SMART: Screen-based Gesture Recognition on Commodity Mobile Devices

Zimo Liao[1], Zhicheng Luo[2], Qianyi Huang[2,5], Linfeng Zhang[3],
Fan Wu[1], Qian Zhang[4], Yi Wang[2,5]

Shanghai Jiao Tong University[1]
Southern University of Science and Technology[2] Tsinghua University[3]
Hong Kong University of Science and Technology[4] Peng Cheng Laboratory[5]

MobiCom 2021

In-air gesture control is natural and contactless

Gesture control via hardware on mobile devices

Image

[CVPR’13,UIST’14]

Acoustic

[CHI’12,MobiCom’16]

Wi-Fi

[Mobicom’15,Ubicomp’16]

Privacy concerns

Background noise

Gesture control via hardware modification

Huawei
Mate40

depth camera

Google
Pixel 4

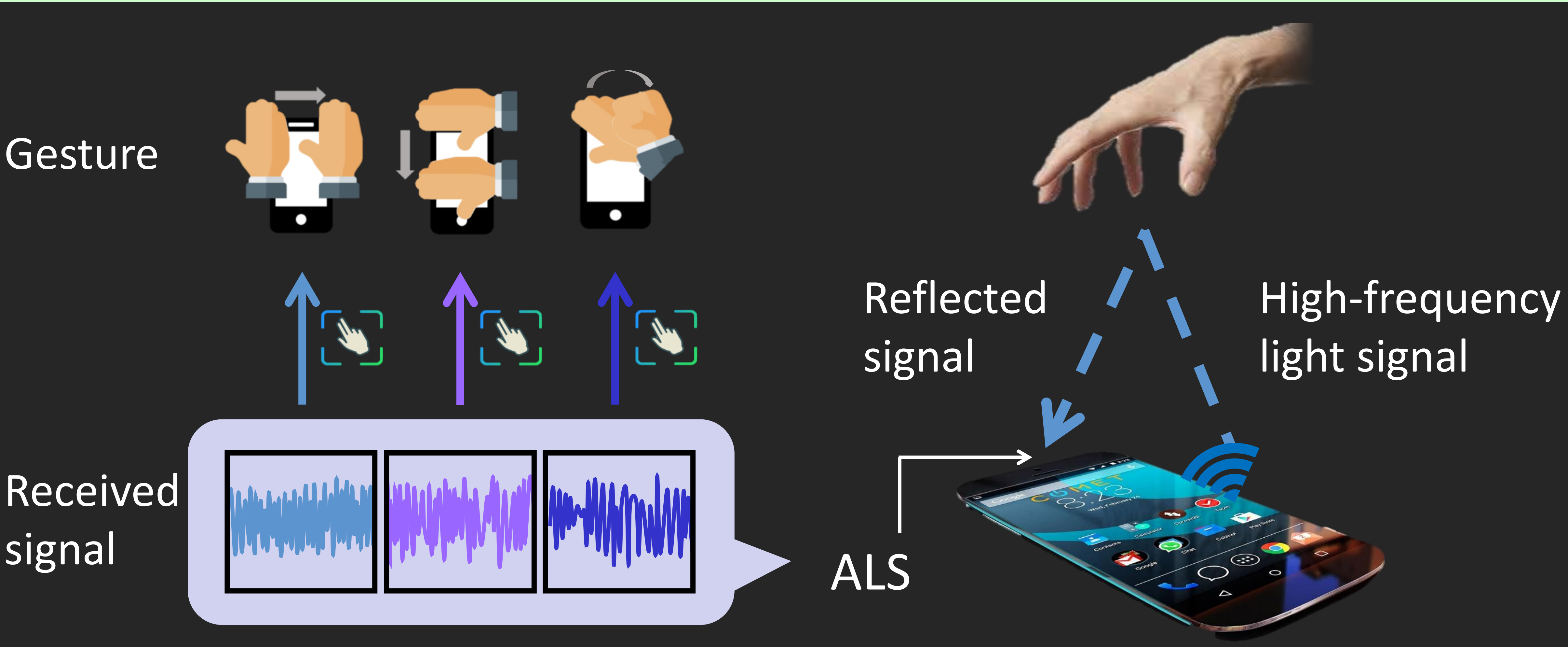
mmWave radar

Can we support in-air gesture recognition on legacy devices ***without hardware modification****?*

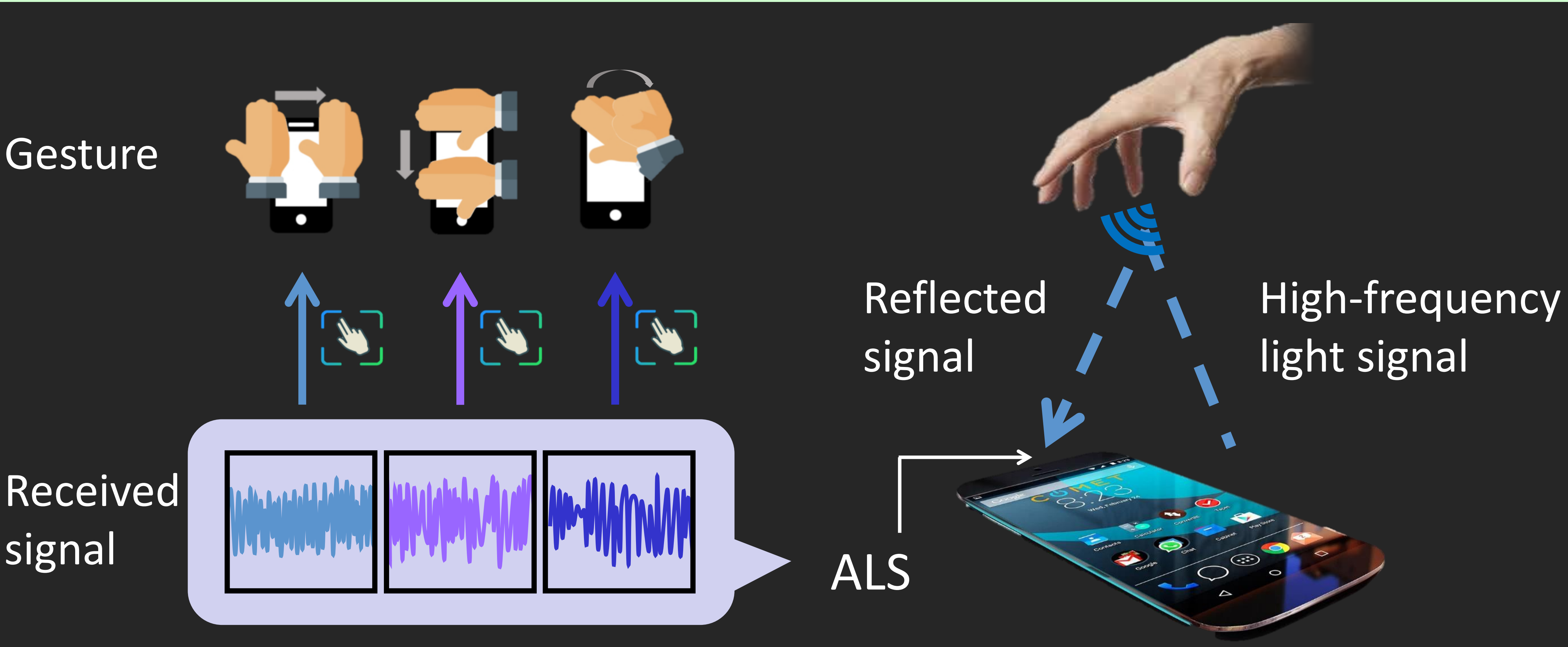
*Can we support in-air gesture recognition on legacy devices **without hardware modification**?*

SMART: A gesture recognition system leveraging **the screen and ambient light sensor(ALS)**.

SMART: A gesture recognition system leveraging **the screen and ambient light sensor(ALS)**.



SMART: A gesture recognition system leveraging **the screen and ambient light sensor(ALS)**.



Gesture

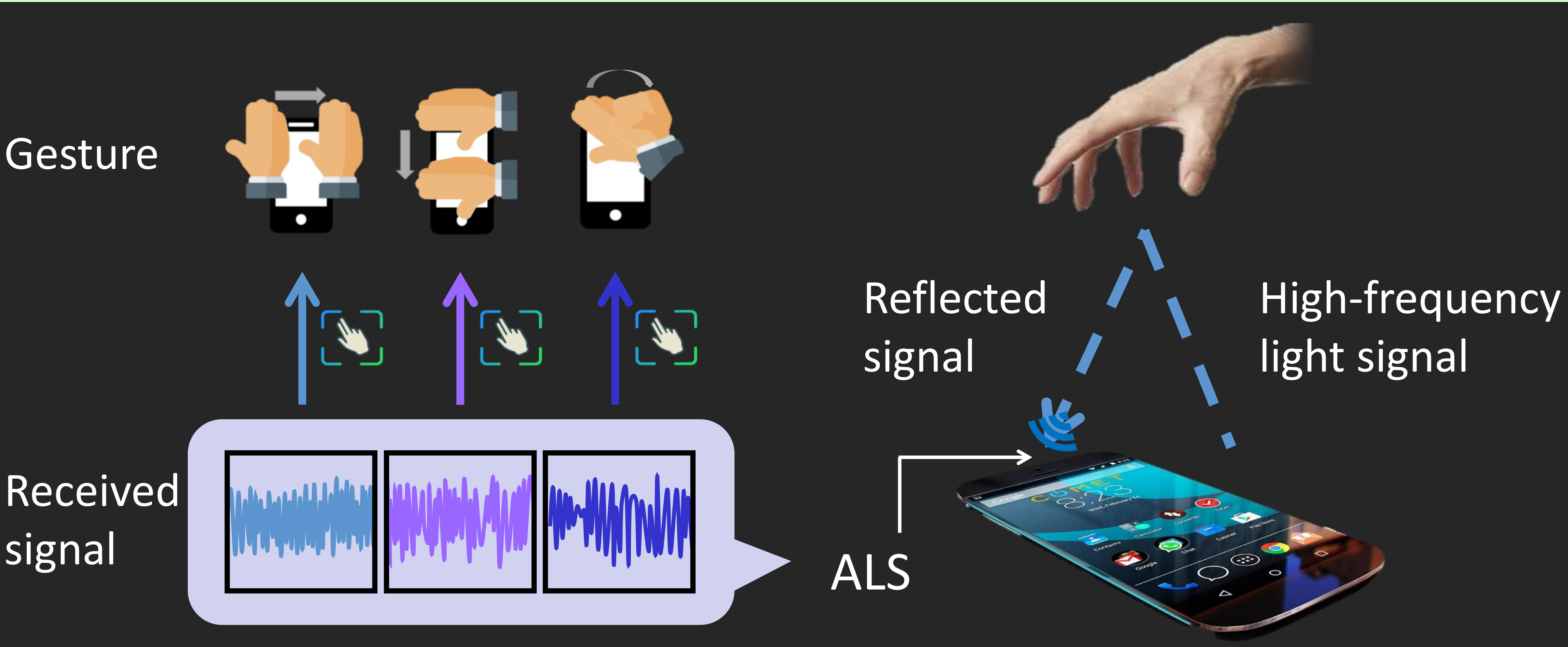
Received signal

Reflected signal

High-frequency light signal

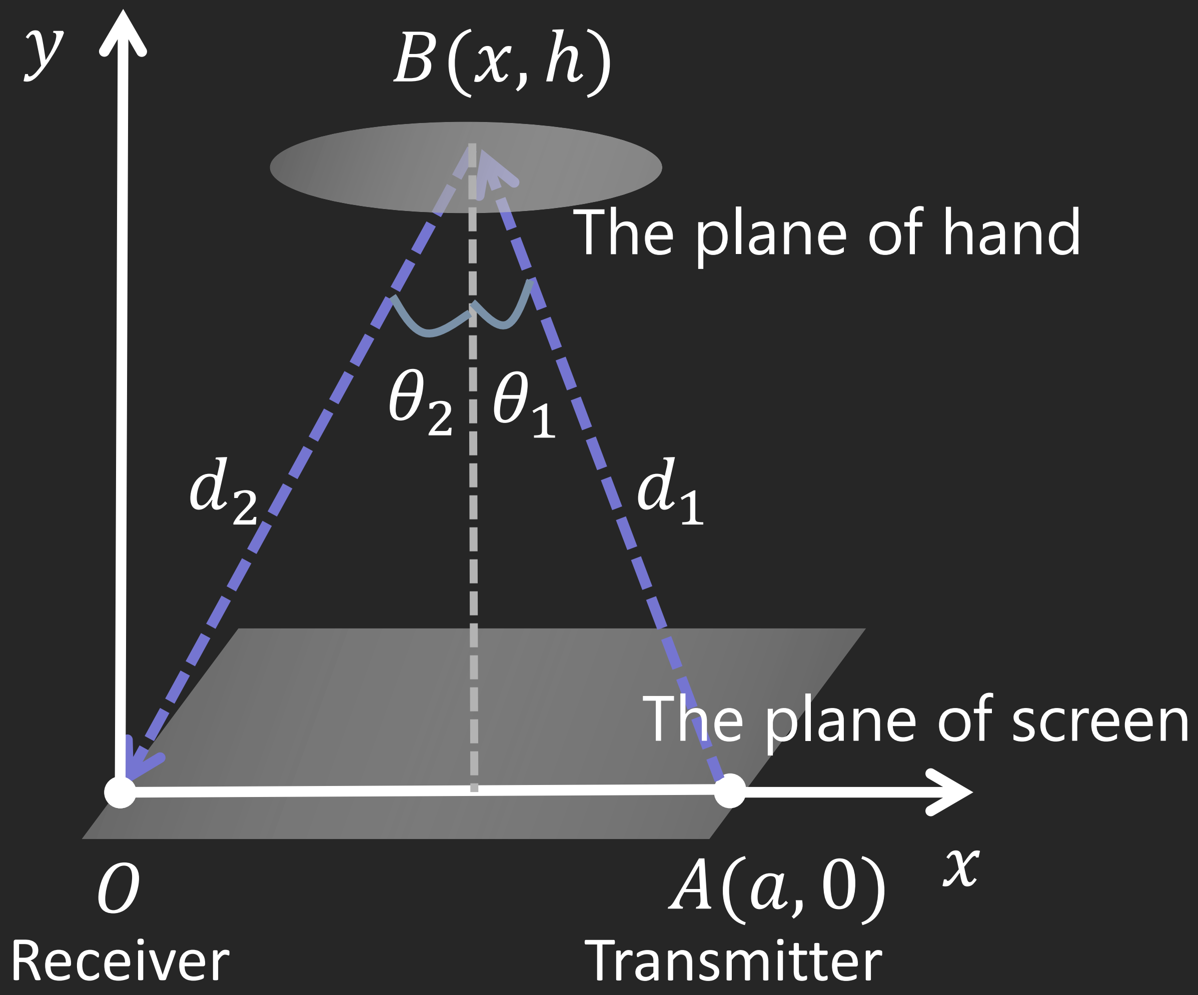
ALS

SMART: A gesture recognition system leveraging **the screen and ambient light sensor(ALS)**.



What is the relationship between the received light power and the hand gesture?

Model the “Screen-Hand-ALS” channel



Calculate received power

$$I_O = \boxed{I_A} \cdot \boxed{l_{AB}} \cdot \boxed{l_B} \cdot \boxed{l_{BO}} \cdot \boxed{l_O}$$

Lambert radiation

$$= I_A \cdot c \cdot \frac{\cos\theta_1}{{d_1}^2} \cdot \frac{\cos{\theta_2}^2}{{d_2}^2}$$

$$= I_A \cdot c \cdot \frac{h^3}{\left((x-a)^2+h^2\right)^{\frac{3}{2}}\left(x^2+h^2\right)^2}$$

I_O: signal power received by point O.

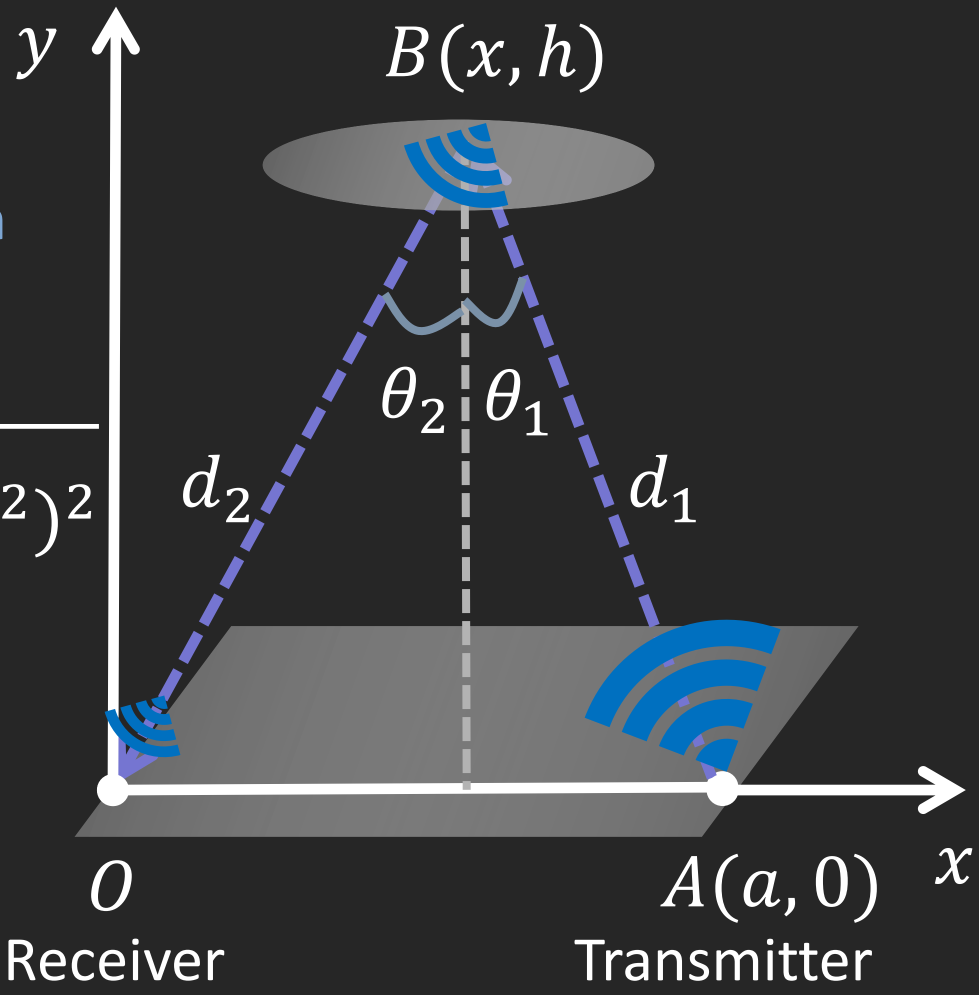
I_A: signal power transmitted by point A.

l_{AB}: loss from A to B.

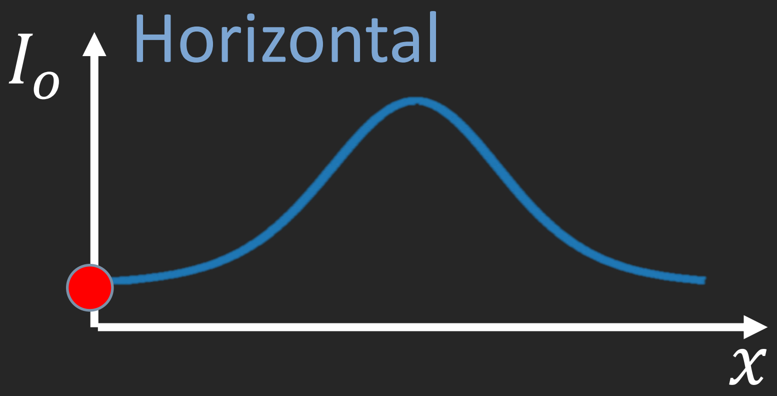
l_{BO}: loss from B to O.

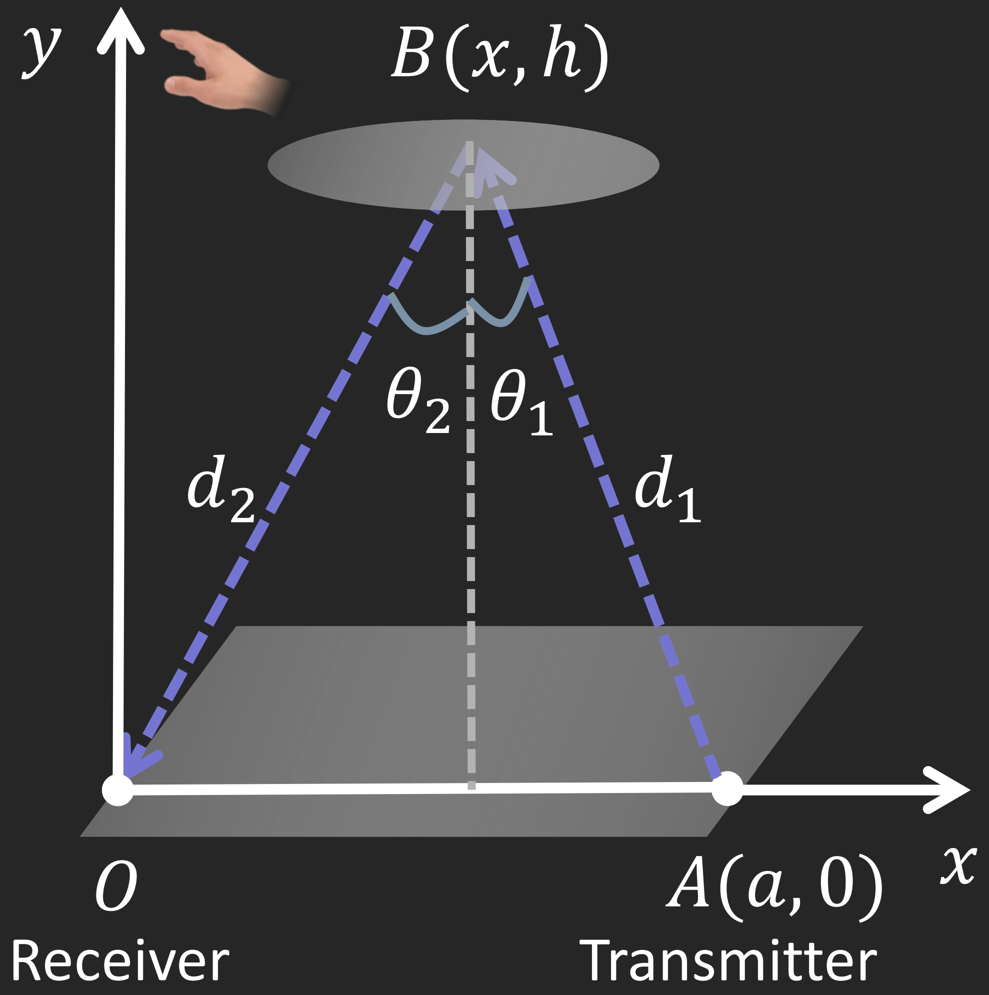
l_B: loss at point B.

l_O: loss at point O.

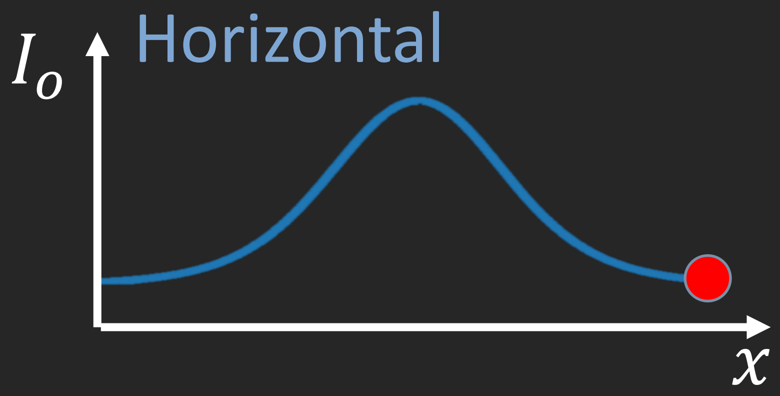


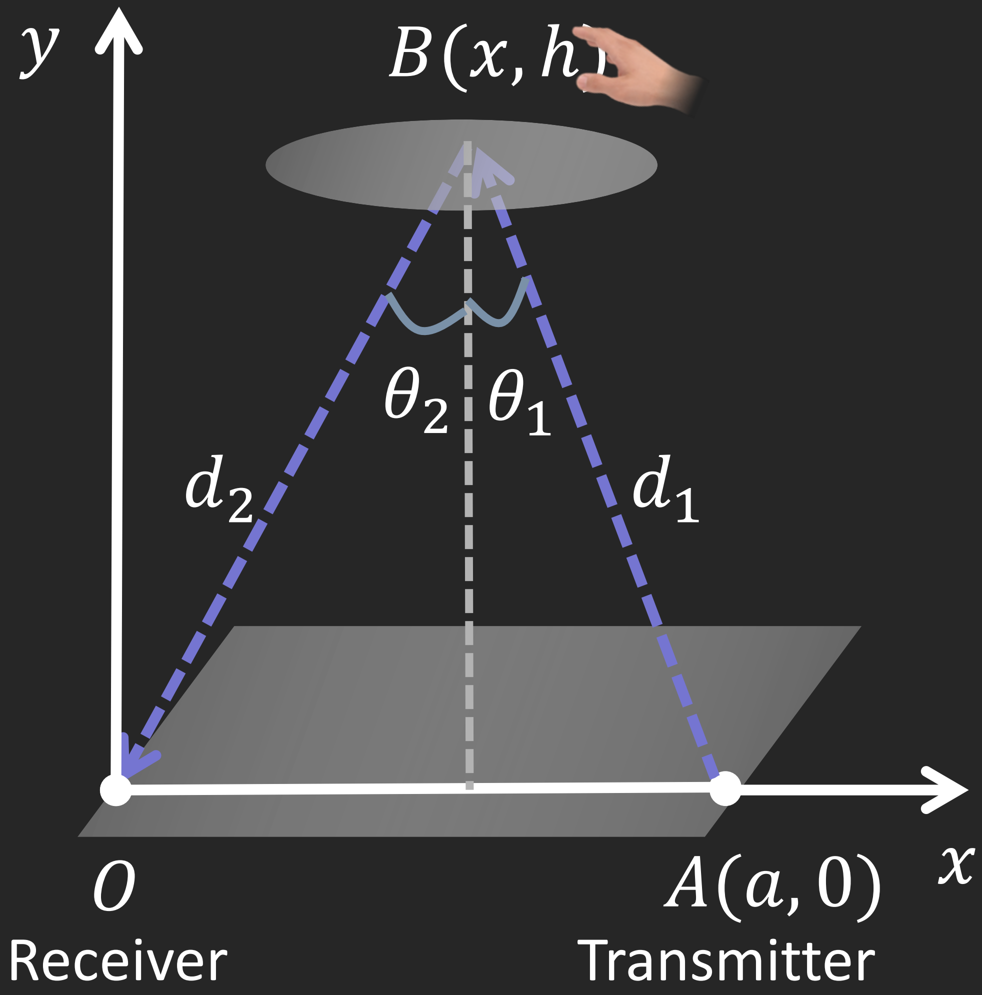
Hand movement and received power



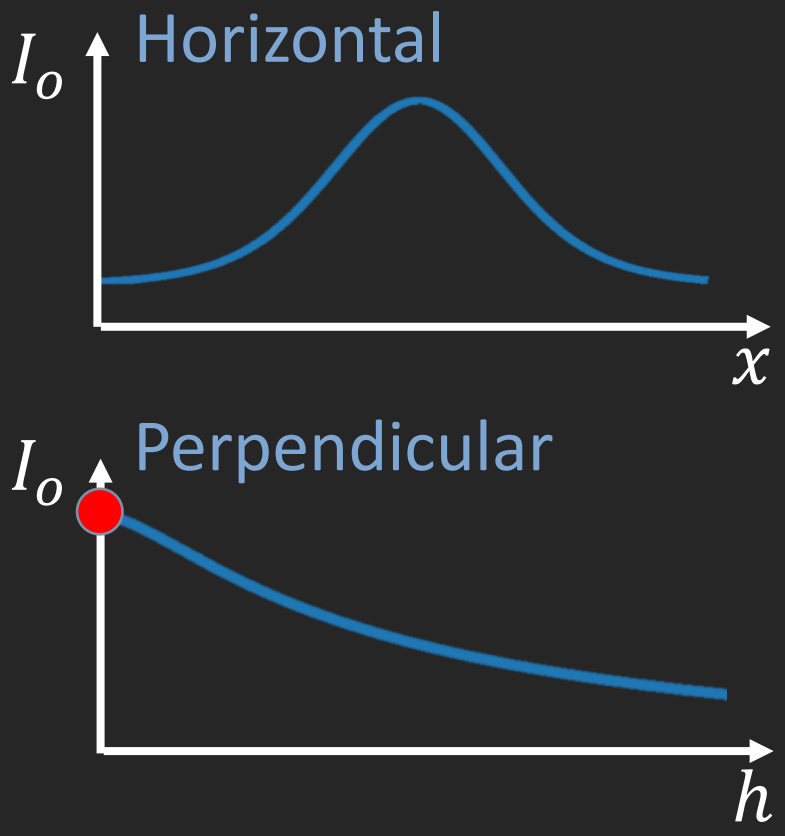


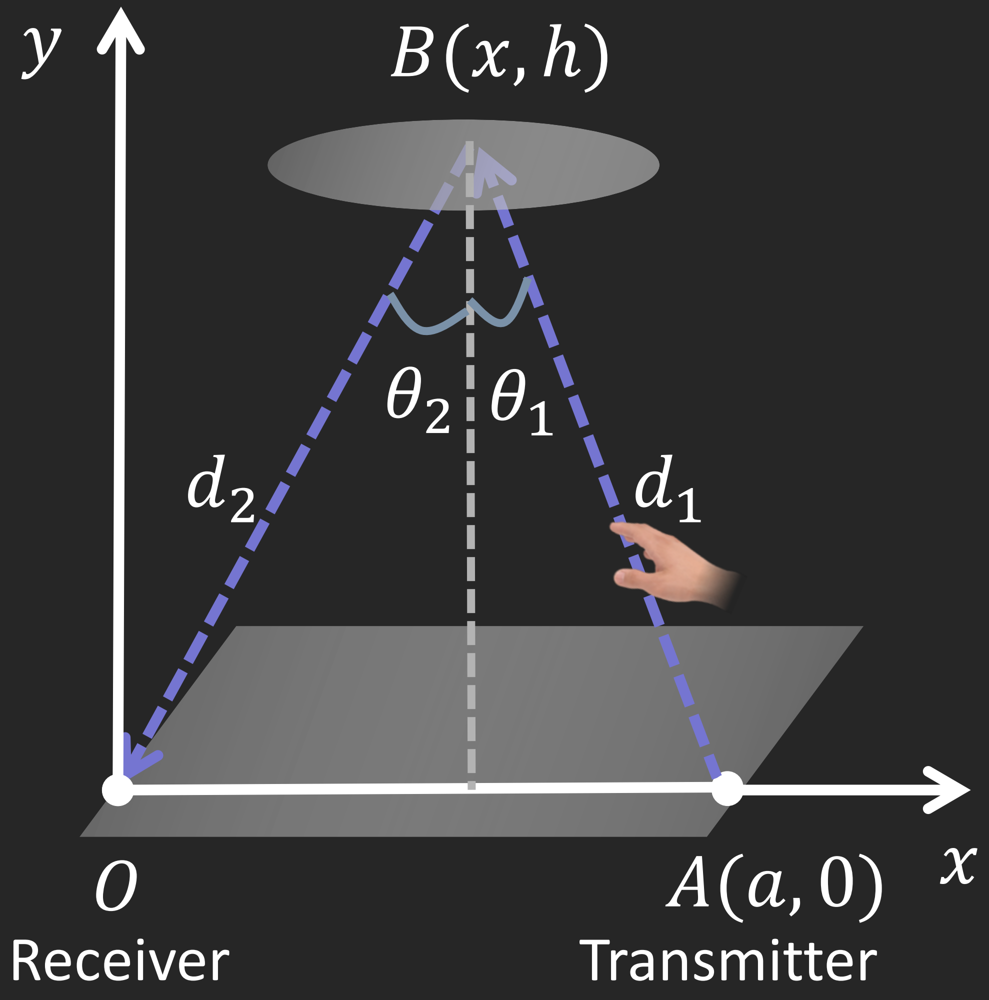
Hand movement and received power



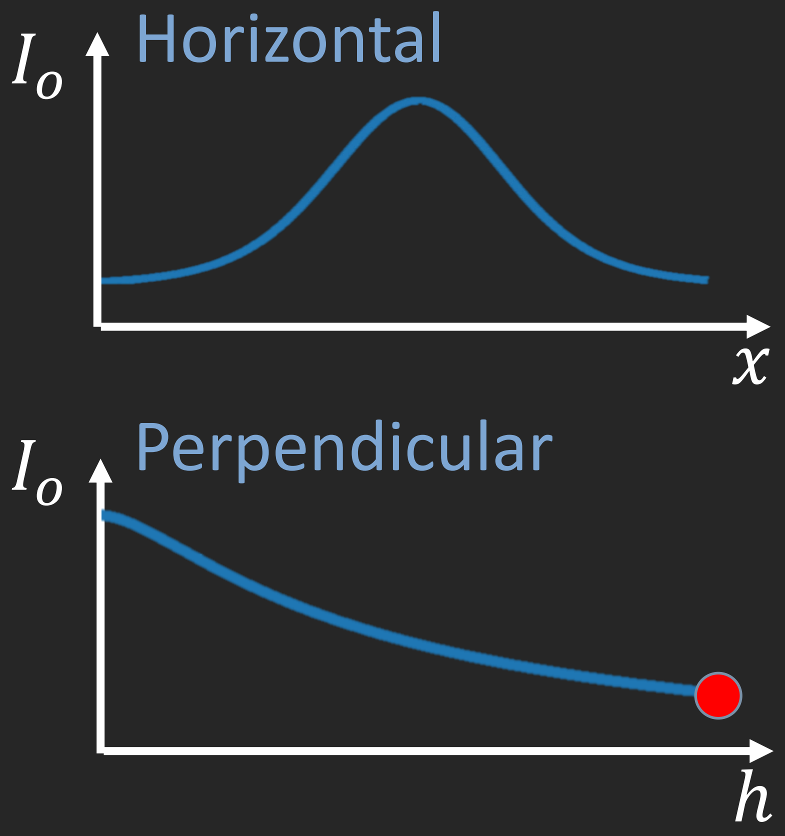


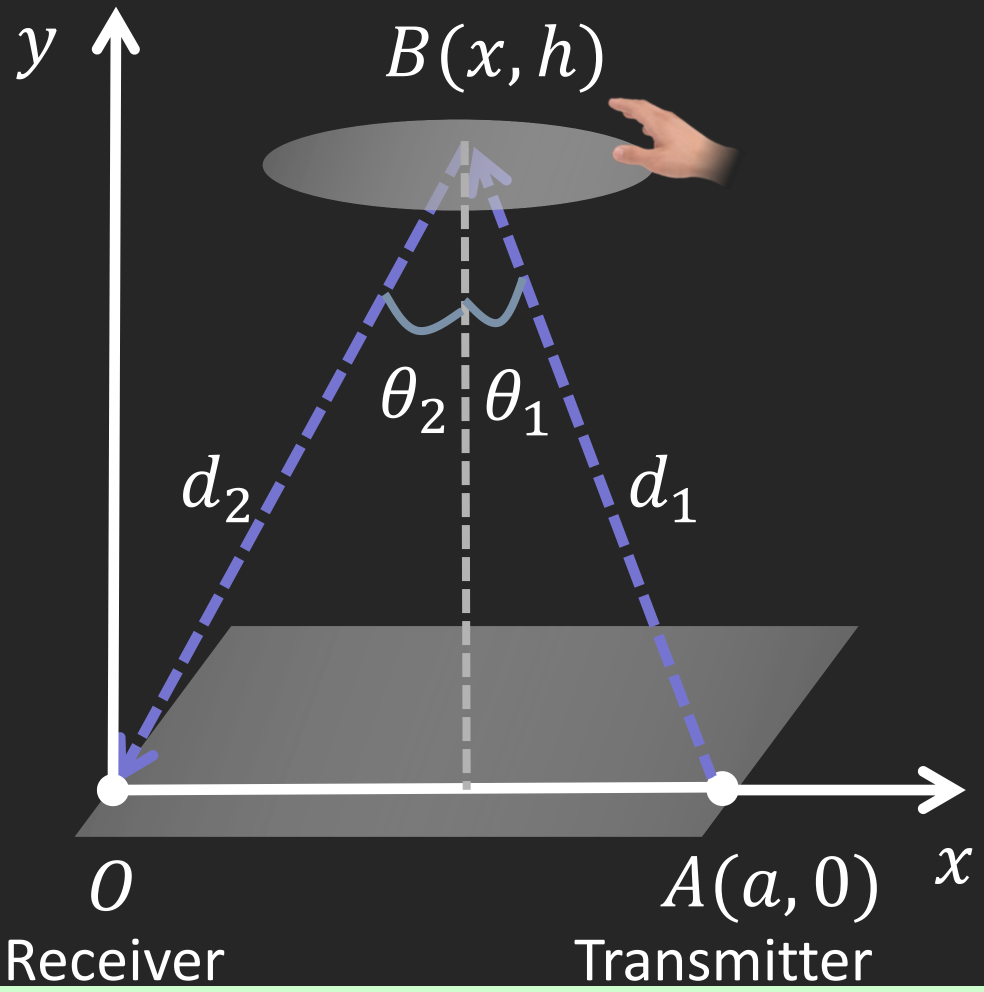
Hand movement and received power





Hand movement and received power





The fundamental working principle of SMART

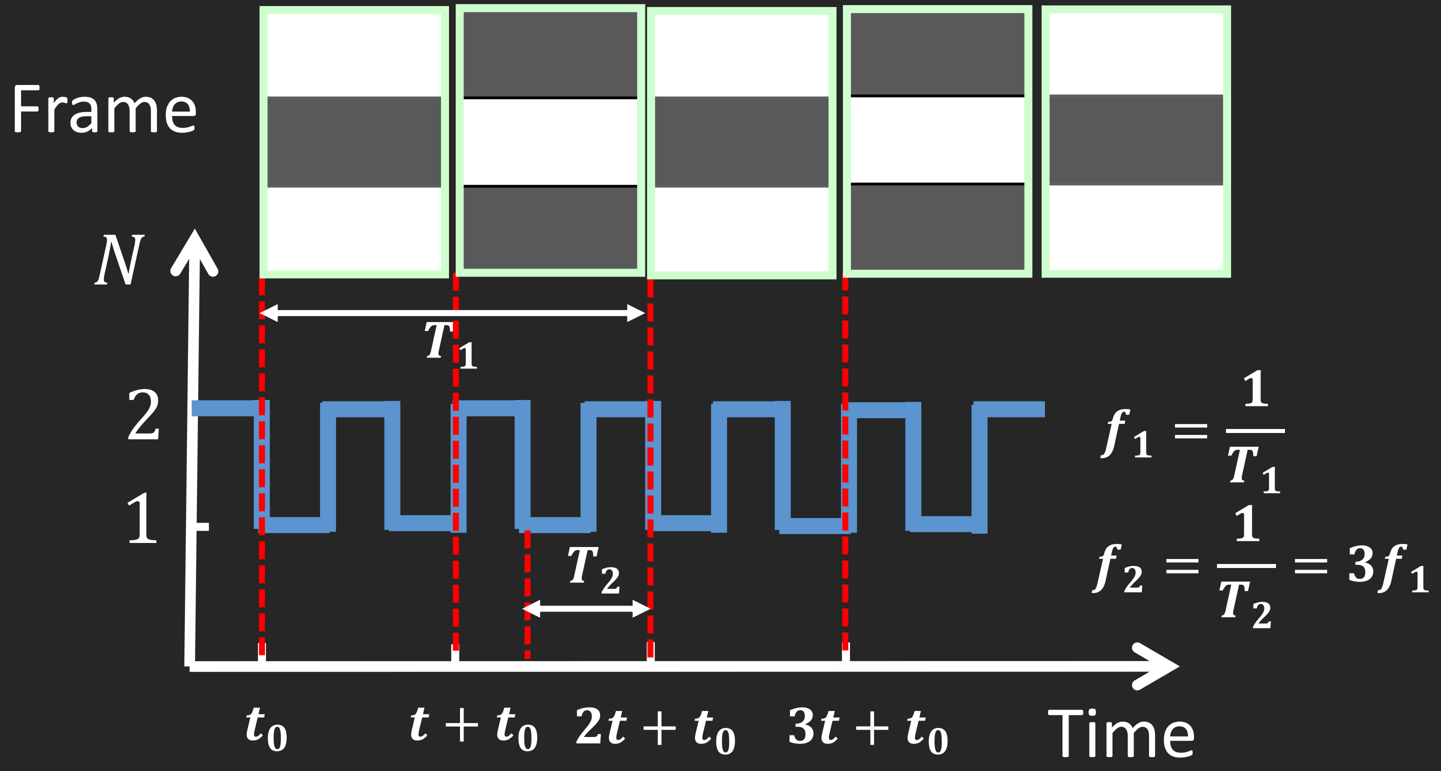
Screen’s refresh rate limits modulated frequency

Refresh time per frame	$\boldsymbol{t}(=\frac{\mathbf{1}}{f_r})$	$\mathbf{2}\boldsymbol{t}$	$\mathbf{3}\boldsymbol{t}$	**… …**	$\mathbf{n}\boldsymbol{t}$
Frequency	f_1	$\frac{f_1}{2}$	$\frac{f_1}{3}$	**… …**	$\frac{f_1}{n}$

Lower frequency

Higher frequency light signals are needed since human eyes are sensitive to low frequency flickering

Transmit high frequency signal



$$f_1 = \frac{1}{T_1}$$

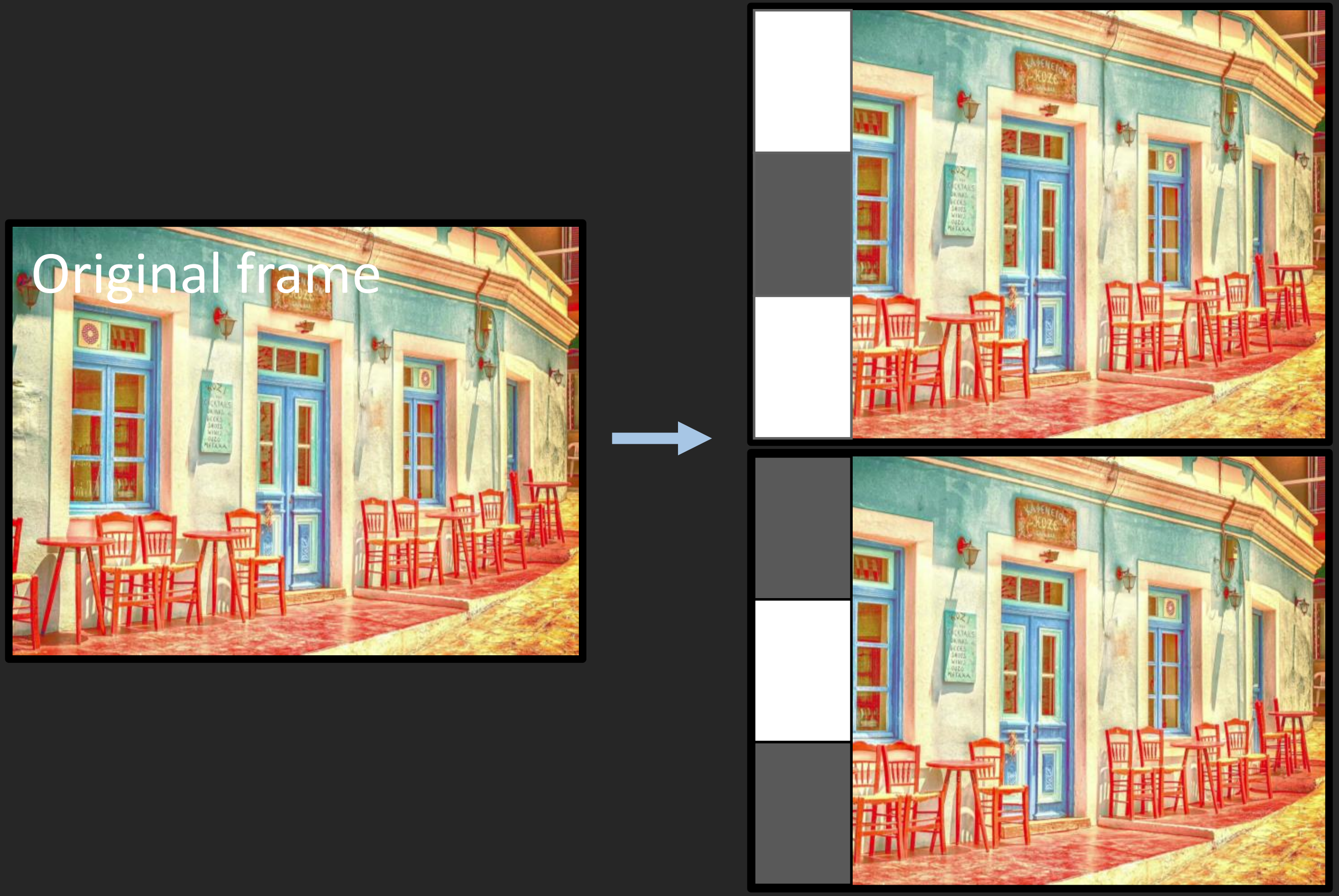
$$f_2 = \frac{1}{T_2} = 3f_1$$

N: the number of bright blocks

t: refreshing time per frame

Hide signals in the screen content

Color decomposition of each pixel



Color space: RGB -> CIE 1931

$$\max \Delta Y = |Y_1 - Y_2|$$

$$s.t. \quad x_1 = x_2 = x_0,$$

$$y_1 = y_2 = y_0,$$

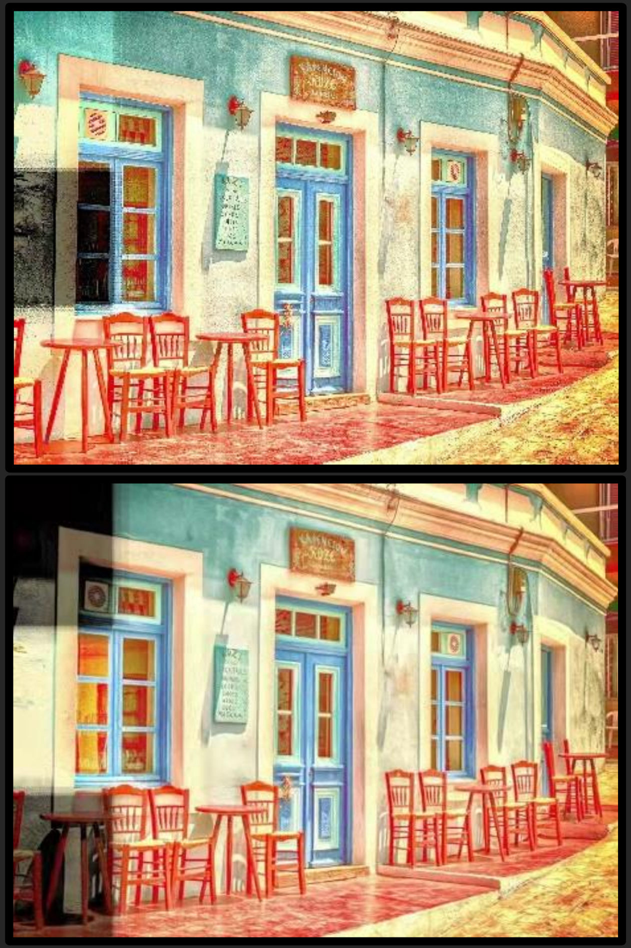
$$Y_0 = \frac{Y_1 + Y_2}{2}$$

Maximum luminance change.

The same chromaticity.

Color addictive rule.

Color decomposition of each pixel



Color space: RGB -> CIE 1931

$$\max \Delta Y = |Y_1 - Y_2|$$

$$s.t. \quad x_1 = x_2 = x_0,$$

$$y_1 = y_2 = y_0,$$

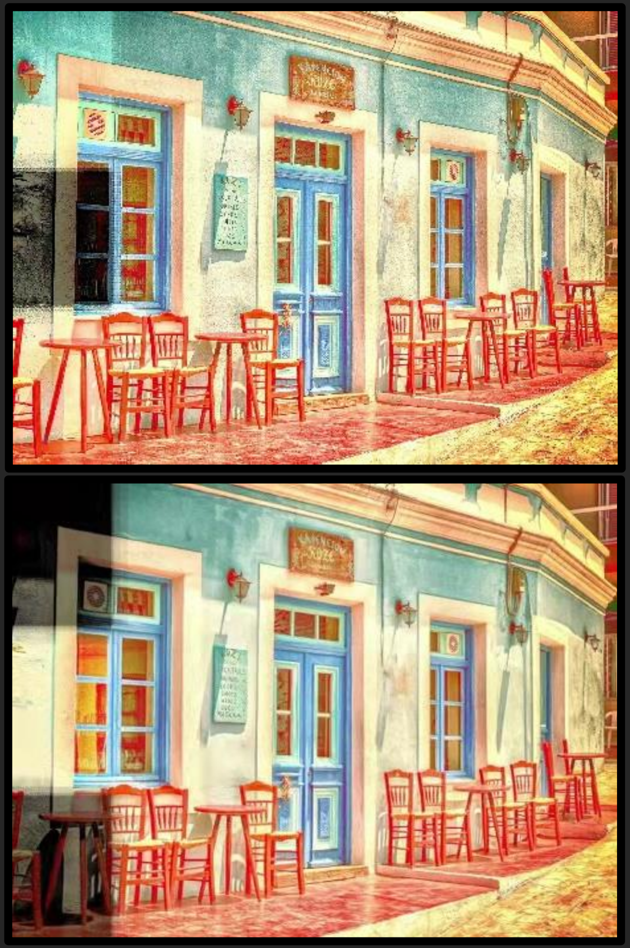
$$Y_0 = \frac{Y_1 + Y_2}{2}$$

Maximum luminance change.

The same chromaticity.

Color addictive rule.

Color decomposition of each pixel



Color space: RGB -> CIE 1931

$$\max \Delta Y = |Y_1 - Y_2|$$

$$s.t.\quad x_1 = x_2 = x_0,$$

$$y_1 = y_2 = y_0,$$

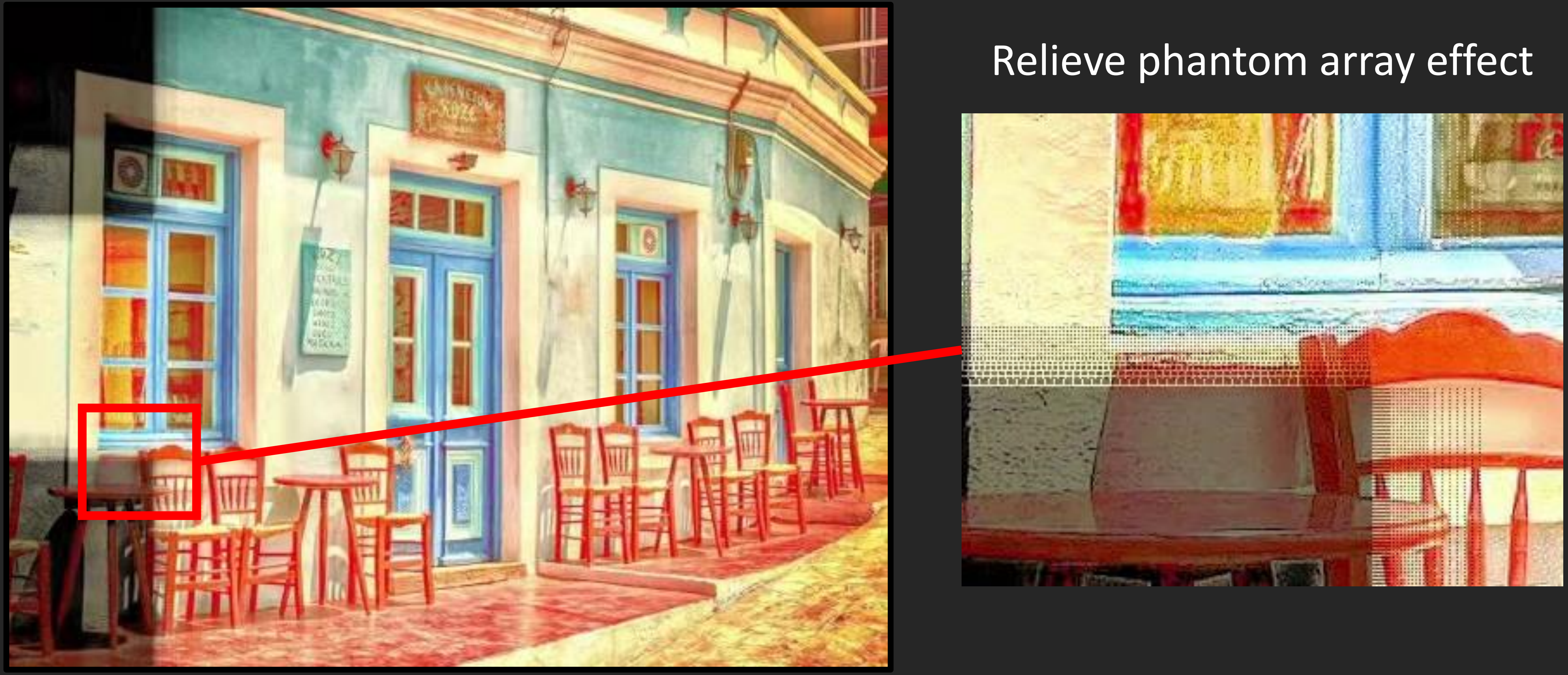
$$Y_0 = \frac{Y_1+Y_2}{2}$$

Maximum luminance change.

The same chromaticity.

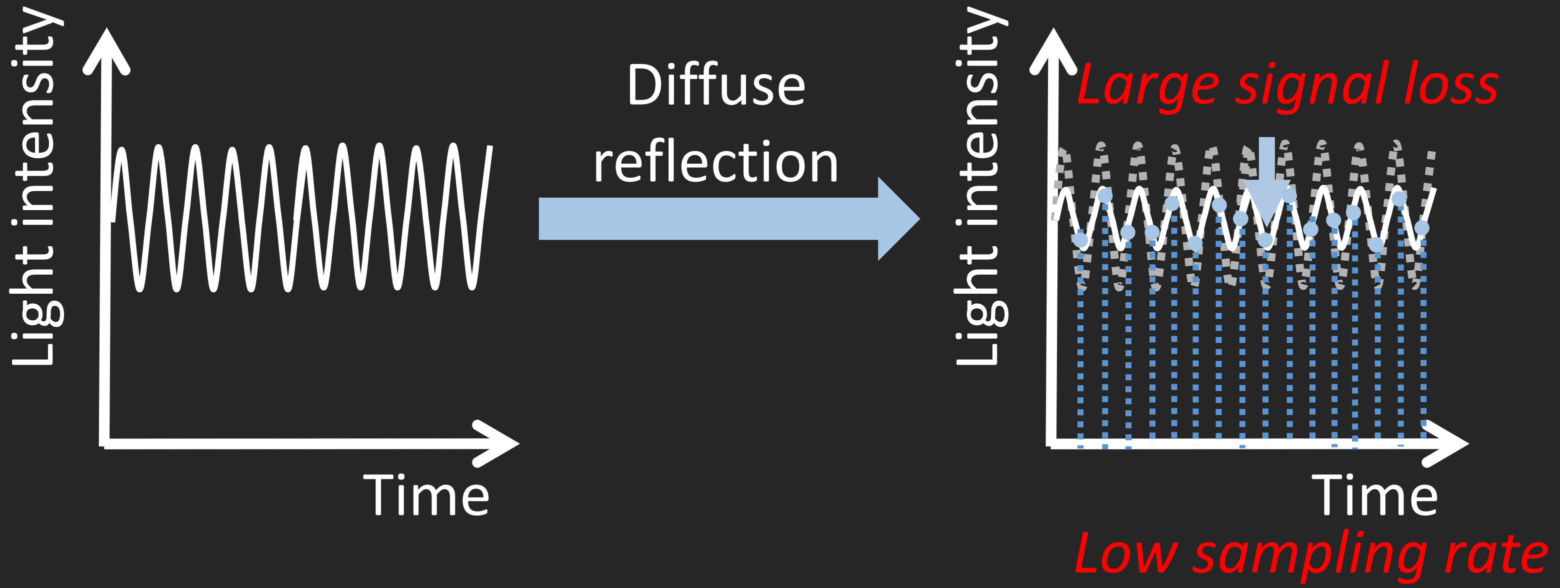
Color addictive rule.

Edge smoothing



Relieve phantom array effect

Signal received by ALS is low-quality



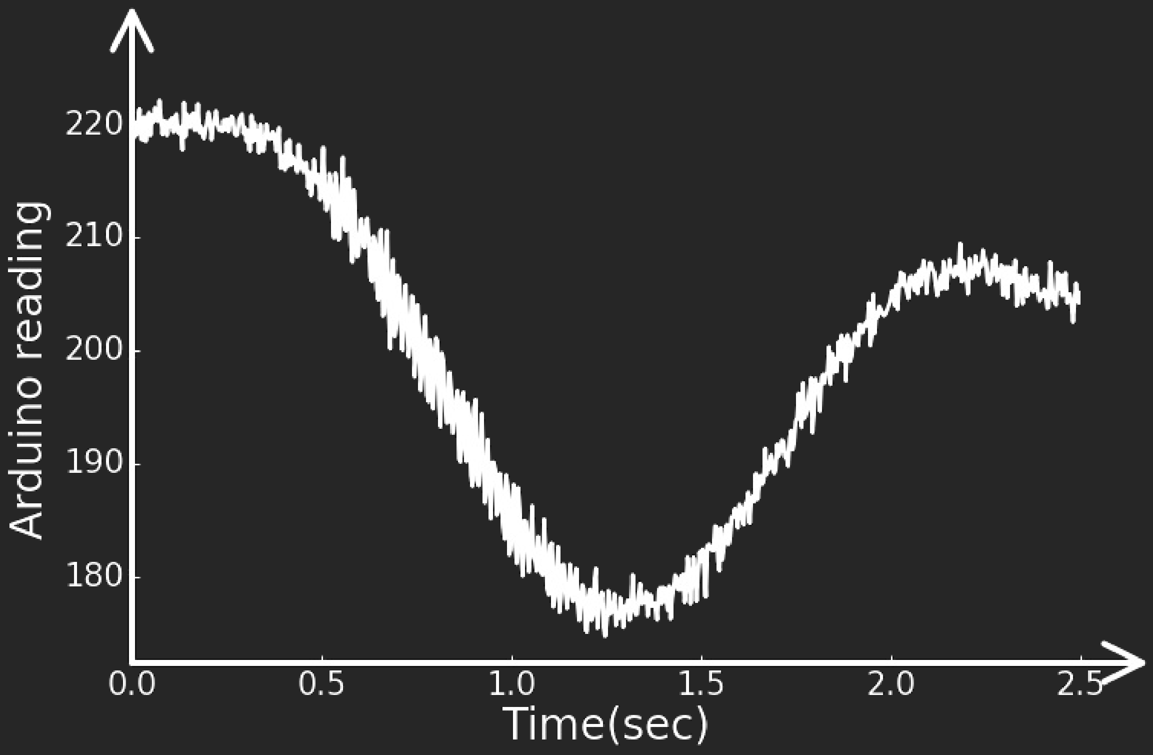
Diffuse reflection

Large signal loss

Low sampling rate

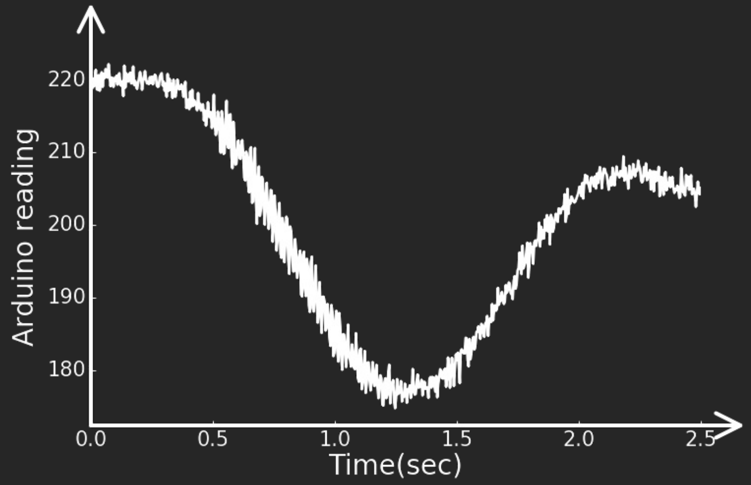
Segmentation according to reflected power

Raw signal



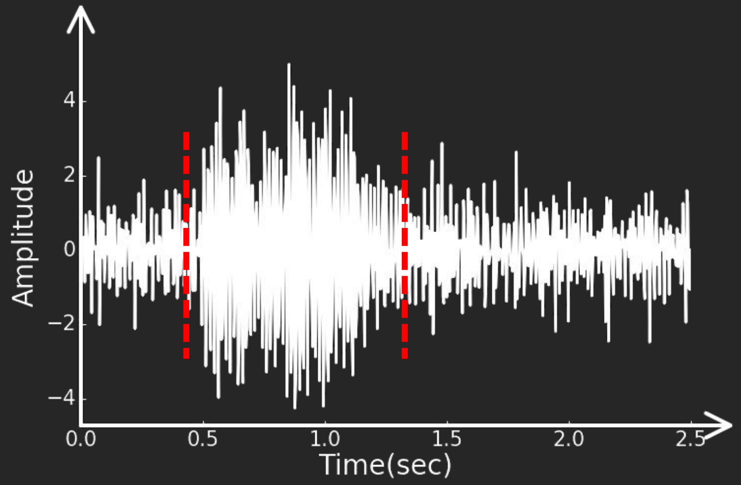
Segmentation according to reflected power

Raw signal

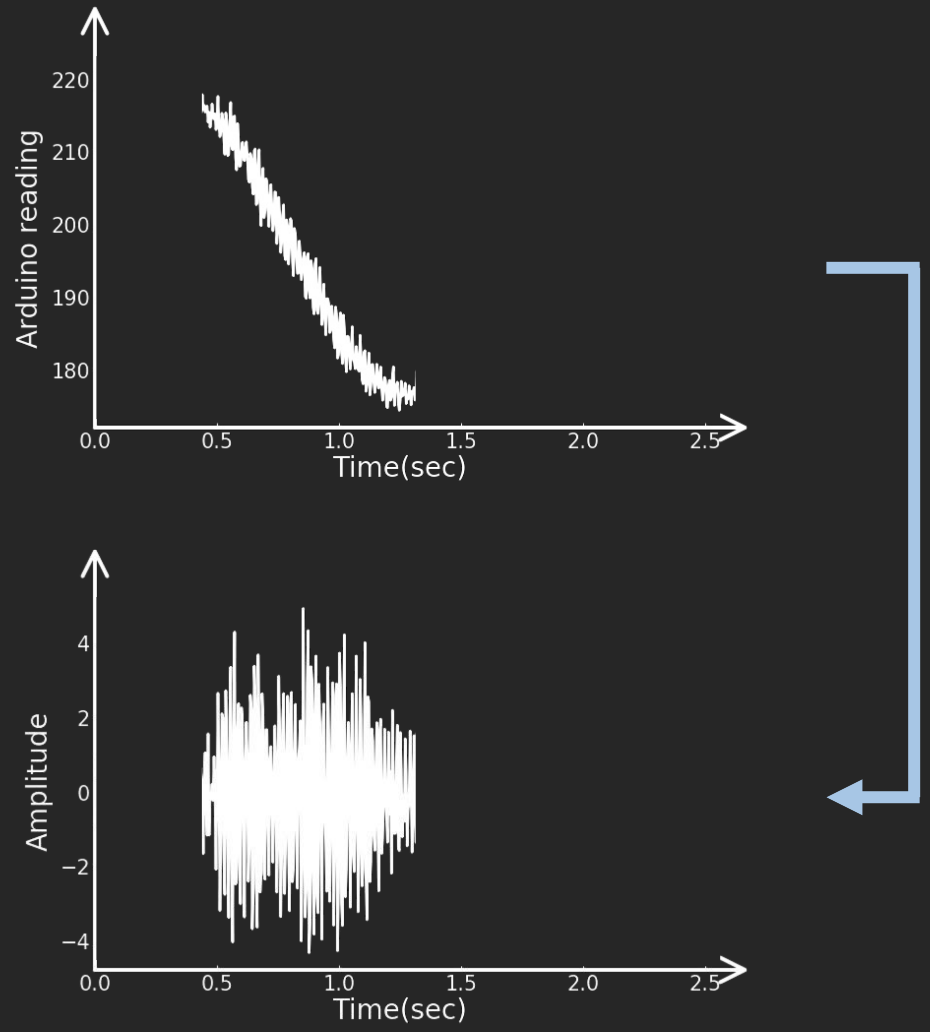


High-pass filter

Reflected signal



Segmentation according to reflected power

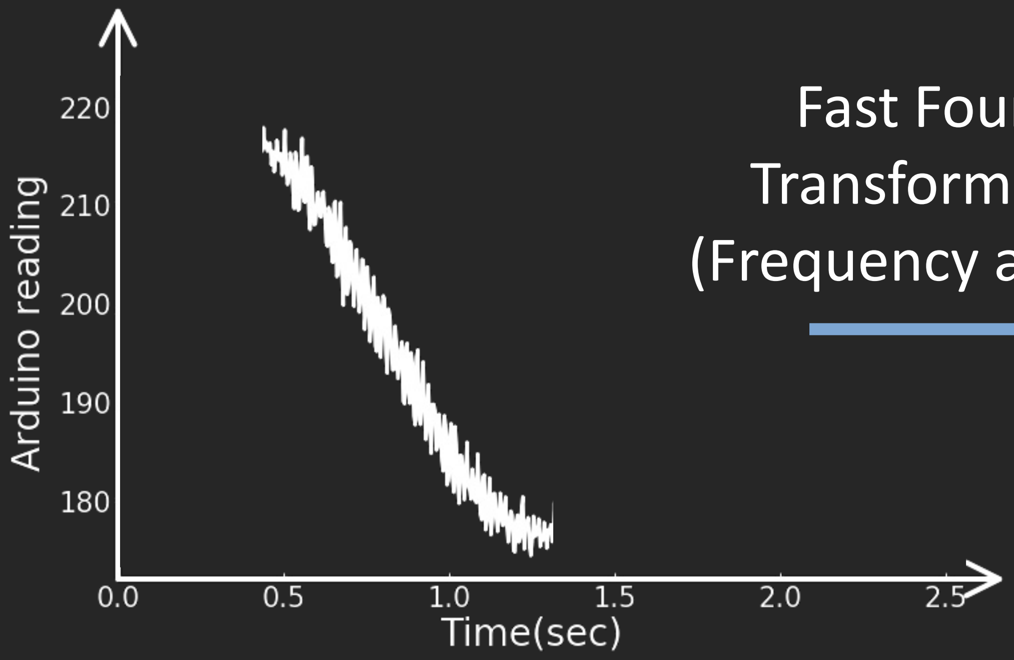


Raw signal

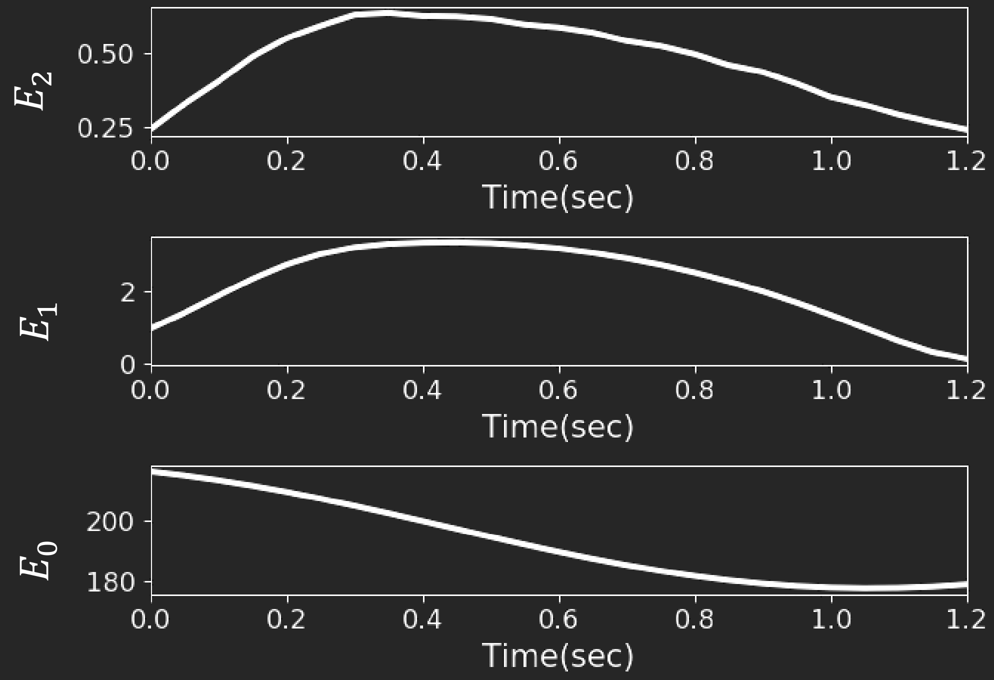
High-pass filter

Reflected signal

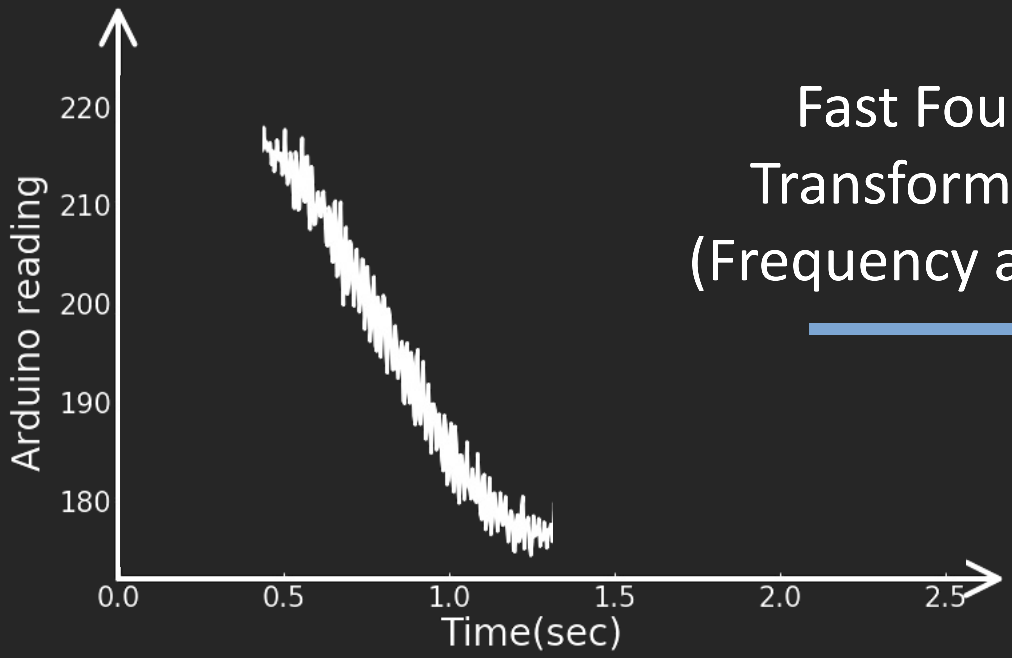
Signal pre-processing and classification



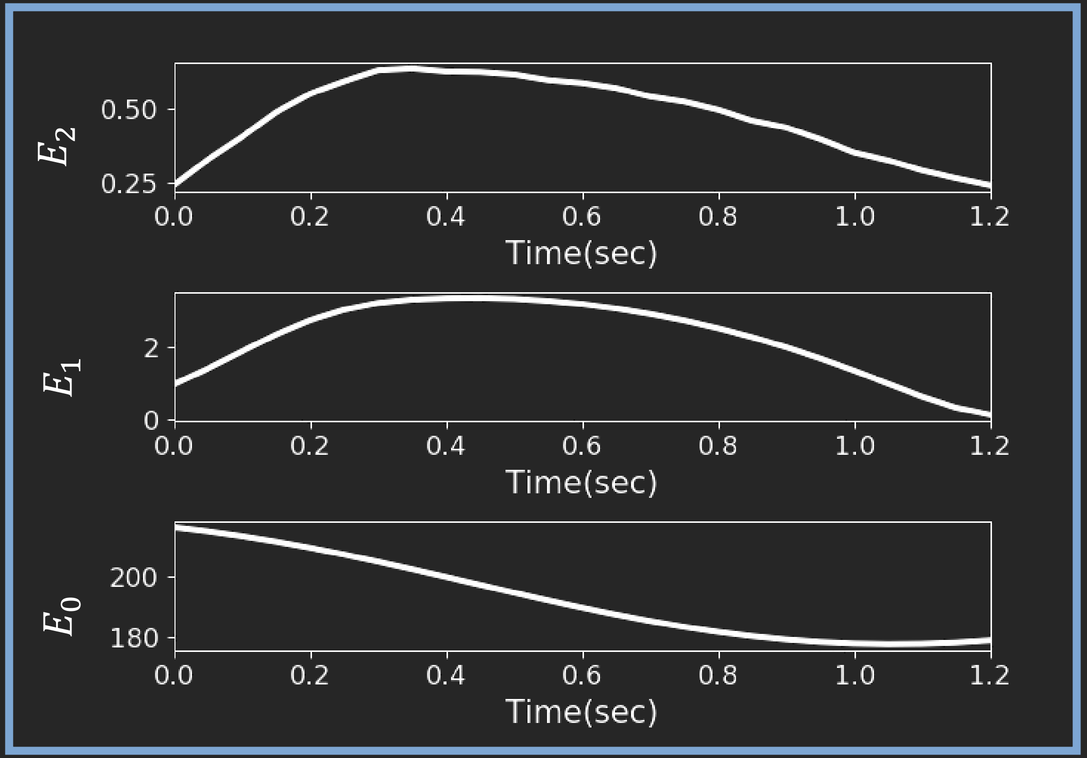
Fast Fourier Transformation (Frequency aliasing)



Signal pre-processing and classification



Fast Fourier Transformation (Frequency aliasing)

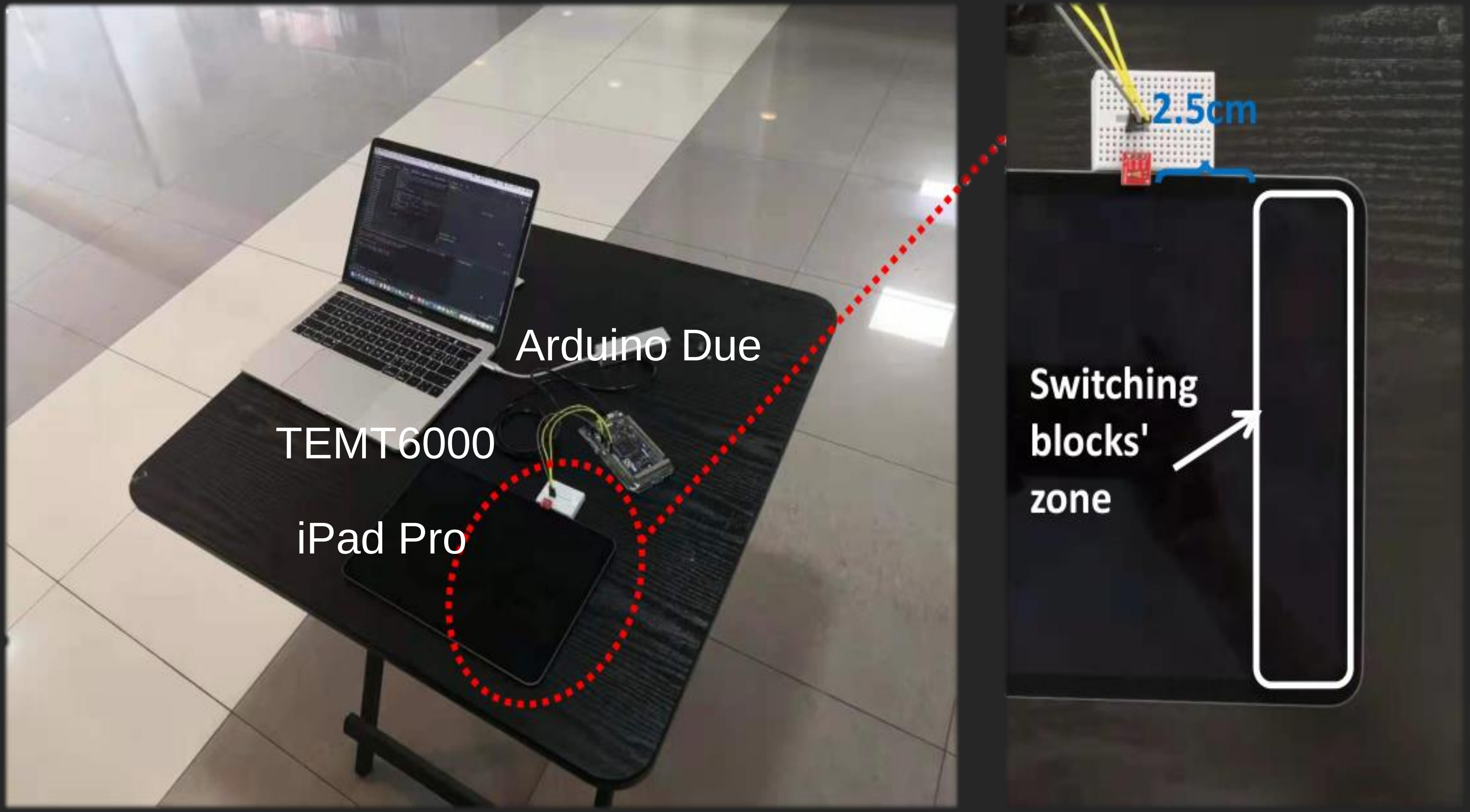


Feature extraction & Classification

Gesture recognition

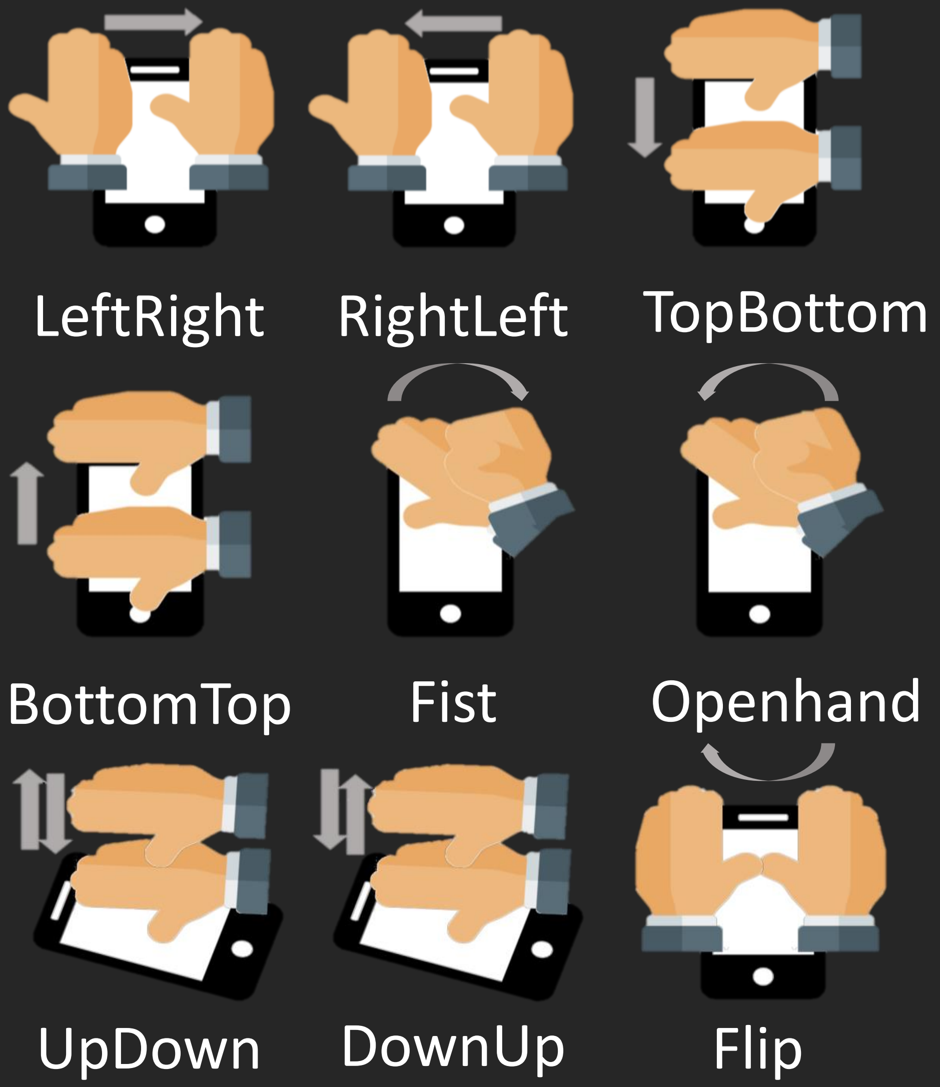
Evaluation

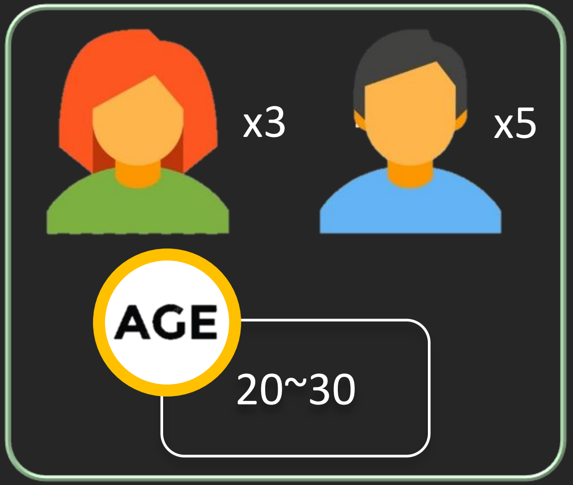
- Prototype
  - Transmitter: iPad Pro 11;
  - Receiver: TEMT6000(250Hz); Arduino Due;
- Experiment setting
  - 9 gestures;
  - 8 users;
  - 5 static & 2 dynamic lighting environments;

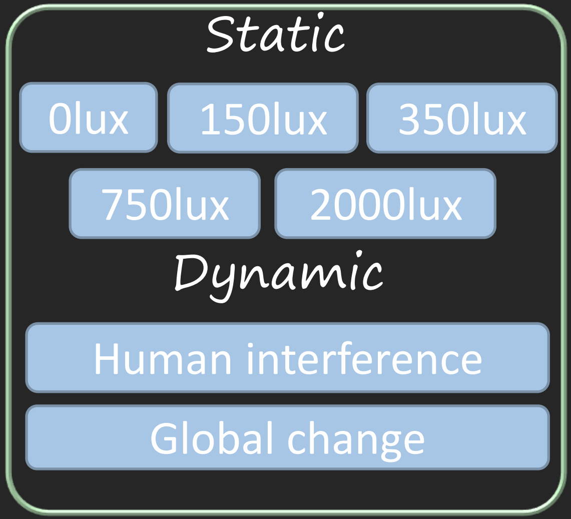


Evaluation

- Prototype
  - Transmitter: iPad Pro 11;
  - Receiver: TEMT6000(250Hz); Arduino Due;
- Experiment setting
  - 9 gestures;
  - 8 users;
  - 5 static & 2 dynamic lighting environments;







Static

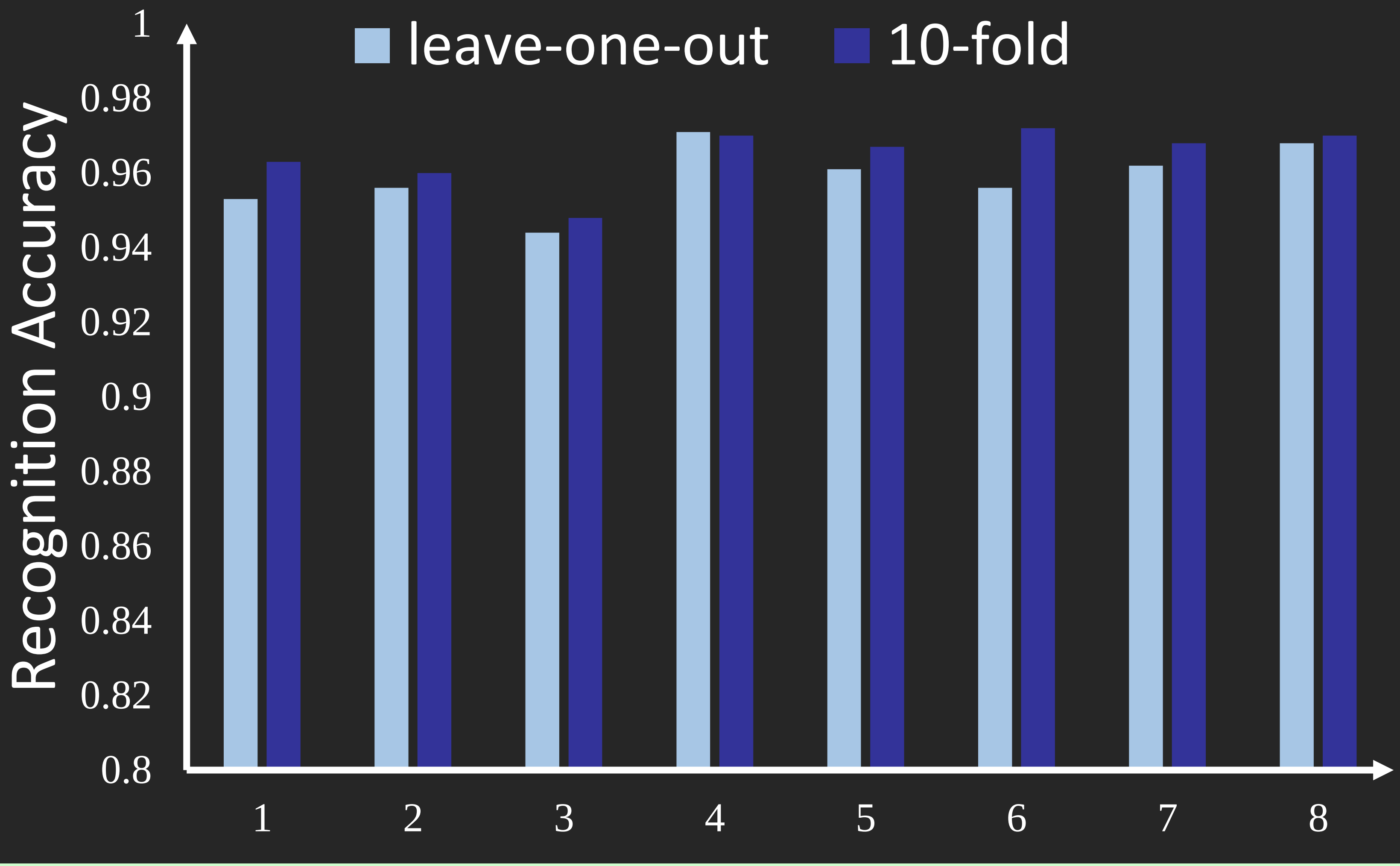
0lux | 150lux | 350lux | 750lux | 2000lux

Dynamic

Human interference

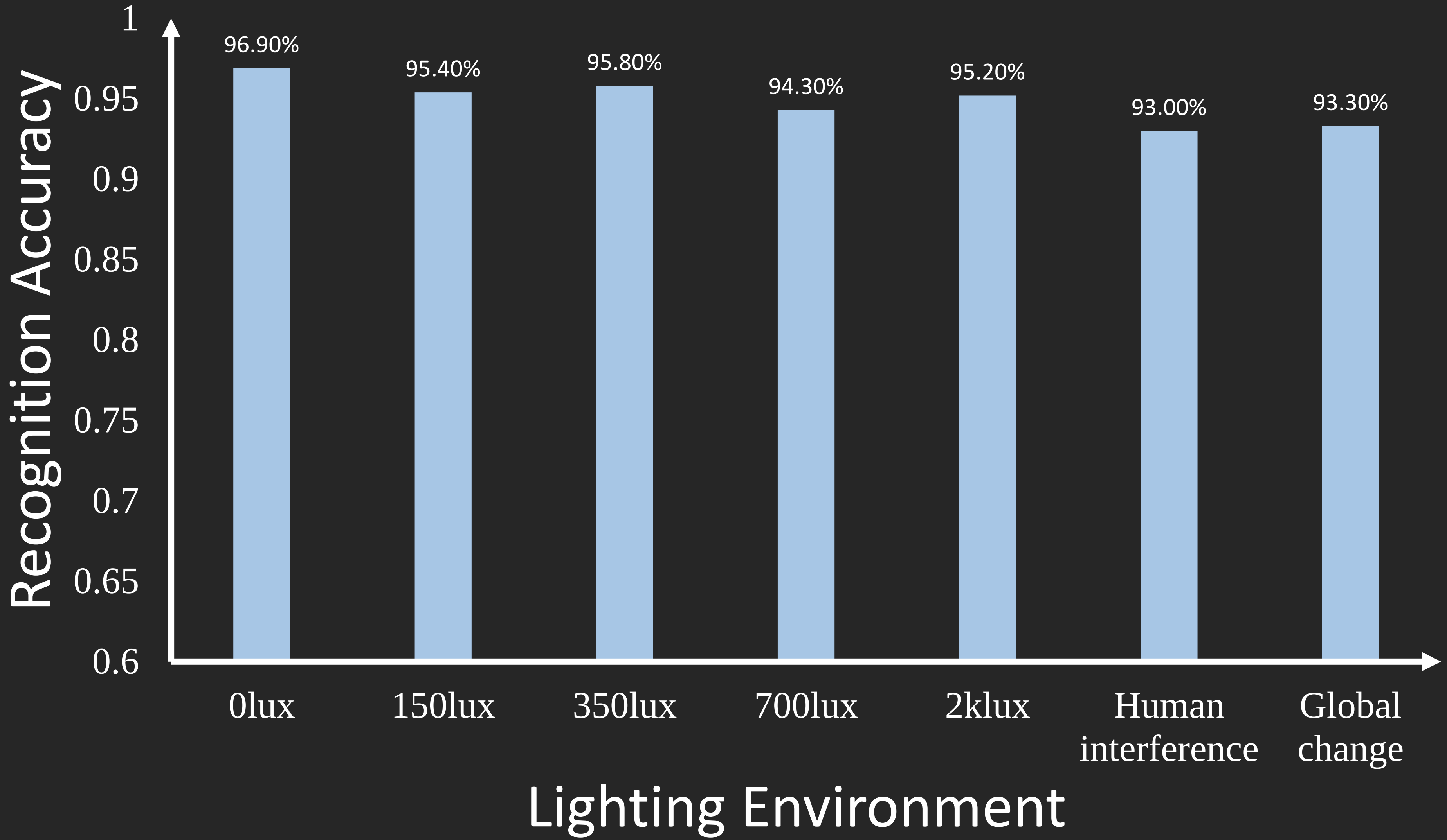
Global change

Accuracy v.s. Different users

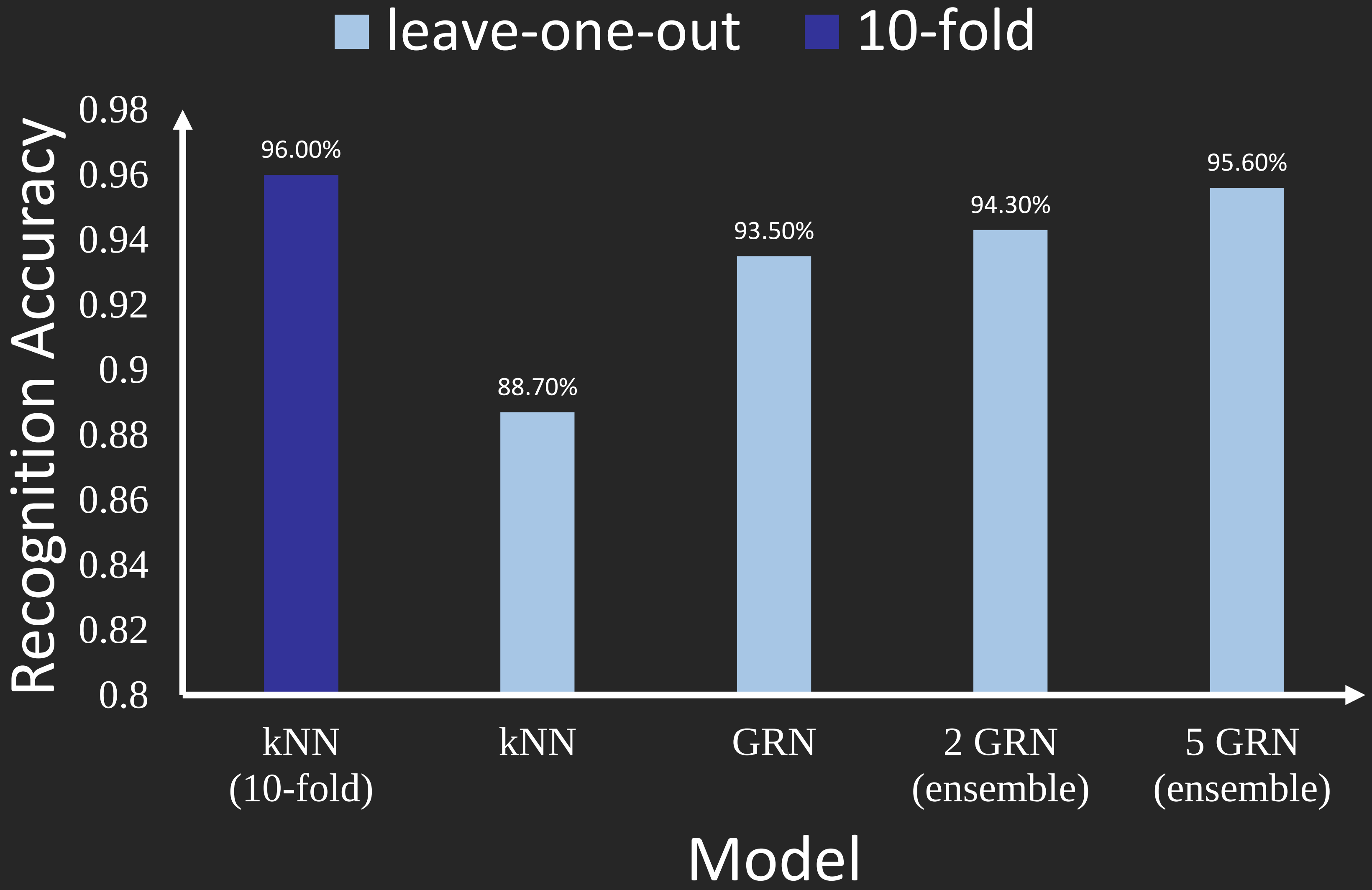


SMART is a generic model.

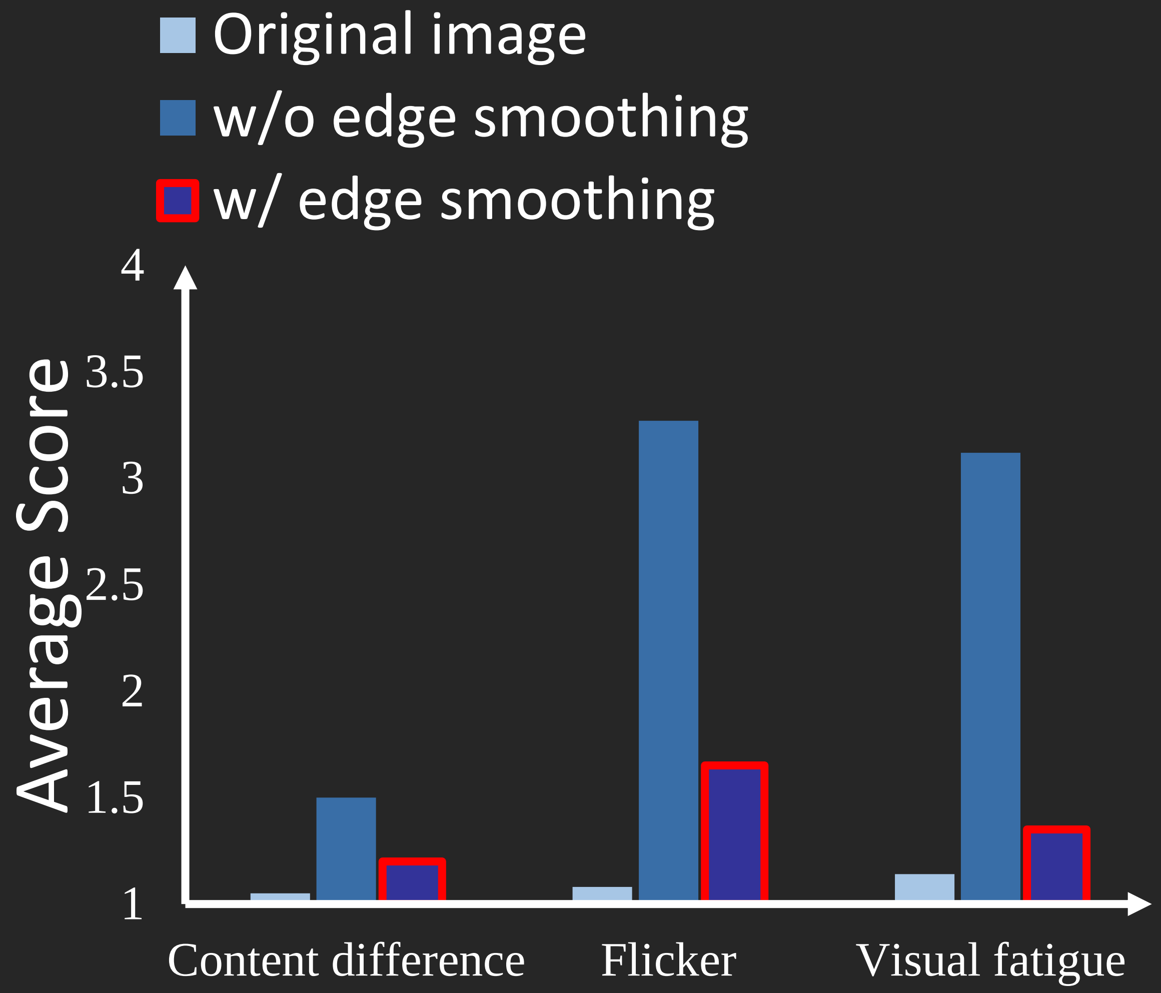
Accuracy v.s. Different lighting environments



Accuracy v.s. Unseen lighting environments



User perception

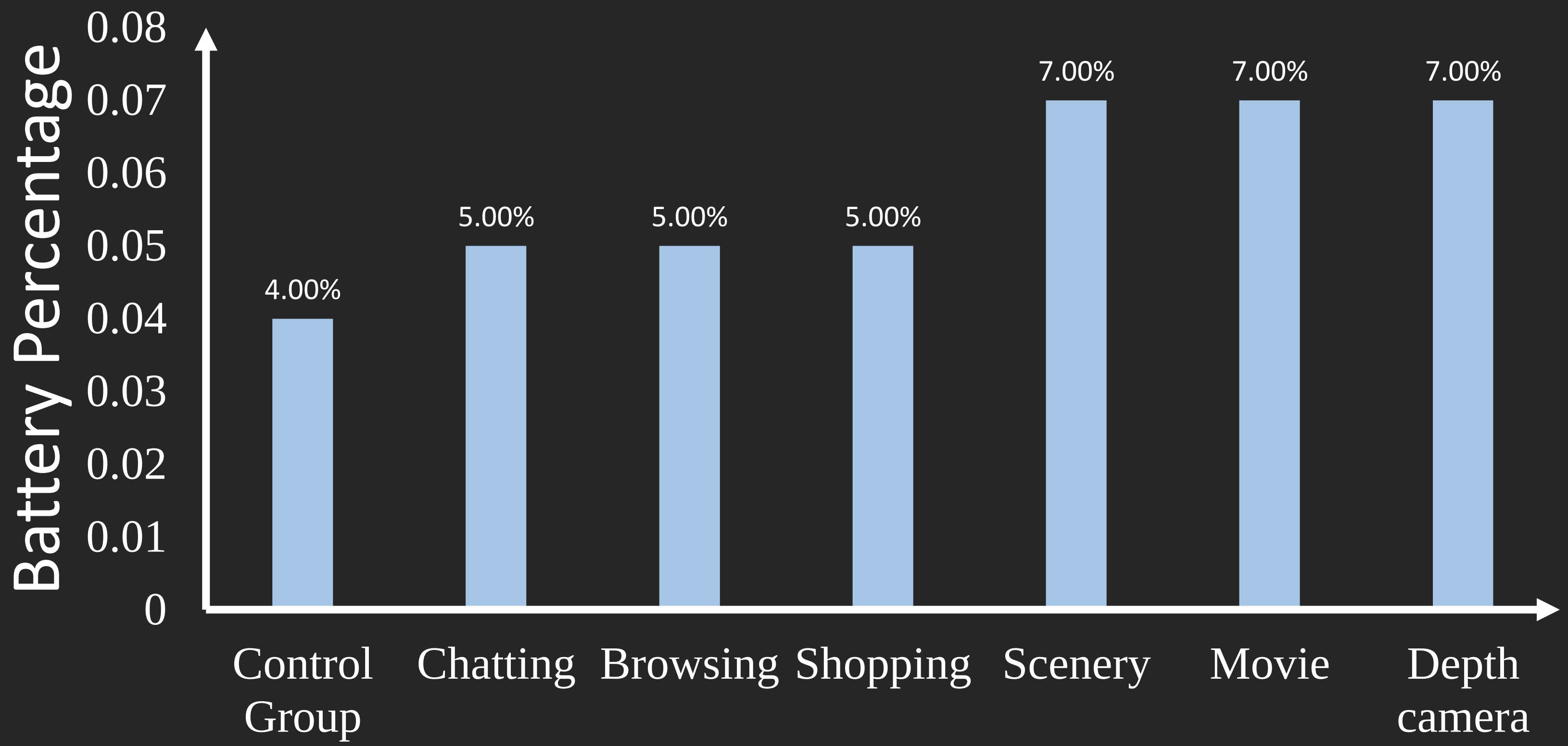


15 volunteers, 6 different images

Score	Content difference	Flicker	Visual fatigue
1	The same	No flickering	No fatigue
2	↓	↓	↓
3			
4	Evidently different	Evident flickering	Strong fatigue

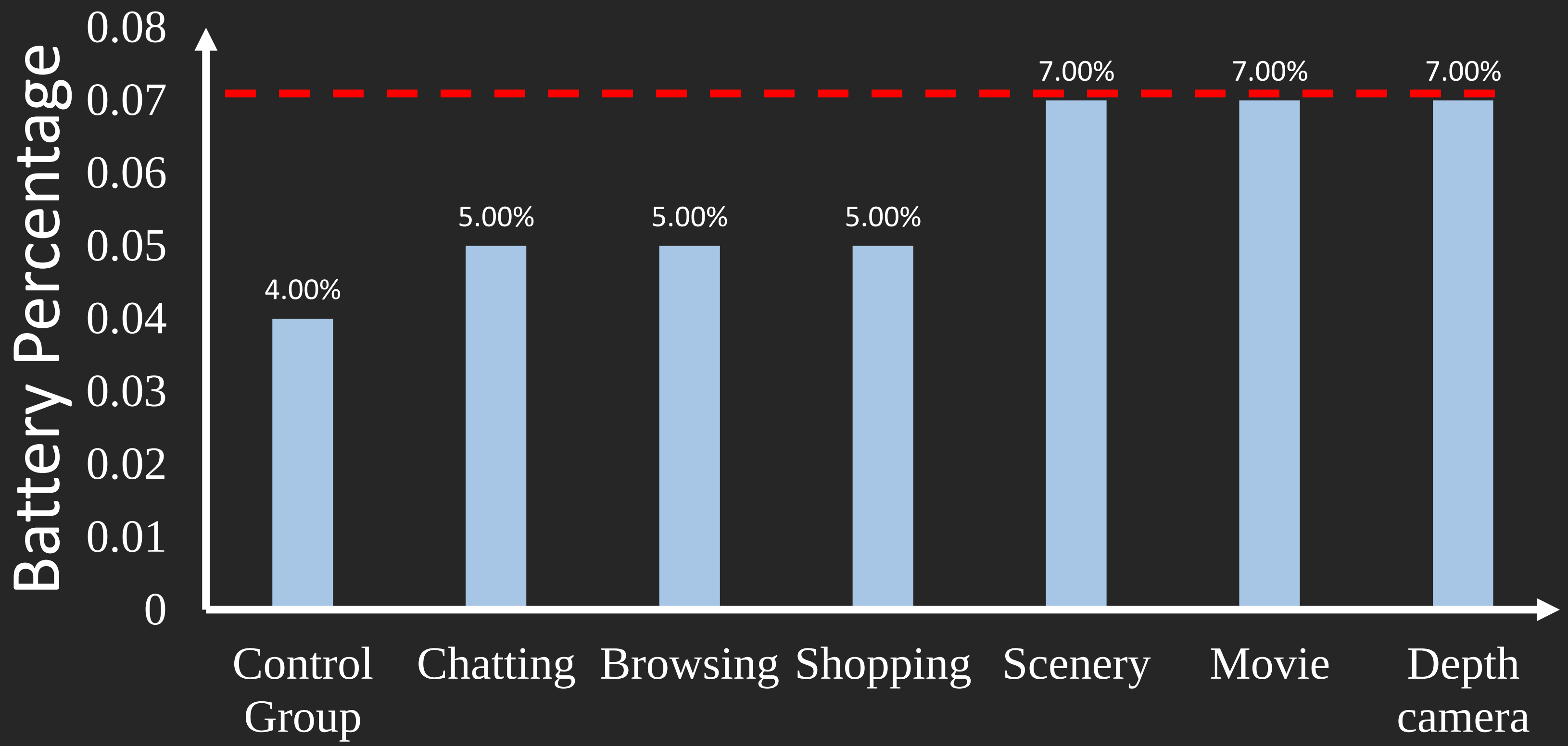
Power consumption comparison with depth camera

Huawei
Mate30 Pro



Power consumption comparison with depth camera

Huawei Mate30 Pro



SMART's power consumption is lower than depth-camera

Thanks for your attention!

Q&A