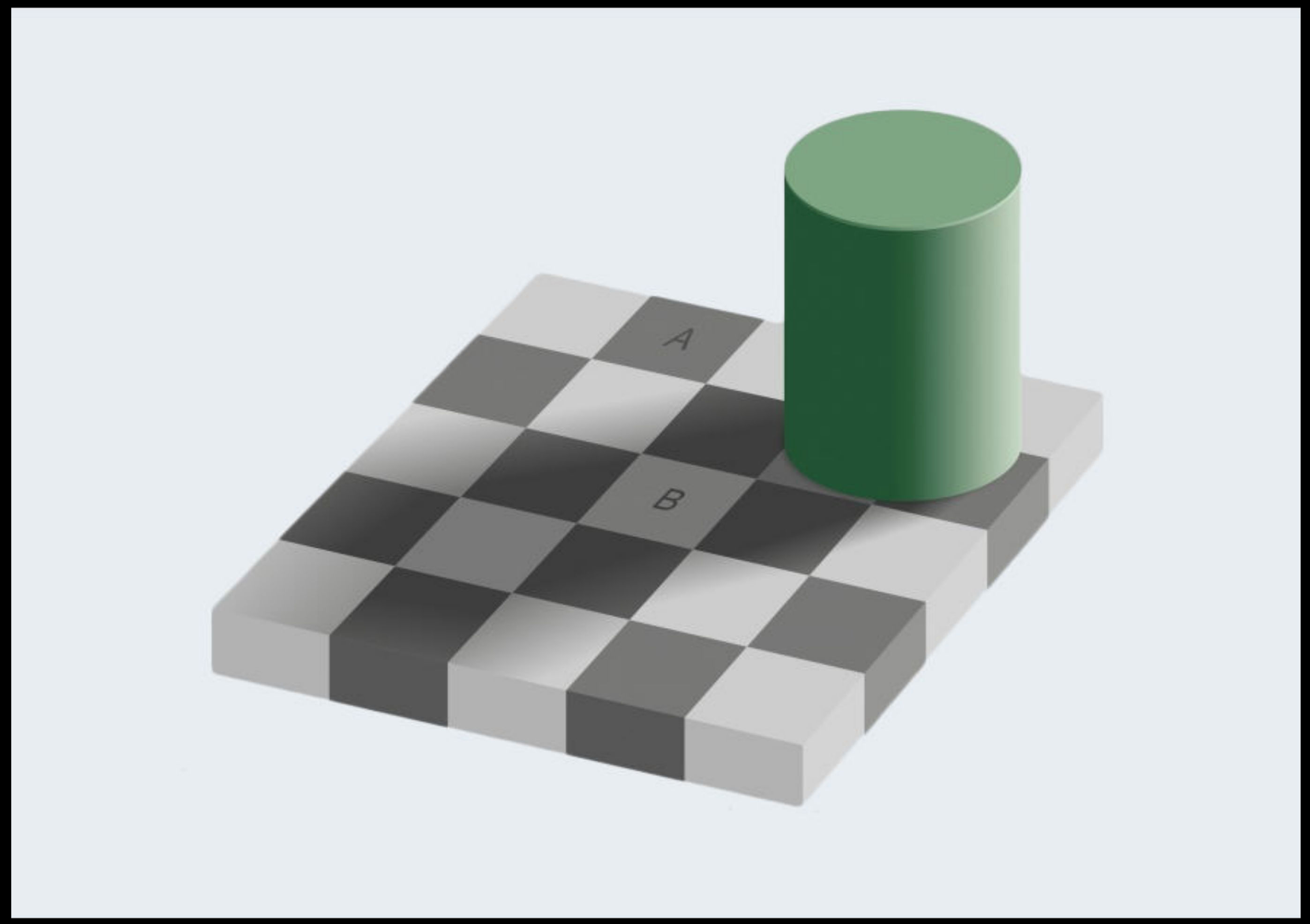
5. Better theoretical understanding of Rate-Distortion-Perception formulation

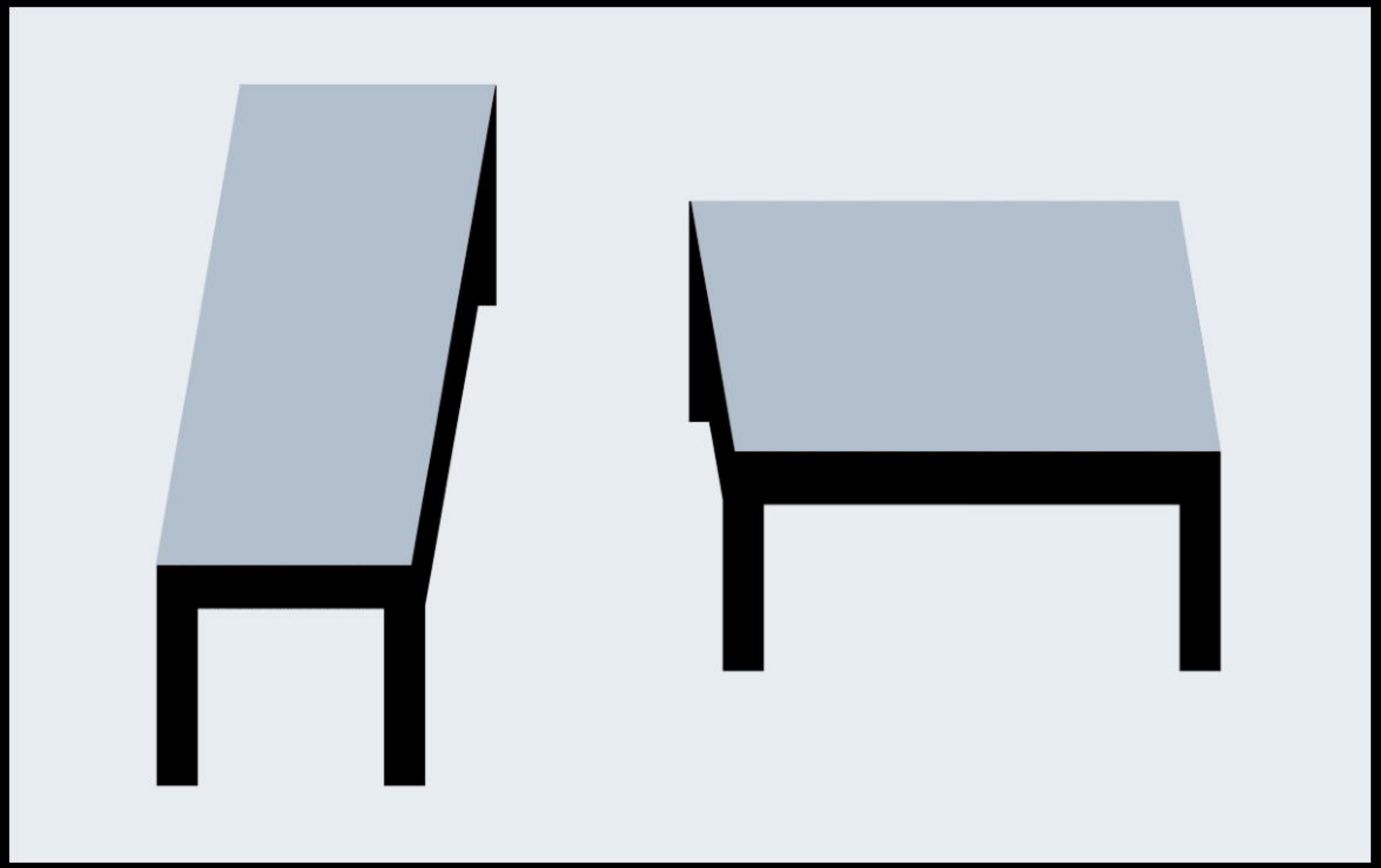
6. Incorporating perceptual optimizations in traditional and learnt codecs

...

Hard interesting problems because a lot of not-so-well-defined variables involved such as human perception, image representation, display properties, viewing conditions, resolutions, human-variance, etc.







“Perception is an ***interpretation*** of the retinal image, not a description”

Foundations of Vision, Brian A. Wandell