Screen Perturbation: Adversarial Attack and Defense on Under-Screen Camera

*Hanting Ye**, *Guohao Lan**, *Jinyuan Jia*†, *Qing Wang**

**Delft University of Technology*

†The Pennsylvania State University

October 3rd, 2023

European Training Network in Low-Energy Visible Light IoT Systems

Evolution of smartphone screens

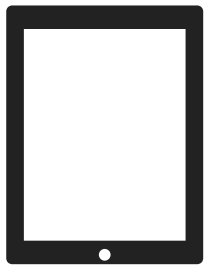
Notch Screen

Punch-Hole

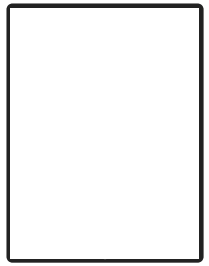
Drop Notch

Full-screen

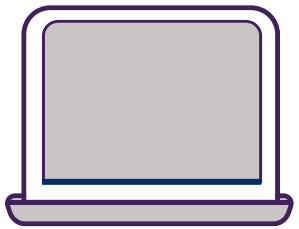
ZTE AXON,
Samsung Fold,
Xiaomi MIX,
Redmi
(2022 - ⋯)



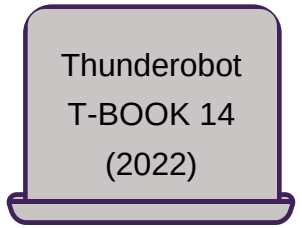
Traditional tablet



Full-screen tablet

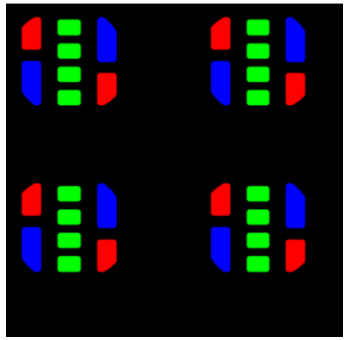


Traditional laptop

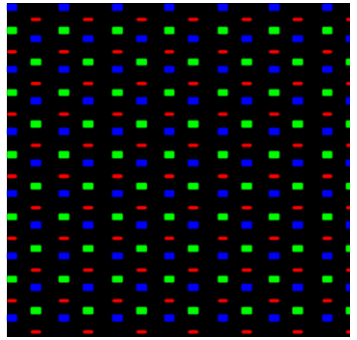


Full-screen laptop

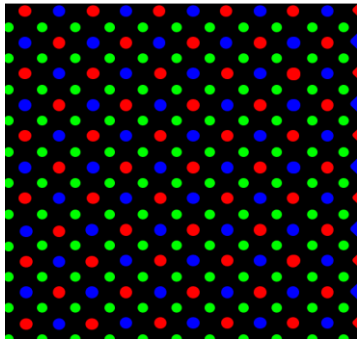
Under-screen camera



Samsung

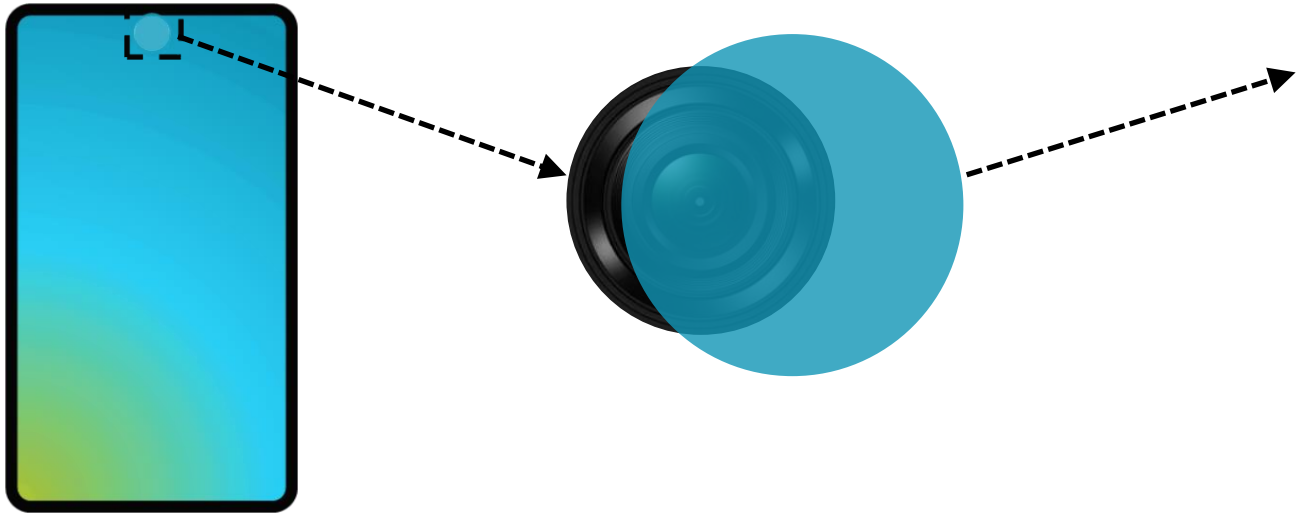


ZTE



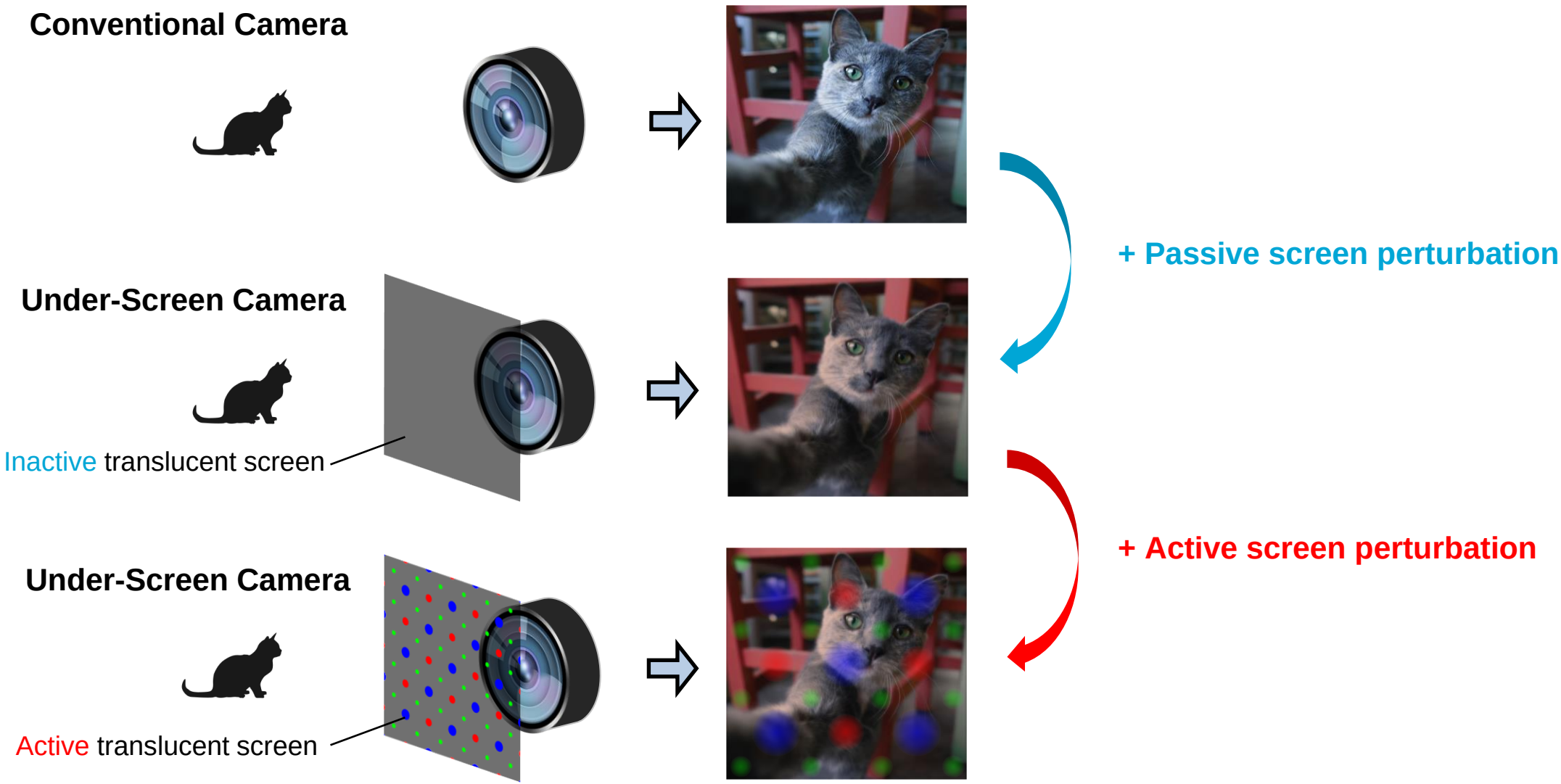
Xiaomi

Full-Screen Smartphone **Under-Screen Camera** **Translucent Screen Region**

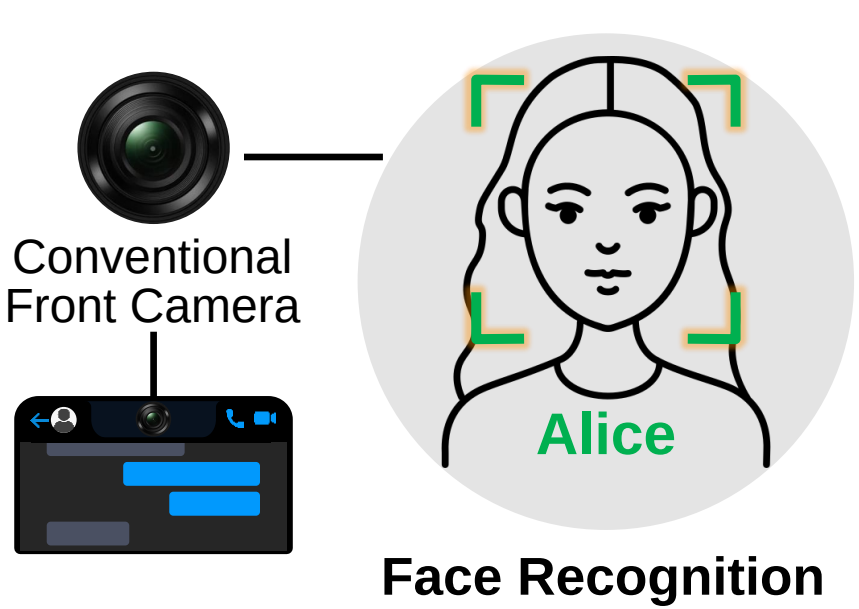


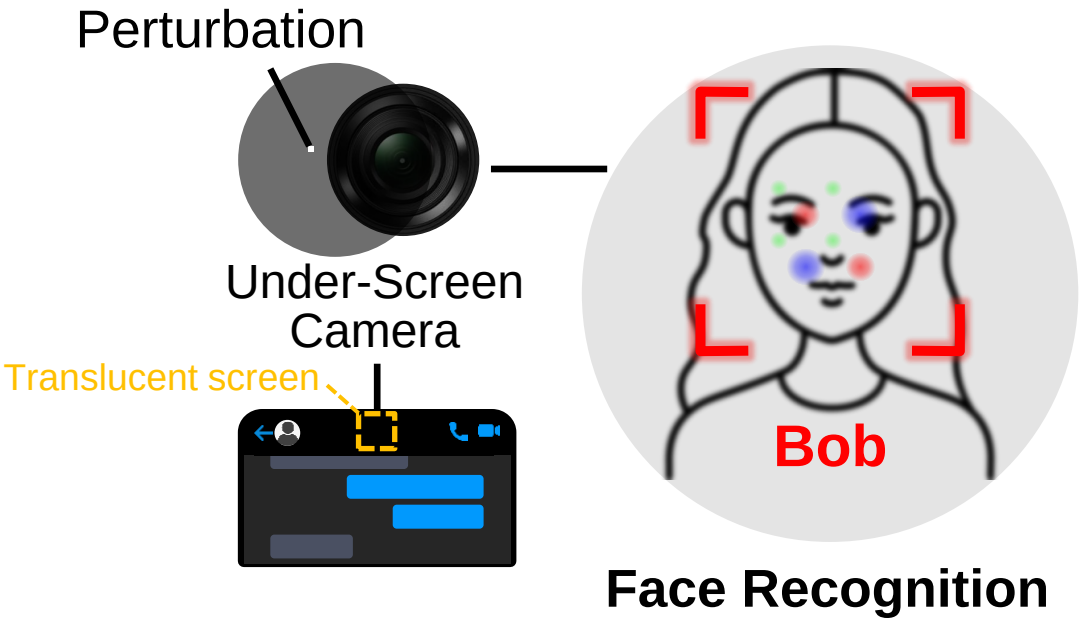
Normal Screen Region

Images formed in different scenarios



Screen perturbation



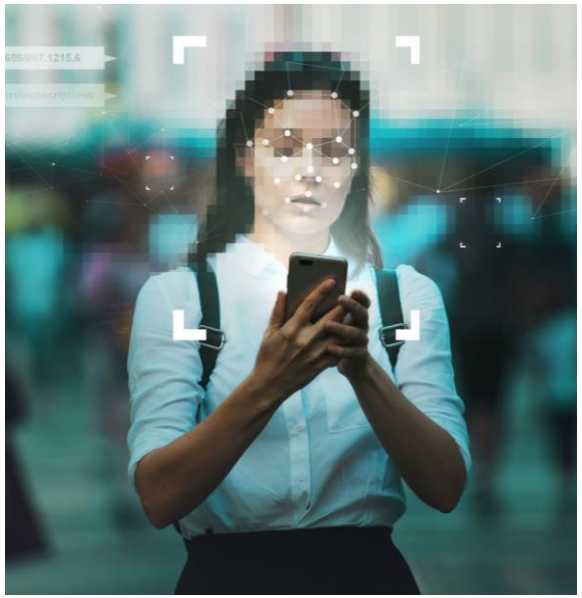


Defenders can proactively activate screen-pixels to thwart unauthorized ML models

Attackers can manipulate imperceptible screen-pixels to disrupt legitimate ML models

Screen perturbation: A software-defined perturbation that modifies the pixels displayed on the translucent screen region to nullify ML models

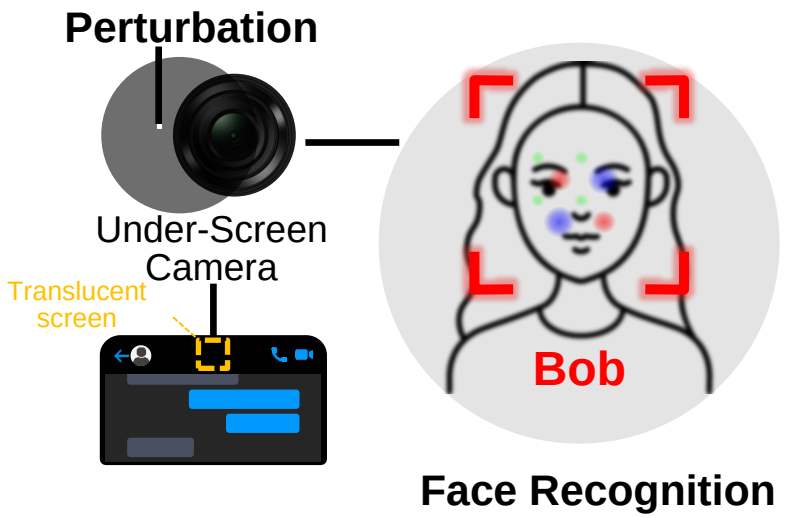
Key difference for screen perturbation



Privacy fears as MILLIONS of photos used to train facial recognition AI without users' consent *(Source: Daily Mail Online)*

Camera cover solution:

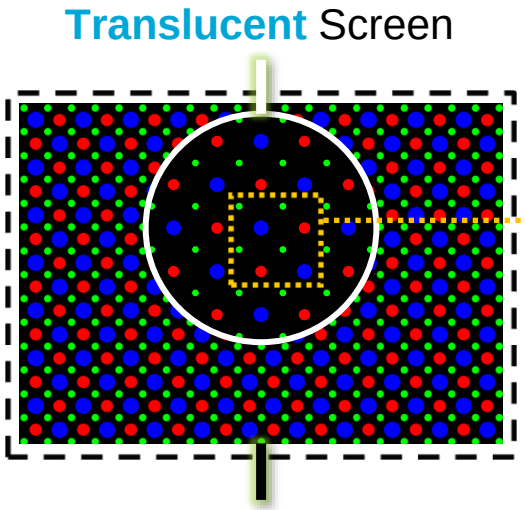
1) Secure but block everything
2) Obstruct the screen display



Proposed screen perturbation:

1) *Software-defined solution*
2) *Imperceptible changes*

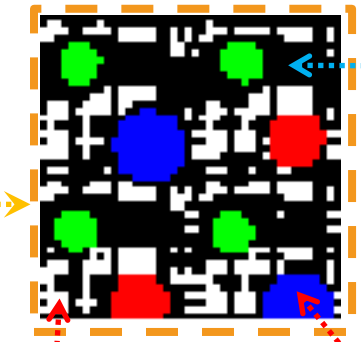
Image formation model of under-screen camera



Translucent Screen

Normal Screen Region

Screen-pixel unit



Control unit area

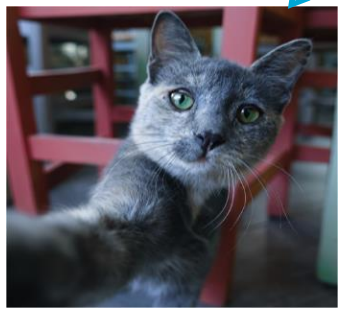
Screen-non-pixel area

RGB subpixel set

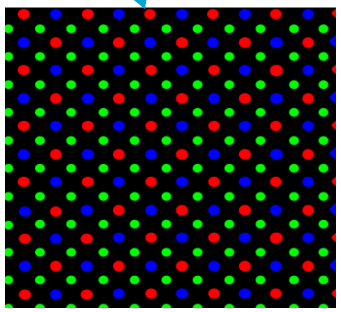
$$\boldsymbol{I} \oplus \boldsymbol{z} = \underbrace{\boldsymbol{I} \otimes k\left(F_p\right)}_{\text{Passive Perturbation}} + \underbrace{\boldsymbol{z} \otimes \sum\nolimits_c k\left(F_a\left(c\right)\right)}_{\text{Active Perturbation}} + noise, \quad c \in \{R, G, B\}$$

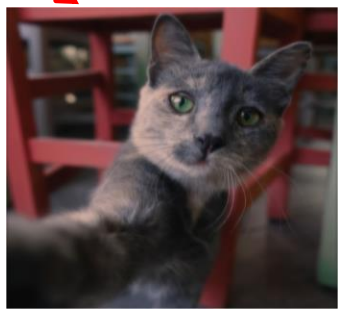
$k\left(F\right)$: Blur kernel

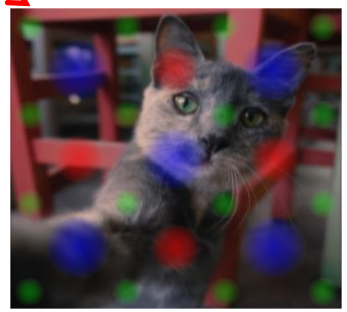
Original images



Lit-up screen-pixels

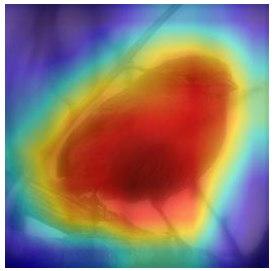




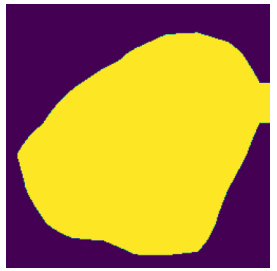


Chromaticity destruction

Q1: How to **localize** the region that has the highest influence on the decision making of the ML model?

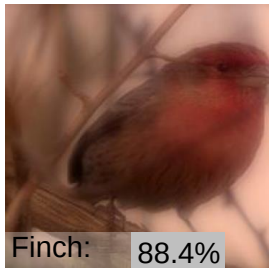


Grad-CAM heatmap



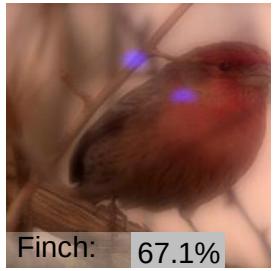
Attack region mask

Q2: Which **color** should we pick for increasing adversarial strength?



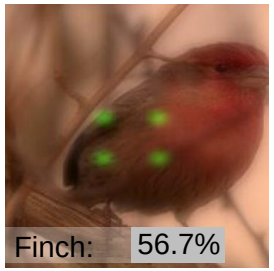
Passive perturbation

Red subpixel set

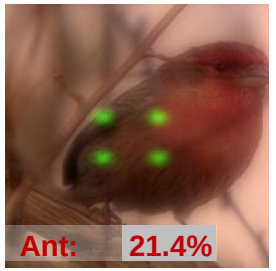


Blue subpixel set

Q3: **One** color or **Two** colors



Green, 30% bright



Green, 80% bright

G & B, 10% bright

Morphology destruction

New screen configuration to generate screen perturbation

Original screen configuration of no screen perturbation

Growth rate

$$s\left(\hat{l}, j\right) = \frac{\Phi\left(\boldsymbol{I} \oplus \boldsymbol{z}_{j,c,\boldsymbol{b}}, \hat{l}\right) - \Phi\left(\boldsymbol{I} \oplus \boldsymbol{z}, \hat{l}\right)}{\Phi\left(\boldsymbol{I} \oplus \boldsymbol{z}, \hat{l}\right)}$$

j : screen-pixel ***location***

$\boldsymbol{c}$: screen-pixel *color*

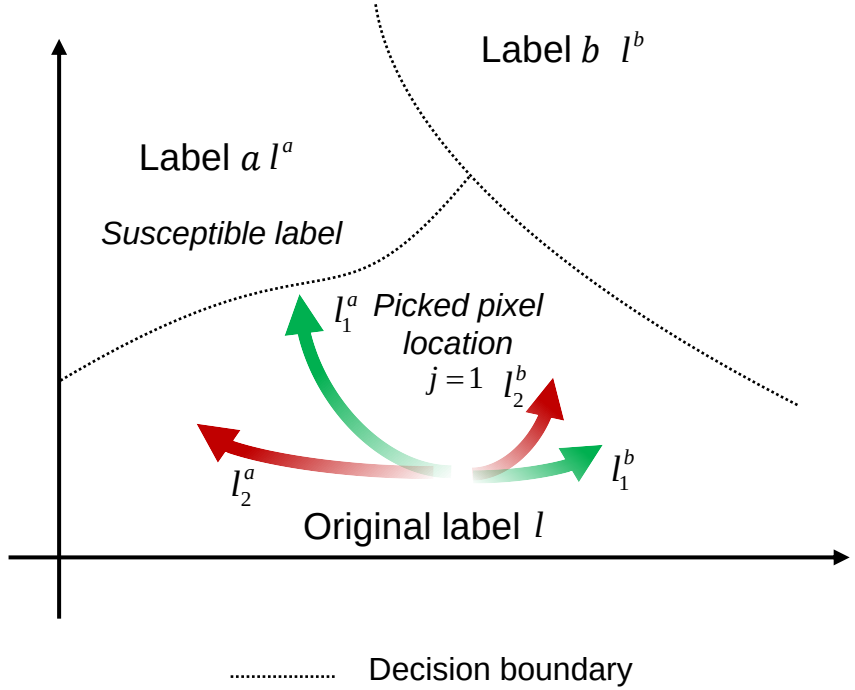
$\boldsymbol{b}$: screen-pixel *brightness*

The most susceptible label

$$\tilde{l}_j = \arg\max_{\hat{l}} s\left(\hat{l}, j\right)$$

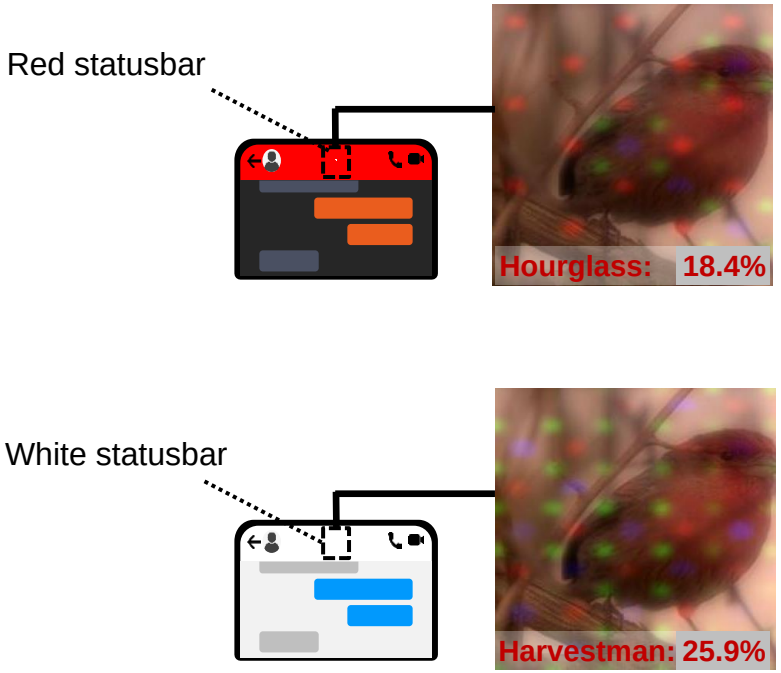
Picked screen-pixel location

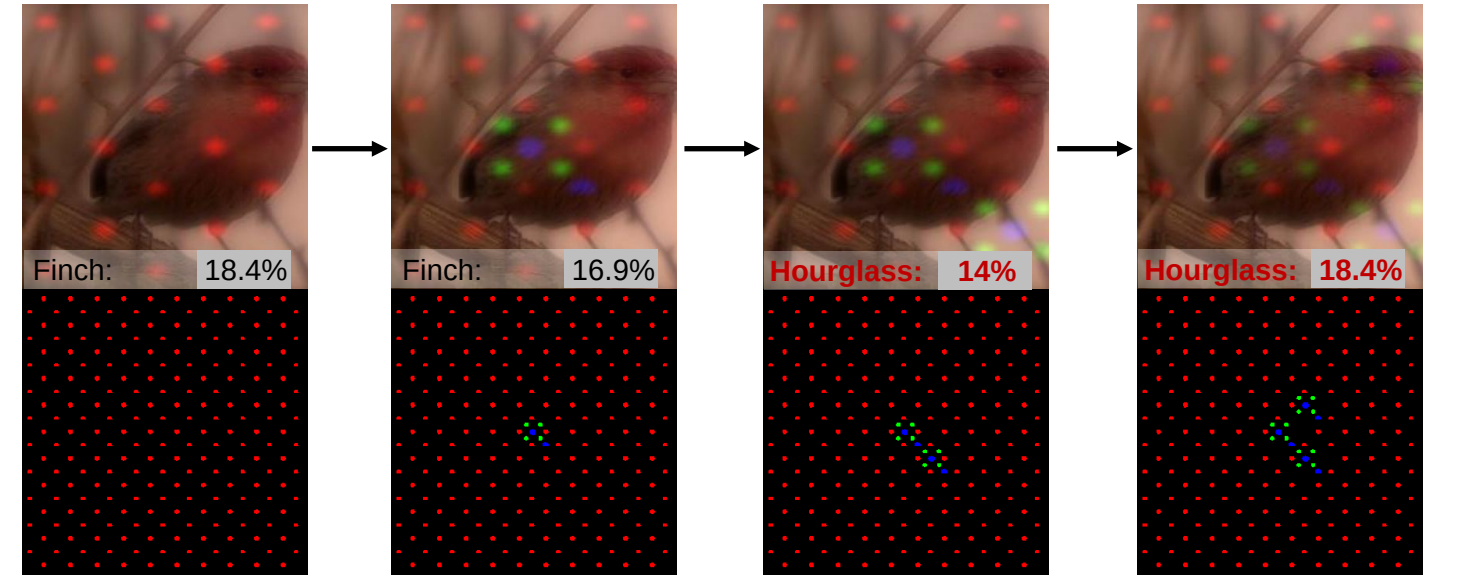
$$j = \arg\max_{\hat{j}} s\left(\tilde{l}_j, \hat{j}\right)$$



$\Phi\left(\boldsymbol{I}, l\right)$: Predicting probability of classifier Φ in classifying the image $\boldsymbol{I}$ with label l

Multiple-pixel perturbation





① Red statusbar: no pixels changed

② Red statusbar: **one** pixel changed

③ Red statusbar: **two** pixels changed

④ Red statusbar: **three** pixels changed

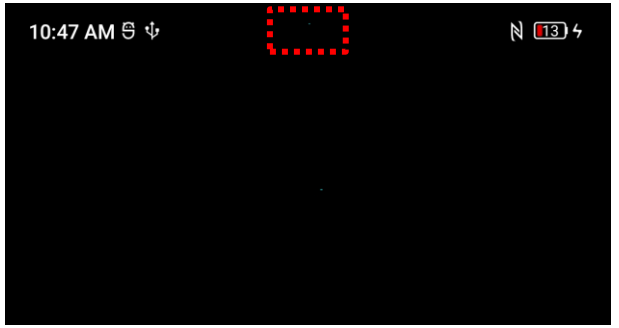
*Regional aggregation effect: **Adjacent** screen-pixel units have similar effect on the probabilities of predicted labels*

Built testbed & images results

ZTE AXON30

Xiaomi MIX4

Samsung Fold4



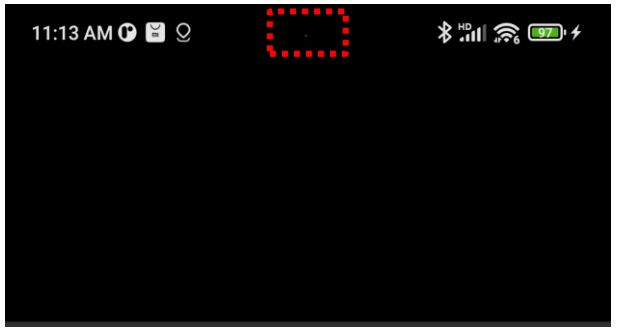


Image classification task

Clean images:

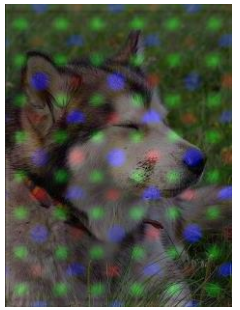
Blurred images (Passive):

One-pixel perturbed image

Multiple-pixel perturbed image

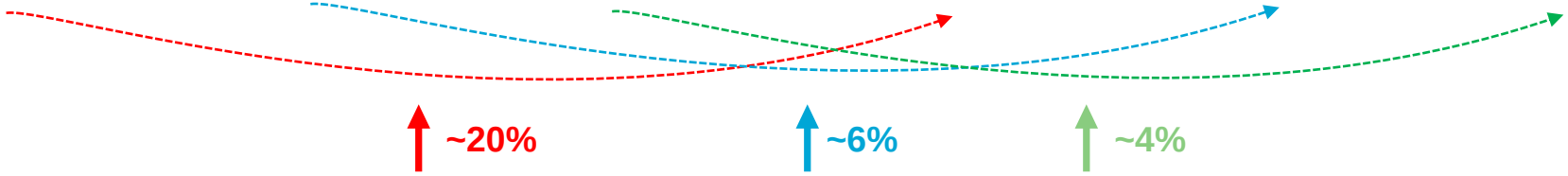
Deblurred images (Passive):

Deblurred One-pixel perturbed image



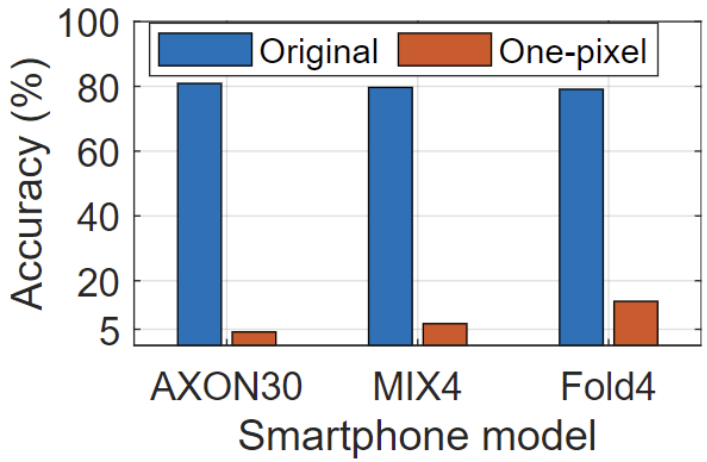
Deblurred Multiple-pixel perturbed image

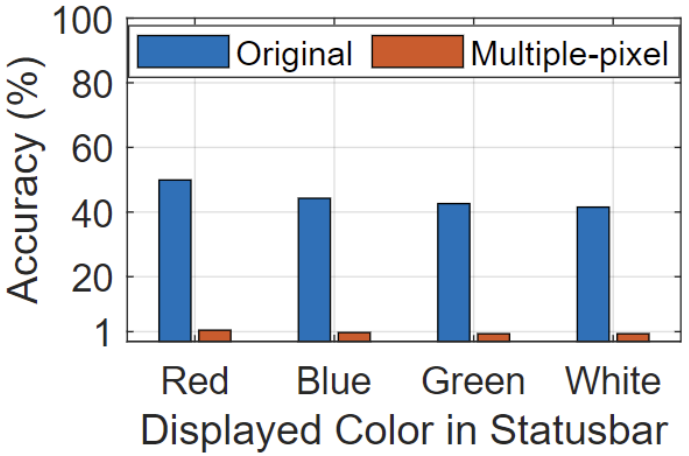
ResNet: *94.2%* *74.4%* *20.8%* *5.8%* *94.6%* *26.8%* *9.2%*



More results under different target models (MobileNet, ShuffleNet, IncepResNet) can be found in the paper

Face recognition task





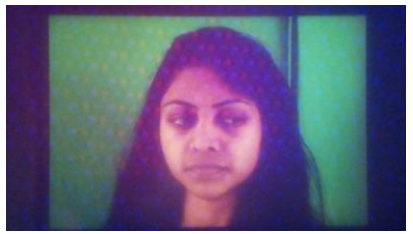
ZTE AXON30

Xiaomi MIX4

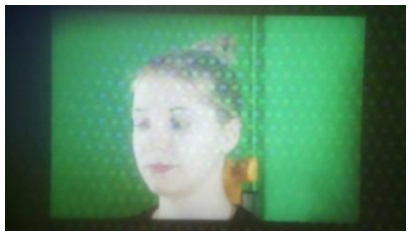
Samsung Fold4

Red statusbar

Green statusbar



Blue statusbar

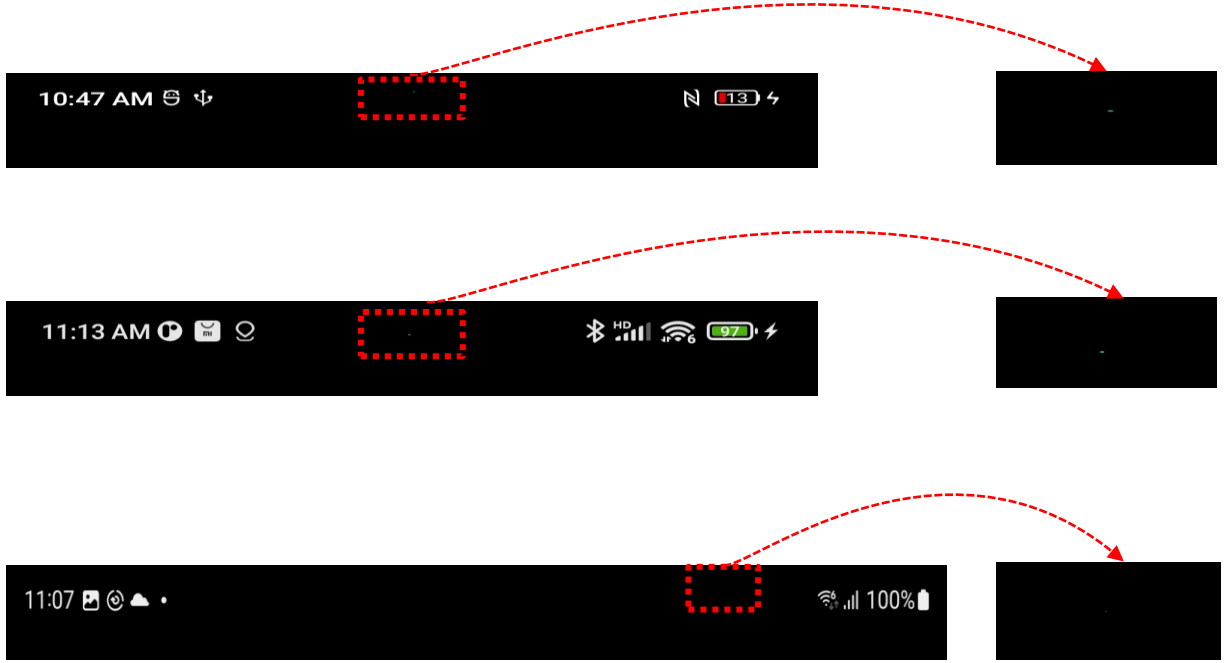


White statusbar

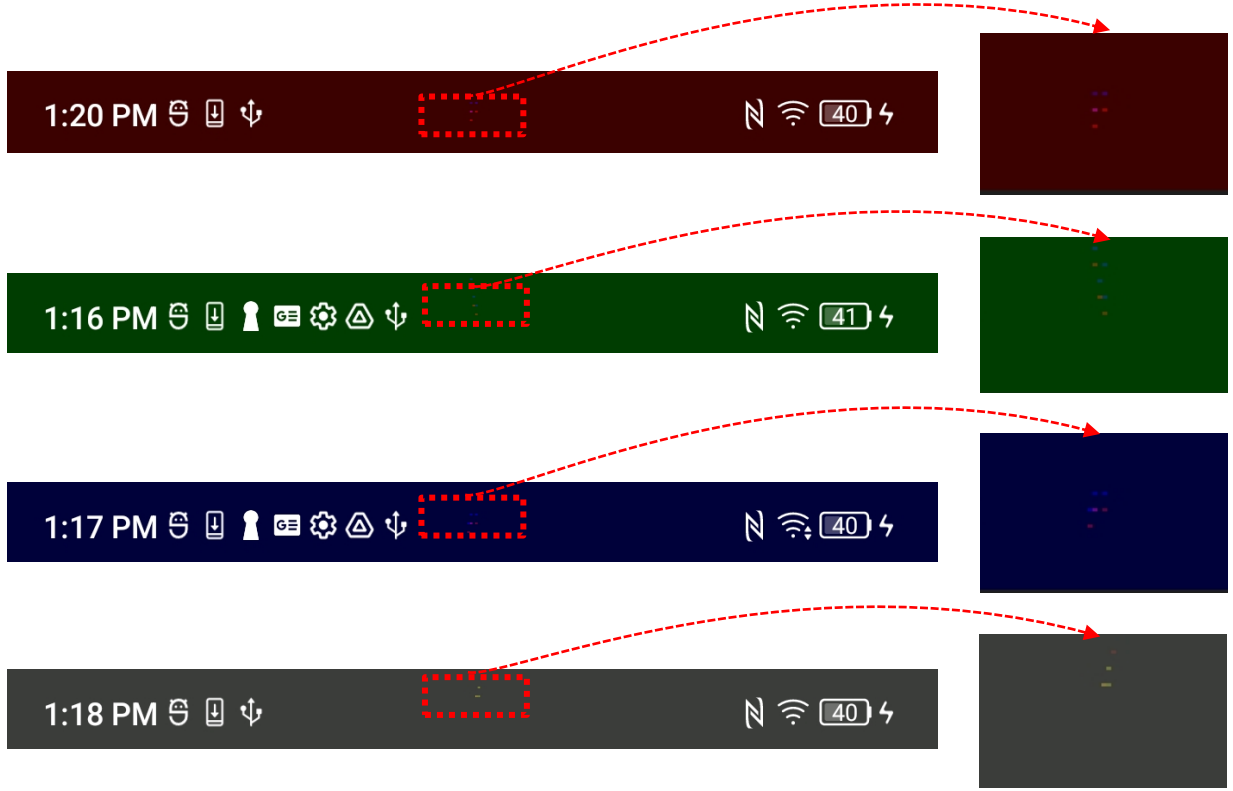
User study

30 Participants: 13 female and 17 male, aged between 20 and 45

One-pixel screen perturbation



Multiple-pixel screen perturbation



__No one__ notice any screen perturbation during the use of the smartphones!

Ineffective counter measures

Adversarial training

Training images
(1:1 clean: perturbed ratio)

Classifier

Clean images

94% → 87%

Perturbed images

<5% → <10%

Randomized smoothing

Training images
(with isotropic Gaussian noise)

Smoothed classifier

Majority vote

Clean images

94% → 94%

Perturbed images

<5% → <7%

This work has been funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement ENLIGHTEM No. 814215

THANK YOU

TU Delft

PennState