Looking into the dark: from image to video

Zibo Meng
Deep Learning Scientist
OPPO US R&D
Entered 40+ markets

Ranked #5 in 2020 Q1*

More than 350 million users

OPPO Research Institute (Headquarters · Shenzhen)

Dongguan Research Center

Beijing Research Center

Japan Research Center

Shanghai Research Center

US Research Center
What’s hidden in the dark
What’s hidden in the dark
What’s hidden in the dark

“Lux” is the standard unit of measure for illumination (i.e. brightness) of a surface at a given point.
“Lux” is the standard unit of measure for illumination (i.e. brightness) of a surface at a given point.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Illuminance</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct sunlight</td>
<td>~100K lux</td>
</tr>
<tr>
<td>daylight (non-direct sunlight)</td>
<td>~10K lux</td>
</tr>
<tr>
<td>dark (e.g. moonlight)</td>
<td>&lt; 1 lux</td>
</tr>
</tbody>
</table>
Traditional Camera Pipeline

- sub-optimal results due to low signal-to-noise ratio
- (low photon counts in the dark)

RAW → ISP: 
- Demosaicking
- White balance
- Color correction
- Gamma correction
- Tone mapping ← RGB
Imaging in the dark

Raise ISO sensitivity to gain brightness?
noise amplification in the electronic signal

Increase exposure time?
blur due to hand shake or object motion

Use flash?
reflections, glare, shadows
Let AI do the work

RAW → AI → RGB
What’s inside

AI
What’s inside

AI

Convolutional Neural Network

Loss Functions
A Convolutional Neural Network

A U-net: an encoder-decoder network
Works great for image-to-image translation, but yields results with color inconsistency.

Chen et. al. & Zamir et. al.
A Convolutional Neural Network

A U-net: an encoder-decoder network
Works great for image-to-image translation, but yields results with color inconsistency.

OPPO dark sight net
A Convolutional Neural Network

A U-net: an encoder-decoder network
Works great for image-to-image translation, but yields results with color inconsistency.

Chen et. al. & Zamir et. al.
A Convolutional Neural Network

A U-net: an encoder-decoder network
Works great for image-to-image translation, but yields results with color inconsistency.

Chen et al. & Zamir et al.

OPPO dark sight net
Loss Functions

Some Choices of Loss Functions

\[ L_1 = \frac{1}{N} \sum_{p=1}^{N} \left| \hat{y}_p - y_p \right| \]

\[ MSE = \frac{1}{N} \sum_{p=1}^{N} \left( \hat{y}_p - y_p \right)^2 \]

\[ L_{SSIM} = 1 - \frac{1}{N} \sum_{p=1}^{N} SSIM(\hat{y}_p, y_p) \]

\[ L_{MS-SSIM} = 1 - \frac{1}{N} \sum_{p=1}^{N} MS-SSIM(\hat{y}_p, y_p) \]

\[ L_{perceptual} = \frac{1}{N} \sum_{p=1}^{N} \left( \phi(\hat{y}_p) - \phi(y_p) \right)^2 \]
Loss Functions

\[ PSNR = 20 \cdot \log_{10} \left( \frac{\text{MaxPixelValue}}{\sqrt{\text{MSE}}} \right) \]

Simple to calculate

Has clear physical meanings

Does not correlate very well with human’s perceived visual quality
Loss Functions

Structural Similarity (SSIM)

Natural images are highly structured: pixels exhibit strong dependencies on each other

Sensitivity of human visual system (HVS) is related to:
- luminance (mean),
- contrast (variance),
- structure (covariance)

\[
SSIM(\hat{y}_p, y_p) = \text{luminance} \cdot \text{contrast} \cdot \text{structure} = l(\hat{y}_p, y_p) \cdot c(\hat{y}_p, y_p) \cdot s(\hat{y}_p, y_p) = \frac{2\mu_{\hat{y}_p} \cdot \mu_{y_p} + C_1}{\mu_{\hat{y}_p}^2 + \mu_{y_p}^2 + C_1} \cdot \frac{2\sigma_{\hat{y}_p} \cdot \sigma_{y_p} + C_2}{\sigma_{\hat{y}_p}^2 + \sigma_{y_p}^2 + C_2} \cdot \frac{\sigma_{\hat{y}_p, y_p}^2 + C_3}{\sigma_{\hat{y}_p} \cdot \sigma_{y_p} + C_3}
\]

\[
= \frac{2\mu_{\hat{y}_p} \cdot \mu_{y_p} + C_1}{\mu_{\hat{y}_p}^2 + \mu_{y_p}^2 + C_1} \cdot \frac{2\sigma_{\hat{y}_p} \cdot \sigma_{y_p} + C_2}{\sigma_{\hat{y}_p}^2 + \sigma_{y_p}^2 + C_2} \cdot \frac{\sigma_{\hat{y}_p, y_p}}{\sigma_{\hat{y}_p} \cdot \sigma_{y_p}} = l(\hat{y}_p, y_p) \cdot cs(\hat{y}_p, y_p)
\]

\[
C_1 = (K_1 \cdot \text{MaxPixelValue})^2 C_2 = (K_2 \cdot \text{MaxPixelValue})^2 C_3 = C_2/2
\]
Loss Functions

Multi-Scale Structural Similarity (MS-SSIM)

\[
SSIM(y, \hat{y}) = \frac{2\mu_y \mu_{\hat{y}} + C_1}{\mu_y^2 + \mu_{\hat{y}}^2 + C_1} \cdot \frac{2\sigma_{y\hat{y}} + C_2}{\sigma_y^2 + \sigma_{\hat{y}}^2 + C_2} = l(y, \hat{y}) \cdot cs(y, \hat{y})
\]

\[
C_1 = (K_1 \cdot \text{MaxPixelValue})^2 \quad C_2 = (K_2 \cdot \text{MaxPixelValue})^2
\]

\[
MS-SSIM(y_p, \hat{y}_p) = \left[ l_M(\hat{y}_p, y_p) \right]^{\gamma M} \cdot \prod_{k=1}^{M} \left[ cs_k(\hat{y}_p, y_p) \right]^\eta k
\]

MS-SSIM extends SSIM by computing variance and covariance components at M scales.

\( k^{th} \) scale image = sub-sampling original image by factor of 2 in both spatial dimensions for \((k-1)\) times

More flexible than single-scale SSIM: incorporated variations of image resolution and viewing conditions.
Loss Functions

**Perceptual Loss**

\[ L_{\text{perceptual}} = \frac{1}{N} \sum_{p=1}^{N} (\phi(y \hat{p}) - \phi(y_p))^2 \]

Measures the difference between deep feature representations of the output and ground-truth images, each extracted from a pre-trained neural network on ImageNet.

Enhances semantic similarity at deep feature representation level and serves as a perceptual metric.

Zhang et. al. arXiv:1801.03924, CVPR 2018
Loss Functions

Chen et. al. (arXiv:1805.01934):

\[ L_1 \text{(also tried } L_2 \text{ and } L_{SSIM} \text{ separately)} \]

Zamir et. al. (arXiv:1904.05939):

\[ \alpha (\beta L_1 + (1 - \beta) L_{MS-SSIM}) + (1 - \alpha) L_{perceptual} \quad \alpha = 0.9 \quad \beta = 0.99 \]

OPPO: \[ L_1 \quad L_{MS-SSIM} \quad ???? \]
Quantitative Results

\[
PSNR = 10 \cdot \log_{10} \left( \frac{MaxPixelValue^2}{MSE} \right) \quad MeanSquaredError(MSE) = \frac{1}{N} \sum_{p=1}^{N} \left( y_p - \hat{y}_p \right)^2
\]

<table>
<thead>
<tr>
<th>Camera Type</th>
<th>Sensor Type</th>
<th>Resolution</th>
<th>Chen et. al. (arXiv:1805.01934)</th>
<th>Zamir et. al. (arXiv:1904.05939)</th>
<th>OPPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sony 7SII camera</td>
<td>Bayer sensor</td>
<td>4240 x 2832</td>
<td>28.88</td>
<td>29.43</td>
<td>29.72</td>
</tr>
<tr>
<td>Fujifilm X-T2 camera</td>
<td>APS-C X-Trans sensor</td>
<td>6000 x 4000</td>
<td>26.61</td>
<td>27.63</td>
<td>28.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.787</td>
<td>-</td>
<td>0.795</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.68</td>
<td>-</td>
<td>0.722</td>
</tr>
</tbody>
</table>
Qualitative Results

Fujifilm X-T2 camera (APS-C X-Trans sensor, 6000 x 4000)
ISO = 6400
exposure time = 100ms
Qualitative Results

Chen et. al.  

Zamir et. al.  

OPPO  

Ground truth
Deliver to Cell Phones
Light up the darkness
Light up the darkness
Light up the darkness
More demonstrations
More demonstrations
More demonstrations
More demonstrations
More demonstrations

Ultra Dark

Night

1X
More demonstrations
Move to videos

Hard to produce paired data for videos

The algorithm should run in real time

The processed frames should be temporally consistent
Move to videos

Hard to produce paired data for videos
Adopted the model based on single images

The algorithm should run in real time

The processed frames should be temporally consistent
Move to videos

Hard to produce paired data for videos
Adopted the model based on single images

The algorithm should run in real time
The model should be extensively compressed

The processed frames should be temporally consistent
Move to videos

Hard to produce paired data for videos
Adopted the model based on single images

The algorithm should run in real time
The model should be extensively compressed

The processed frames should be temporally consistent
A temporal filtering approach should be employed
Thank you!