Compound Quality Mapping on High-resolution Images and Videos

Danda Pani Paudel & Zhiwu Huang
Computer Vision Lab @ ETH Zurich
Compound Quality Mapping

- Color Adjustment
- Illumination Enhancement
- Texture Sharpening
- etc.
Data Collection for Compound Image and Video Quality Mapping
Data Collection I: Paired Image Retouching (expensive expert effort)
Data Collection II: Weakly-paired Collection (expensive alignment)

<table>
<thead>
<tr>
<th>Camera</th>
<th>Sensor</th>
<th>Image size</th>
<th>Photo quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPhone 3GS</td>
<td>3 MP</td>
<td>2048 × 1536</td>
<td>Poor</td>
</tr>
<tr>
<td>BlackBerry Passport</td>
<td>13 MP</td>
<td>4160 × 3120</td>
<td>Mediocre</td>
</tr>
<tr>
<td>Sony Xperia Z</td>
<td>13 MP</td>
<td>2592 × 1944</td>
<td>Average</td>
</tr>
<tr>
<td>Canon 70D DSLR</td>
<td>20 MP</td>
<td>3648 × 2432</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

an overlapping region is determined by SIFT descriptor matching

non-linear transform and a crop resulting in two images of the same resolution representing the same scene
Data Collection II: Weakly-paired Video Collection? (more expensive)

Vid3oC Dataset

- Three cameras
- 82 recordings
- 328 videos
- Includes stereo depth

- Canon 5D Mark IV high quality DSLR
- Huawei P20 high-end smartphone
- ZED stereo camera
- AIM 2019 Video SR

The Vid3oC and IntVID Datasets for Video Super Resolution and Quality Mapping. Sohyeong Kim, Guanju Li, Dario Fuoli, Martin Danelljan, Zhiwu Huang, Shuhang Gu and Radu Timofte. ICCV 2019 Workshops.
Data Collection II: Unpaired Collection (Cheaper)

No Correspondence

Poor-quality Image Dataset

High-quality Image Dataset
Supervised Deep Learning Methods for Compound Image Quality Mapping
Deep Bilateral Learning for Real-Time Image Enhancement (HDRNet)

**Idea:** consumes a low-resolution version of the input image, followed by an edge-preserving upsampling to the full-resolution image in a bilateral filtering fashion

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{i} \| I_i - O_i \|^2$$
Deep Bilateral Learning for Real-Time Image Enhancement (HDRNet)

12 megapixel 16-bit linear input (tone-mapped for visualization)
tone-mapped with HDR+ 400 – 600 ms
processed with our algorithm 61 ms, PSNR = 28.4 dB

Underexposed Photo Enhancement (UPE)

**Idea:** learn an image-to-illumination (instead of image-to-image) mapping

Underexposed Photo Enhancement (UPE)

Loss function:

\[ \mathcal{L} = \sum_{i=1}^{N} \omega_r \mathcal{L}_r^i + \omega_s \mathcal{L}_s^i + \omega_c \mathcal{L}_c^i \]

Reconstruction 
\[ \mathcal{L}_r^i = \| I_i - S \ast \tilde{I}_i \|^2, \]
\[ \text{s.t. } (I_i)_c \leq (S)_c \leq 1, \forall \text{ pixel channel } c \]

Smoothness 
\[ \mathcal{L}_s^i = \sum_p \sum_c \omega_{x,c} (\partial_x S_p)_c^2 + \omega_{y,c} (\partial_y S_p)_c^2 \]

Color 
\[ \mathcal{L}_c^i = \sum_p \angle((\mathcal{F}(I_i))_p, \tilde{I}_i)_p \]

Underexposed Photo Enhancement (UPE)

(a) Input  
(b) JieP [4]  
(c) HDRNet [13]  
(d) DPE [9]  
(e) White-box [15]  
(f) Distort-and-Recover [22]  
(g) Our result  
(h) Expert-retouched

Visual Comparison on MIT-Adobe FiveK

Underexposed Photo Enhancement (UPE)

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDRNet [2]</td>
<td>21.96</td>
<td>0.866</td>
</tr>
<tr>
<td>DPE [1]</td>
<td>22.15</td>
<td>0.850</td>
</tr>
<tr>
<td>White-Box [3]</td>
<td>18.57</td>
<td>0.701</td>
</tr>
<tr>
<td>Distort-and-Recover [4]</td>
<td>20.97</td>
<td>0.841</td>
</tr>
<tr>
<td>Ours w/o $\mathcal{L}_r$, w/o $\mathcal{L}_s$, w/o $\mathcal{L}_c$</td>
<td>21.97</td>
<td>0.867</td>
</tr>
<tr>
<td>Ours with $\mathcal{L}_r$, w/o $\mathcal{L}_s$, w/o $\mathcal{L}_c$</td>
<td>22.31</td>
<td>0.871</td>
</tr>
<tr>
<td>Ours with $\mathcal{L}_r$, with $\mathcal{L}_s$, w/o $\mathcal{L}_c$</td>
<td>22.89</td>
<td>0.884</td>
</tr>
<tr>
<td>Ours</td>
<td>23.04</td>
<td>0.893</td>
</tr>
</tbody>
</table>

Quantitative Comparison on MIT-Adobe FiveK

DSLR Photo Enhancement (DSLR-PE)

Idea: learn the translation function using a residual convolutional neural network with a composite perceptual error function that combines content, color and adversarial texture losses

\[ \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{content}} + 0.4 \cdot \mathcal{L}_{\text{texture}} + 0.1 \cdot \mathcal{L}_{\text{color}} + 400 \cdot \mathcal{L}_{\text{tv}} \]

\[ \mathcal{L}_{\text{color}}(X, Y) = \| X_b - Y_b \|^2_2 \]

\[ \mathcal{L}_{\text{content}} = \frac{1}{C_jH_jW_j} \| \psi_j(F_W(I_s)) - \psi_j(I_t) \| \]

\[ \mathcal{L}_{\text{tv}} = \frac{1}{CHW} \| \nabla_x F_W(I_s) + \nabla_y F_W(I_s) \| \]

\[ \mathcal{L}_{\text{texture}} = - \sum_i \log D(F_W(I_s), I_t) \]

DSLR Photo Enhancement (DSLR-PE)

Typical artifacts generated by our method (bottom) compared with original iPhone images (top).

Weakly-Supervised Deep Learning Methods for Compound Image Quality Mapping
Weakly Supervised Photo Enhancer (WESPE)

- Content Consistency Loss

$$\mathcal{L}_{\text{content}} = \frac{1}{C_j H_j W_j} \| \psi_j(x) - \psi_j(\tilde{x}) \|,$$

- Adversarial Color Loss

$$\mathcal{L}_{\text{color}} = - \sum_i \log D_c(G(x)_b).$$

- Adversarial Texture Loss

$$\mathcal{L}_{\text{texture}} = - \sum_i \log D_t(G(x)_g).$$

- Total Variation Loss

$$\mathcal{L}_{\text{tv}} = \frac{1}{CHW} \| \nabla_x G(x) + \nabla_y G(x) \|,$$

WESPE: weakly supervised photo enhancer for digital cameras, Ignatov et al., CVPRW 2018.
Weakly Supervised Photo Enhancer (WESPE)

<table>
<thead>
<tr>
<th>DPED images</th>
<th>APE</th>
<th>Weakly Supervised</th>
<th>Fully Supervised [13]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>WESPE [DIV2K]</td>
<td>WESPE [DPED]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>iPhone</td>
<td></td>
<td>17.28</td>
<td>0.86</td>
</tr>
<tr>
<td>BlackBerry</td>
<td></td>
<td>18.91</td>
<td>0.89</td>
</tr>
<tr>
<td>Sony</td>
<td></td>
<td>19.45</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*WESPE: weakly supervised photo enhancer for digital cameras, Ignatov et al., CVPRW 2018.*
Deep Photo Enhancer (DPE)

-Identity Mapping Loss:

\[ I = \mathbb{E}_{x, y'} \left[ MSE(x, y') \right] + \mathbb{E}_{y, x'} \left[ MSE(y, x') \right]. \]

-Cycle-Consistency Loss:

\[ C = \mathbb{E}_{x, x''} \left[ MSE(x, x'') \right] + \mathbb{E}_{y, y''} \left[ MSE(y, y'') \right]. \]

-Adversarial Loss:

\[ A_D = \mathbb{E}_x \left[ D_X(x) \right] - \mathbb{E}_{x'} \left[ D_X(x') \right] + \mathbb{E}_y \left[ D_Y(y) \right] - \mathbb{E}_{y'} \left[ D_Y(y') \right], \]
\[ A_G = \mathbb{E}_{x'} \left[ D_X(x') \right] + \mathbb{E}_{y'} \left[ D_Y(y') \right]. \]

Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs, Chen et al., CVPR 2018.
Deep Photo Enhancer (DPE)

Preference matrix from AMT user study

<table>
<thead>
<tr>
<th></th>
<th>CycleGAN</th>
<th>DPED</th>
<th>NPEA</th>
<th>CLHE</th>
<th>ours</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CycleGAN</td>
<td>-</td>
<td>32</td>
<td>27</td>
<td>23</td>
<td>11</td>
<td>93</td>
</tr>
<tr>
<td>DPED</td>
<td>368</td>
<td>-</td>
<td>141</td>
<td>119</td>
<td>29</td>
<td>657</td>
</tr>
<tr>
<td>NPEA</td>
<td>373</td>
<td>259</td>
<td>-</td>
<td>142</td>
<td>50</td>
<td>824</td>
</tr>
<tr>
<td>CLHE</td>
<td>377</td>
<td>281</td>
<td>258</td>
<td>-</td>
<td>77</td>
<td>993</td>
</tr>
<tr>
<td>ours</td>
<td>389</td>
<td>371</td>
<td>350</td>
<td>323</td>
<td>-</td>
<td>1433</td>
</tr>
</tbody>
</table>

Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs, *Chen et al., CVPR 2018.*
Limitation for Compound Quality Mapping and High-Resolution Image Treatment

<table>
<thead>
<tr>
<th>Model</th>
<th>Limitation (Compound Quality)</th>
<th>Limitation (High Resolution)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WESPE</td>
<td>Color and Texture (Not sufficient)</td>
<td>Patch-wise Enhancement</td>
</tr>
<tr>
<td>DPE</td>
<td>No consideration on mixed-perceptual improvement</td>
<td>Down-scaling</td>
</tr>
</tbody>
</table>
Divide-and-Conquer Adversarial Learning for High-resolution Image and Video Enhancement

Divide-and-Conquer Inspired Method

Step 1#: divide

Step 2#: conquer

Step 3#: combine

Mixed-Perception

Divide-and-Conquer Inspired Method

Additive

Multiplicative
Network Design

(a) Enhancer for Perception-based Division

Multiplicative connection: \( E(x) = x \cdot s^{-1}_x \)

Global concatenation

Local concatenation

Additive connection: \( E(x) = x + r_x \)

(b) Discriminator for Freq- and Dim-based Division

Frequency-based input

Dim-based approximation

\( F^M \)

\( \frac{H}{2} \times \frac{W}{2} \)

\( D^1_i \)
Perception-based Division

(a) Input
(b) Additive component
(c) Multiplicative component
(d) Additive map
(e) Multiplicative map
(f) Fused Map
Frequency-based Division

Dimension-based Division and Optimization

(1) project n-dim PDFs to 1-dim marginal PDFs by orthogonal matrix

(2) target 1-dim marginal PDF
source 1-dim marginal PDF

(3) match n-dim PDFs by comparing 1-dim marginal PDFs

target n-dim PDF
source n-dim PDF
Loss Design

Adaptive SWGAN loss

\[
\min \max_E \int_{\theta \in \mathbb{R}^{n-1}} \left( \mathbb{E}_{y \sim P_y}[C(y)] - \mathbb{E}_{\hat{y} \sim P_{\hat{y}}}[C(E(x))] \right) + \lambda \mathbb{E}_{y \sim P_y} \left[ \max(0, \|\nabla_{\hat{y}} C(\hat{y})\|_2 - 1) \right]
\]

\[
\nabla_{\hat{y}} C(\hat{y}) = \eta \nabla_{\hat{y}} C(\hat{y}) + (1 - \eta) \frac{\nabla_{\hat{y}} C(\hat{y})}{\lambda}
\]

Adaptive Penalty
Evaluation on Toy Data

2.5k iterations

5k iterations

(a) WGAN  (b) AdaWGAN  (c) SWGAN  (d) Proposed AdaSWGAN
High-resolution Issue for Image Enhancement

Input

Downscaling (low-res, noisy, blurry)
Deep Photo Enhancer (DPE) [Chen et al in CVPR’18]

Patch-wise Enhancement (spatial inconsistency)
Weakly Supervised Photo Enhancer (WESPE) [our work in CVPR’18 workshop]

Multi-scale Photo Enhancement (MUSPE) Our current work
Multi-scale Extension of DACAL for Image Enhancement

(a) high-scale enhancer $E_h$

(b) high-scale discriminator $C_h$

![Diagram showing the multi-scale extension of DACAL for image enhancement]
Table 1: PSNR and SSIM results for the MIT-Adobe FiveK [42] test images. Here, WB and DR indicate the White-Box and Distort-and-Recover methods, respectively. $\text{MUSPE}_{l_1}$, $\text{MUSPE}_{l_2}$, $\text{MUSPE}_{l_3}$ and $\text{MUSPE}_l$ represent the use of individual additive, individual multiplicative, multiplicative cascaded by additive, and our suggested parallel fusion (two-stream strategy), respectively. $\text{MUSPE}_h$ is our higher-scale version. $\text{PSNR}_d/\text{SSIM}_d$ and $\text{PSNR}_f/\text{SSIM}_f$ indicate the results on downscaled images and full-resolution images, respectively.

<table>
<thead>
<tr>
<th></th>
<th>WB</th>
<th>DR</th>
<th>DPED</th>
<th>DPE</th>
<th>$\text{MUSPE}_{l_1}$</th>
<th>$\text{MUSPE}_{l_2}$</th>
<th>$\text{MUSPE}_{l_3}$</th>
<th>$\text{MUSPE}_l$</th>
<th>$\text{MUSPE}_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{PSNR}_d$</td>
<td>18.86</td>
<td>21.64</td>
<td>21.05</td>
<td>22.10</td>
<td>22.73</td>
<td>22.99</td>
<td>23.01</td>
<td>23.52</td>
<td>24.15</td>
</tr>
<tr>
<td>$\text{PSNR}_f$</td>
<td>19.09</td>
<td>21.52</td>
<td>20.86</td>
<td>21.65</td>
<td>22.43</td>
<td>22.69</td>
<td>23.02</td>
<td>23.56</td>
<td>24.07</td>
</tr>
<tr>
<td>$\text{SSIM}_d$</td>
<td>0.928</td>
<td>0.936</td>
<td>0.922</td>
<td>0.947</td>
<td>0.958</td>
<td>0.942</td>
<td>0.949</td>
<td>0.959</td>
<td>0.962</td>
</tr>
<tr>
<td>$\text{SSIM}_f$</td>
<td>0.920</td>
<td>0.922</td>
<td>0.916</td>
<td>0.894</td>
<td>0.948</td>
<td>0.942</td>
<td>0.940</td>
<td>0.954</td>
<td>0.956</td>
</tr>
</tbody>
</table>

Table 2: PSNR and SSIM results for the DPED [14] test $100 \times 100$ image patches. Here, $l$, $f$, $d$ for MUSPE represent the use of our proposed sliced-perception, sliced-frequency and sliced-dimension learning respectively. $\text{MUSPE}_h$ is our higher-scale version.

<table>
<thead>
<tr>
<th></th>
<th>WESPE</th>
<th>DPE</th>
<th>$\text{MUSPE}_l$</th>
<th>$\text{MUSPE}_{l+f}$</th>
<th>$\text{MUSPE}_{l+f+d}$</th>
<th>$\text{MUSPE}_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{PSNR}_{100}$</td>
<td>17.45</td>
<td>18.53</td>
<td>19.62</td>
<td>20.01</td>
<td>20.43</td>
<td>20.90</td>
</tr>
<tr>
<td>$\text{SSIM}_{100}$</td>
<td>0.854</td>
<td>0.861</td>
<td>0.868</td>
<td>0.869</td>
<td>0.872</td>
<td>0.874</td>
</tr>
</tbody>
</table>
Input

WESPE [Ignatov, CVPRW’18]

DPE [Chen, CVPR’18]

Proposed MUSPE [ICLR’20 submission]
Input

WESPE [Ignatov, CVPRW’18]

DPE [Chen, CVPR’18]

Proposed MUSPE [ICLR’20 submission]
Video Quality Mapping = Image Quality Mapping + Temporal Smoothing
Recurrent Extension of DACAL for Video Enhancement

(a) recurrent extension of Generator

(b) recurrent extension of Discriminator
Perframe-DACAL

Recurrent-DACAL (fine-tuned on Retouched&DSLR images)
Perframe-DACAL

Recurrent-DACAL (fine-tuned on Retouched&DSLR images)
Conclusion

- Supervision
  - Weak-supervision is cheaper
- Compound quality mapping
  - Divide-and-conquer inspired algorithm is promising
- High-resolution image treatment
  - Multiscaled training is helpful
- Video enhancement
  - Recurrent model works well