Deep Low-Contrast Image Enhancement using Structure Tensor Representation

Hyungjoo Jung,[1,2] **Hyunsung Jang,**[1,3] **Namkoo Ha,**[3] **and Kwanghoon Sohn**[1*]
[1] Yonsei University
[2] Korea Institute of Science and Technology (KIST)
[3] LIG Nex1
jhj0220@kist.re.kr, hyunsung.jang@yonsei.ac.kr, namkoo.ha@lignex1.com, khsohn@yonsei.ac.kr

Abstract

We present a new deep learning framework for low-contrast image enhancement, which trains the network using the multi-exposure sequences rather than explicit ground-truth images. The purpose of our method is to enhance a low-contrast image so as to contain abundant details in various exposure levels. To realize this, we propose to design the loss function using the structure tensor representation, which has been widely used as high-dimensional image contrast. Our loss function penalizes the difference of the structure tensors between the network output and the multi-exposure images in a multi-scale manner. Eventually, the network trained by the loss function produces a high-quality image approximating the overall contrast of the sequence. We provide in-depth analysis on our method and comparison with conventional loss functions. Quantitative and qualitative evaluations demonstrate that the proposed method outperforms the existing state-of-the art approaches in various benchmarks.

Introduction

Camera sensors often fail to capture clear images or videos under poor conditions, such as insufficient lighting environment, limited camera performance, and inappropriate setting for the equipment. The lost details and low contrast not only cause unpleasant subjective feelings, but also degenerate the performance of high-level vision algorithms such as detection and stereo matching. The demand for low-contrast image enhancement has tremendously grown for both human and machine perceptions. Over the last decades, various methods have been proposed to enhance the low-contrast images. Traditional enhancement techniques can be mainly categorized into two groups: histogram-based (Arici, Dikbas, and Altunbasak 2009; Celik and Tjahjadi 2011) and Retinex-based (Fu et al. 2016; Guo, Li, and Ling 2017) methods. However, they require manual parameter adjustments, and fail to produce a high-quality results due to the complex structures in natural scenes.

Recently, convolutional neural networks (CNNs) have shown a massive advance in many low-level vision tasks (Dong et al. 2015; Kim et al. 2017; Kupyn et al. 2018), due to their powerful representation ability to image structure. They utilize a set of degraded and ground-truth image pairs as training dataset, where the degraded images are usually obtained by corrupting the clean ones. With the paired training examples, the deep networks can be trained to learn a high-dimensional non-linear mapping from the degraded observations to their ground-truths. The most common choice of loss functions is mean square error (MSE) or mean absolute error (MAE), which measure per-pixel intensity difference between CNN outputs and the ground-truths. However, these loss functions suffer from poor results due to their independent property on local characteristics of the image. Beyond the pointwise quality metrics, several loss functions (Zhao et al. 2016; Dosovitskiy and Brox 2016) have considered the neighboring pixels to obtain perceptually enhanced outputs. Especially, the perceptual loss (Dosovitskiy and Brox 2016) ensures the output image to have the similar high-level feature with that of the ground-truth image.

In the area of single image contrast enhancement, we can train the CNNs with the above loss functions to map the low-contrast input into the corresponding reference image. The paired training dataset has been constructed by retouching low-contrast images through human experts (Bychkovsky et al. 2011) or photography algorithms (Cai, Gu, and Zhang 2018). However, there may be no unique or well-defined high-quality ground-truth given a low-contrast image. Furthermore, the results obtained by the direct network suffer from various distortions, such as noise amplification, false color, and wrong contrast. Such distortions come from the difficulties in balancing both the contrast enhancement and color reconstruction with single ground-truth images. The recent CNN-based enhancement approaches have resolved the distortions assisted by additional networks (Cai, Gu, and Zhang 2018; Ren et al. 2019), refinement (Wei et al. 2018), or adversarial training (Ignatov et al. 2017).

In this paper, we introduce a novel deep learning framework for single image contrast enhancement, which uses the multi-exposure image sequences, not an explicit high-quality image, as weak supervision for network training. Although each exposure image has only weak contrast, the entire sequences can provide a rich details for contrast enhancement. Our goal is to enhance the low-contrast images using abundant details in the multi-exposure sequence. To realize this, we first define the contrast of the sequence using

*Corresponding author

the structure tensor representations (Di Zenzo 1986), which have been widely used as the fusion strategy of variational fusion methods (Socolinsky and Wolff 2002; Gu et al. 2012; Connah, Drew, and Finlayson 2014). Different from variational image fusion methods, we propose to design the loss function using the structure tensor. The loss function minimizes the structure tensor difference between the network output and the multi-exposure sequences. Eventually, our method can produce high-quality results approximating the overall contrast of the multiple images.

Our main contributions can be summarized as follows:

- We propose a multi-scale structure tensor loss function for single image contrast enhancement. The proposed loss function effectively enforces the contrast of enhanced result to approximate the overall contrast of multi-exposure image sequences.
- We provide a performance comparison of the structure tensor loss with other functions, such as MAE, SSIM, and perceptual losses. Furthermore, we show that the proposed loss function is more suitable for single image enhancement than image fusion.
- The experimental results demonstrate the effectiveness of our method in various low-contrast benchmarks, such as low-illumination, under-exposure, and low-end device.

Related Works

Single Image Contrast Enhancement

Traditional contrast enhancement techniques have been classified into two categories: histogram-based (Arici, Dikbas, and Altunbasak 2009; Celik and Tjahjadi 2011) and Retinex-based (Fu et al. 2016; Guo, Li, and Ling 2017) methods. Histogram-based methods redistribute the luminous intensity distribution in global (Arici, Dikbas, and Altunbasak 2009) or local (Celik and Tjahjadi 2011) manner to improve the image contrast. Such simple operation, however, produces unrealistic contrast distortions in the enhanced images due to the lack of structural information. Retinex-based methods (Fu et al. 2016; Guo, Li, and Ling 2017) decompose the input image into reflectance and illumination components based on Retinex theory (Land 1977), and adopt different strategies to enhance each component. However, Retinex-based methods usually destroy the naturalness of images, since the physical model in Retina theory ignores the camera response characteristics.

Recently, the CNNs have been applied to solve the contrast enhancement problem (Cai, Gu, and Zhang 2018; Wei et al. 2018; Ignatov et al. 2017). Motivated by Retinex theory, Wei *et al.* (Wei et al. 2018) proposed the Retinex-Net that consists of decomposition and enhancement modules. Since the enhancement module also amplifies the noise, handcrafted denoising is additionally performed to refine the reflectance. Cai *et al.* (Cai, Gu, and Zhang 2018) proposed a two-stage CNN architecture for contrast enhancement. The first stage enhances the image contrast, followed by an additional refinement to relieve the distortions from the first stage. Ren *et al.* (Ren et al. 2019) proposed an edge stream using recurrent neural network (RNN) to capture additional edge details. Wang *et al.* (Wang et al. 2019) designed the network to estimate illumination map rather than enhanced image directly. The CNN-based methods train the networks using conventional reconstruction losses, i.e., MAE, SSIM, and perceptual losses, with explicit high-quality ground-truth labels. In contrast to the previous works, we propose a new loss function to directly measure the contrast difference between the results and multi-exposure sequence. Our method can generate high-quality enhanced result through a single forward network, which does not require additional refinement or networks.

Variational Image Fusion

Image fusion is a process to generate a fused image containing the important features from the multiple images. Based on the fact that local brightness change is highly related to human perception, variational fusion methods transfer the gradient information from the source images to the fused image. The fundamental process is to compare the local structure among different dimensional images, i.e., source and fused images. Structure tensor (Di Zenzo 1986), which is also called the first fundamental form in differential geometry, is a powerful method to simultaneously describe the local structure of high-dimensional images. Socolinsky and Wolff (Socolinsky and Wolff 2002) first proposed to use the structure tensor to visualize the multi-spectral images into gray-scale image. Gu *et al.* (Gu et al. 2012) applied the structure tensor into the multi-exposure image fusion. The spectral edge (SpE) fusion (Connah, Drew, and Finlayson 2014) extended more general fusion approach, which generates a naturalistic color image from the multi-spectral images. Contrary to these fusion methods, our method uses the structure tensor to enhance the contrast of single image.

Background and Motivation

Structure Tensor Representation

Gradient domain processing has attracted significant interest based on the fact that local brightness changes are commonly related to visually relevant features. However, when dealing with high-dimensional images, the concept of local contrast should be redefined to be applicable to various dimensionality. Structure tensor representation (Di Zenzo 1986) is a powerful tools to describe gradient information across the multi-channel image. The gradient of an N-dimensional image $\mathbf{I}$ at a single pixel (x, y) is defined by the following Jacobian matrix $\nabla\mathbf{I}$

$$\nabla\mathbf{I}(x,y) = \begin{bmatrix} D_1 I_1(x,y) & D_2 I_1(x,y) \\ \vdots & \vdots \\ D_1 I_N(x,y) & D_2 I_N(x,y) \end{bmatrix}, \quad (1)$$

where I_n is the n^{th} gray-level image of $\mathbf{I}$. $D_1 I_n$ and $D_2 I_n$ indicate the gradients of I_n with respect to horizontal and vertical directions, respectively. The gradient in the direction of $\mathbf{v} = [\cos\theta \;\; \sin\theta]^T$ is given by $\nabla\mathbf{I}\mathbf{v}$. Assuming Euclidean metric, we define the contrast of $\mathbf{I}$ at (x, y) in the direction of $\mathbf{v}$ as the squared magnitude of $\nabla\mathbf{I}\mathbf{v}$ as follows:

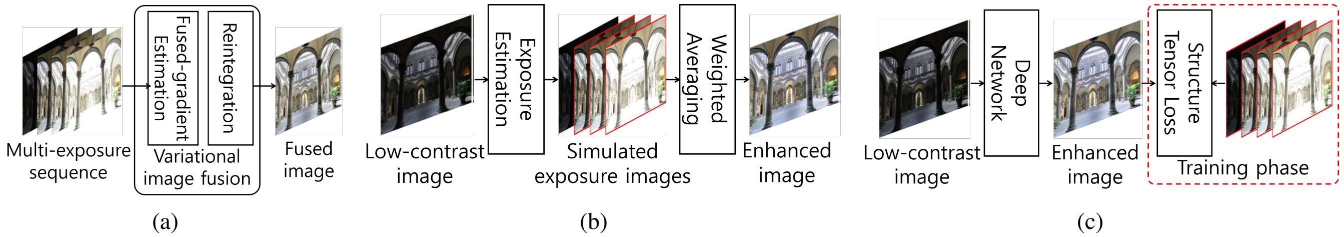


Figure 1: The framework of (a) variational image fusion (Socolinsky and Wolff 2002; Connah, Drew, and Finlayson 2014), (b) exposure estimation based contrast enhancement (Ying, Li, and Gao 2017; Ren et al. 2018; Zhang, Nie, and Zheng 2019), and (c) our contrast enhancement.

$$\mathscr{Z}_{\mathbf{I},\mathbf{v}}(x,y) = \mathbf{v}^T(\nabla\mathbf{I})^T(\nabla\mathbf{I})\mathbf{v}. \tag{2}$$

The 2×2 matrix $\mathbf{Z}_\mathbf{I}(x,y) = (\nabla\mathbf{I})^T(\nabla\mathbf{I})$ is the structure tensor, which summarizes the combined derivative structure of the multi-channel image as follows:

$$\mathbf{Z}_\mathbf{I}(x,y) = \begin{bmatrix} \sum\limits_{n=1}^{N} \left(D_1 I_n(x,y)\right)^2 & \sum\limits_{n=1}^{N} \left(D_1 I_n(x,y)\right)\left(D_2 I_n(x,y)\right) \\ \sum\limits_{n=1}^{N} \left(D_2 I_n(x,y)\right)\left(D_1 I_n(x,y)\right) & \sum\limits_{n=1}^{N} \left(D_2 I_n(x,y)\right)^2 \end{bmatrix}. \tag{3}$$

Since the structure tensor of (3) is a symmetric matrix with real values, it has two real and non-negative eigenvalues. The eigenvectors of the structure tensor indicate the direction of maximal and minimal contrast of (2), and the corresponding eigenvalues denote the rates of change.

Motivation

Once the contrast is defined for an arbitrary dimension of images, it becomes natural to apply the structure tensor into image fusion. The fundamental idea behind structure tensor based image fusion is to find the fused image whose structure tensor is identical to that of source images. The fusion methods sequentially perform the gradient-fusion and reintegration (Fig. 1(a)). Socolinsky and Wolff (Socolinsky and Wolff 2002) defined a scalar gradient field by multiplying the maximum eigenvector of $\mathbf{Z}_\mathbf{I}$ with the maximum eigenvalue of $\mathbf{Z}_\mathbf{I}$. The gray-scale fused image is obtained by solving Poisson equation. Connah *et al.*(Connah, Drew, and Finlayson 2014) proposed more general fusion approach which generates a naturalistic color image from the multi-spectral images. They calculated multi-dimensional gradient field which not only preserves the structure tensor of inputs, but also keeps the naturalistic colors of a reference RGB image. The reintegration step is performed by a look-up-table (LUT) mapping (Finlayson, G. D., and Drew 2011).

Motivated by structure tensor based image fusion, we apply the framework to single image contrast enhancement (Fig. 1(c)). Different from the sequential fusion methods that just employ the structure tensor as the fusion strategy for fused-gradient, the proposed framework uses the metric as the loss function to train the whole network parameters. The multi-exposure images are only required for network training, and our method directly reconstructs detail-enhanced image through a single forward network. Employing the multi-exposure sequence for single image contrast enhancement can be similar to the framework of (Ying, Li, and Gao 2017; Ren et al. 2018; Zhang, Nie, and Zheng 2019). They synthesize the multi-exposure images given low-contrast input, and then combine them to obtain enhanced result (Fig. 1(b)). Our framework can be regarded to combining multi-exposure simulation and fusion processes of (Ying, Li, and Gao 2017; Ren et al. 2018; Zhang, Nie, and Zheng 2019) through a single forward network.

Proposed Method

Given a low-contrast image $\mathbf{f}$, we design a single forward network $\mathcal{D}_{CNN}$ to obtain an enhanced image $\mathbf{u}$, i.e., $\mathbf{u} = \mathcal{D}_{CNN}(\mathbf{f})$. Different from general image enhancement approaches that train the network with single ground-truth images, we propose to train $\mathcal{D}_{CNN}$ using the multi-exposure image sequence $\mathbf{I} = \{\mathbf{I}^1, \cdots, \mathbf{I}^N\}$.

Multi-Scale Structure Tensor Loss Function

We propose a multi-scale structure tensor loss function for directly reconstructing multi-exposure contrast from a single low-contrast image as follows:

$$\mathcal{L}(\mathbf{u},\mathbf{I}^1,\cdots,\mathbf{I}^N) = \sum_{s=0}^{M-1} \sum_{c\in\{r,g,b\}} \sum_{(x,y)\in\Omega} \left\| \mathbf{Z}_{\mathcal{T}_s(\mathbf{u}_c)}(x,y) - \mathbf{Z}_{\mathcal{T}_s(\mathbf{I}_c)}(x,y) \right\|_F^2, \tag{4}$$

where $\left\|\mathbf{A}\right\|_F$ denotes the Frobenius norm of a matrix $\mathbf{A}$, and Ω is a domain for image coordinates. The subscripts c and s indicate RGB index and scale factor, respectively. $\mathbf{I}_c$ is an N-dimensional image concatenating c components $(c \in \{r,g,b\})$ from N multi-exposure image sequence. We minimize structure tensor difference between network output and the sequence at each RGB component, separately. Calculating structure tensors with all RGB components causes the color information missing problem. Furthermore, we compare the structure tensor at M scale image size to ensure wider structural information of the image. The $\mathcal{T}_s(\mathbf{u}_c)$ is a resized image from the original ones with $1/2^s$ scale. The resizing operator $\mathcal{T}_s$ is implemented by blurring the previous scale image $\mathcal{T}_{s-1}(\mathbf{u}_c)$ with a 5×5 Gaussian kernel with standard deviation 1.4, followed by downsampling the blurred one by a scaling factor 2. The averaging with Gaussian weights enables the network output to capture global luminance consistency of the multi-exposure images.

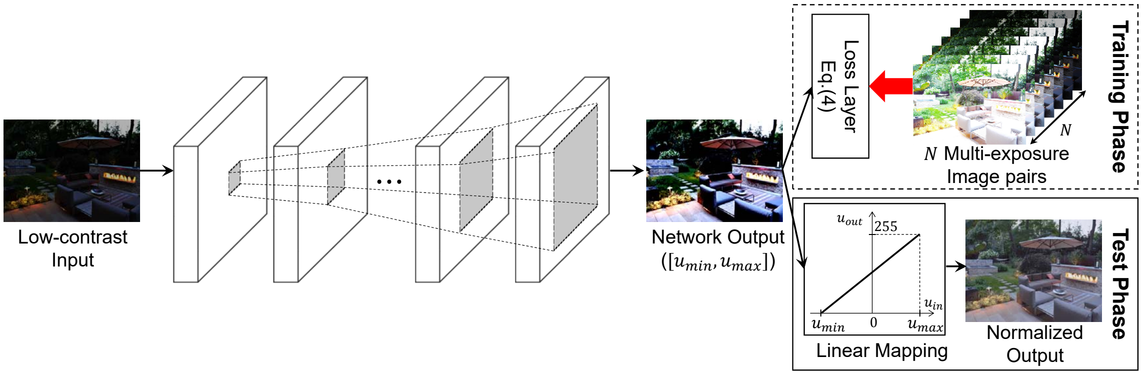


Figure 2: The overall architecture of our contrast enhancement. The network is trained by the proposed loss function of (4) with the multi-exposure sequences. At the test phase, the network output is linearly normalized to visualize within $[0, 255]$.

To apply the standard back-propagation algorithm (Mozer 1989) for network training, the proposed loss function of (4) should be fully differentiable. We provide the partial derivative of $\mathcal{L}$ with respect to $\mathbf{u}_c$. The resizing operator $\mathcal{T}_s$ is omitted to simplify the notions. Let us denote $d^{ij}_{\mathbf{u}_c}$ as the components of $\mathbf{Z}_{\mathbf{u}_c}(x, y)$ ($1 \leq i, j \leq 2$). From (4), $\frac{\partial \mathcal{L}}{\partial \mathbf{u}_c(x,y)}$ is calculated as follows:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{u}_c(x,y)} = 2\sum_{i=1}^{2}\sum_{j=1}^{2}\left(d^{ij}_{\mathbf{u}_c} - d^{ij}_{\mathbf{I}_c}\right)\frac{\partial d^{ij}_{\mathbf{u}_c}}{\partial \mathbf{u}_c(x,y)}. \tag{5}$$

Based on the structure tensor definition in (3), $\frac{\partial d^{ij}_{\mathbf{u}_c}}{\partial \mathbf{u}_c(x,y)}$ can be obtained by:

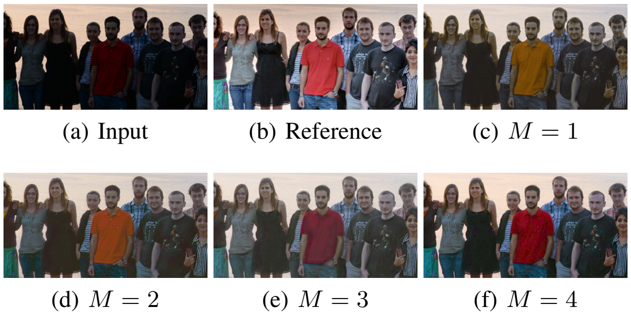
$$\frac{\partial d^{ij}_{\mathbf{u}_c}}{\partial \mathbf{u}_c(x,y)} = D_i^T D_j \mathbf{u}_c(x,y) + D_j^T D_i \mathbf{u}_c(x,y). \tag{6}$$

(6) can be simply implemented by convolutional layer with 1D Laplacian (for $i, j = 1$ or 2) and 2D Laplacian (for $i = 1$, $j = 2$ or $i = 2$, $j = 1$) operators.

Network Architecture

The network with large receptive fields can capture wide contextual information from the image. Generally, there are two methods to ensure large receptive fields of the network: deeper architecture (Zhang et al. 2017) and encoder-decoder (Noh, Hong, and Han 2015). However, the architectures require more parameters to be learned and increase the complexity. Furthermore, encoder-decoder architecture with pooling layers (or strided convolution) losses the subtle details during the resolution reduction at encoder part. In this paper, we instead design our network $\mathcal{D}_{CNN}$ using the multi-scale context aggregation network (CAN) (Yu and Koltun 2015) to make tradeoff between the size of receptive field and the number of parameters. The CAN uses dilated convolutions to aggregate wide contextual information.

We construct the CAN architecture for low-contrast image enhancement using 7 dilated convolution layers as shown in Fig. 2. All convolution layers have the size of 3×3 kernels with 64 intermediate feature maps except the last layer. The last convolution layer applies a 1×1 convolution that predicts the enhanced image $\mathbf{u}$. We set the dilation factor l_k of k^{th} convolution layer to 2^{k-2} for $2 \leq k \leq 6$. The factors of the first and the last layer set to 1 ($l_1 = l_7 = 1$). The receptive fields of $\mathcal{D}_{CNN}$ are 65×65. We use parametric rectified linear unit (PReLU) (He et al. 2015).



(a) Input (b) Reference (c) $M = 1$

(d) $M = 2$ (e) $M = 3$ (f) $M = 4$

Figure 3: The effect of the number of scales M in (4): (a) Low-light input, (b) reference, (c)-(f) results of our method trained with $M = 1$, 2, 3, and 4.

The network $\mathcal{D}_{CNN}$ is trained by the proposed loss function of (4) to approximate the structure tensor of N multi-exposure image pairs. However, an image containing the overall structural information of the multiple images is naturally represented beyond the conventional intensity range $[0, 255]$. As a result, we do not use any activation function at the last convolution layer in order not to limit the intensity range of the network output. This can be regarded our method as high dynamic range (HDR) image generation from low-contrast standard dynamic range (SDR) image. To visualize the network output $\mathbf{u}$ with $[u_{\min}, u_{\max}]$ ($u_{\min} < 0$ and $u_{\max} > 255$), we linearly normalized $\mathbf{u}$ at the test phase.

Experiments

Our enhancement method requires the multi-exposure image sequences to train the network $\mathcal{D}_{CNN}$. Recently, large-scale multi-exposure image dataset (Cai, Gu, and Zhang 2018) has

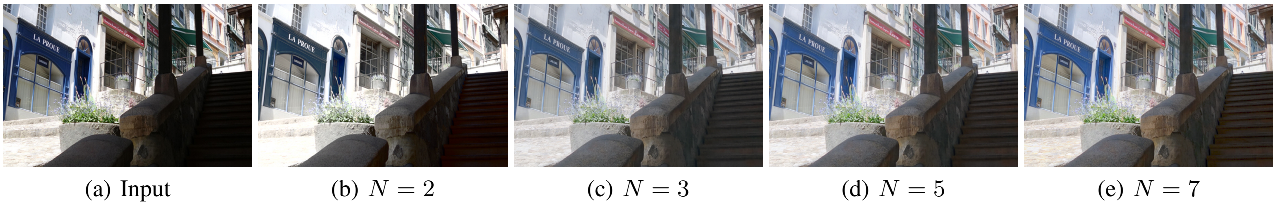
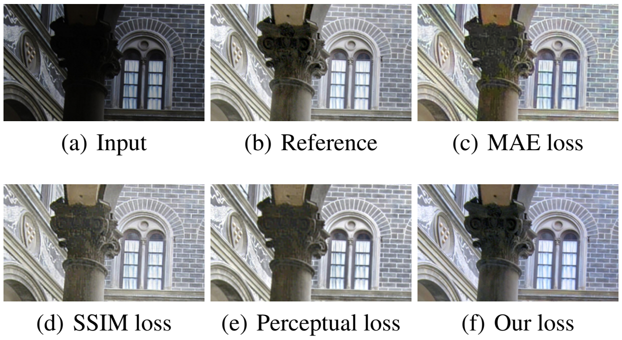

(a) Input (b) $N = 2$ (c) $N = 3$ (d) $N = 5$ (e) $N = 7$

Figure 4: The effect of the multi-exposure image sequence N in (4): (a) Input, (b)-(e) results of our method trained with $N = 2$, 3, 5, and 7. Training the network with larger number of sequences produces an enhanced image with abundant details.


(a) Input (b) Reference (c) MAE loss

(d) SSIM loss (e) Perceptual loss (f) Our loss

Figure 5: Results of $\mathcal{D}_{CNN}$ according to training different loss functions: (a) Low-light image, (b) reference image, (c-f) results trained by MAE, MS-SSIM (Zhao et al. 2016), perceptual (Dosovitskiy and Brox 2016), and our losses.

been constructed including both indoor and outdoor scenes. We trained our network using 7 multi-exposure sequences for each image from (Cai, Gu, and Zhang 2018), which covers most of the exposure levels. The TensorFlow library with 12GB NIVIDIA Titan GPU is used for network construction and training (our code will be made publicly available). The loss function of (4) is minimized with the Adam solver (Kingma and Ba 2014) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ setting). We randomly cropped 5×10^4 patches with 128×128 size from our training dataset, and trained our network using the patches. The learning rate was initialized as 10^{-3} and halved every 10 epoches until 100 epoches.

Analysis

Effect of scale number M The proposed method could be vulnerable to capture natural color information when training the network with a single scale manner ($M = 1$), since the loss function of (4) only compares the structural information between the network output and the multi-exposure images through the structure tensor. The multi-scale design of our loss ensures that the network reconstructs the global color consistency of the multi-exposure images due to the weighted averaging process. Fig. 3 shows the results of our method according to the number of scales M. We used 7 multi-exposure sequences to train the network. When we set $M = 1$, the enhanced result does not contain enough color information (Fig.3(c)). Our method reconstructs color information well as M increases. We empirically confirmed that the network trained by $M = 4$ is enough to produce high-quality results with abundant colors (Fig. 3(f)).

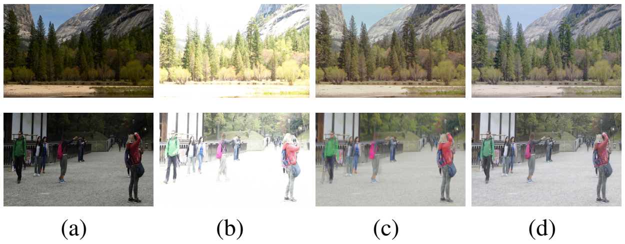

(a) (b) (c) (d)

Figure 6: Comparison of multi-exposure fusion and single enhancement of our loss function (4): (a) Under-exposure, (b) over-exposure, (c) $\mathcal{D}_{CNN}^{fusion}(\mathbf{I})$, and (d) $\mathcal{D}_{CNN}(\mathbf{f})$.

Effect of image sequence number N We trained our network varying the number of image sequence N as 2, 3, 5, and 7 to analyze the effect of N. The number of scale M in (4) is fixed to 4. Fig. 4 shows the visual results of our method according to different N. For $N = 2$ (Fig. 4(b)), the results do not match the correct color and contrast information, since the normal exposure images are not included when training the network. The results trained with large number of sequences ($N = 7$, Fig. 4(e)) show visually best results since wide range of exposure levels can provide abundant details for the contrast enhancement.

Comparison with Different Loss Functions Fig. 5 shows the results of $\mathcal{D}_{CNN}$ which are trained with various loss functions, including MAE, MS-SSIM (Zhao et al. 2016), and perceptual (Dosovitskiy and Brox 2016) losses. We used the high-quality images (Fig. 5(b)) from the training dataset (Cai, Gu, and Zhang 2018) as the ground-truth for the losses. The result with MAE loss suffers from color distortions (Fig. 5(c)) due to the independent property on local characteristics of the loss. Even though the SSIM and perceptual losses consider the neighboring pixels, they also produce the enhanced image with amplified noises (Fig. 5(d)-(e)). Different from other losses, our method ensures the high-quality enhanced results without such distortions (Fig. 5(f)). The visual results demonstrate that our loss function with multi-exposure image sequence outperforms the conventional losses with explicit ground-truths.

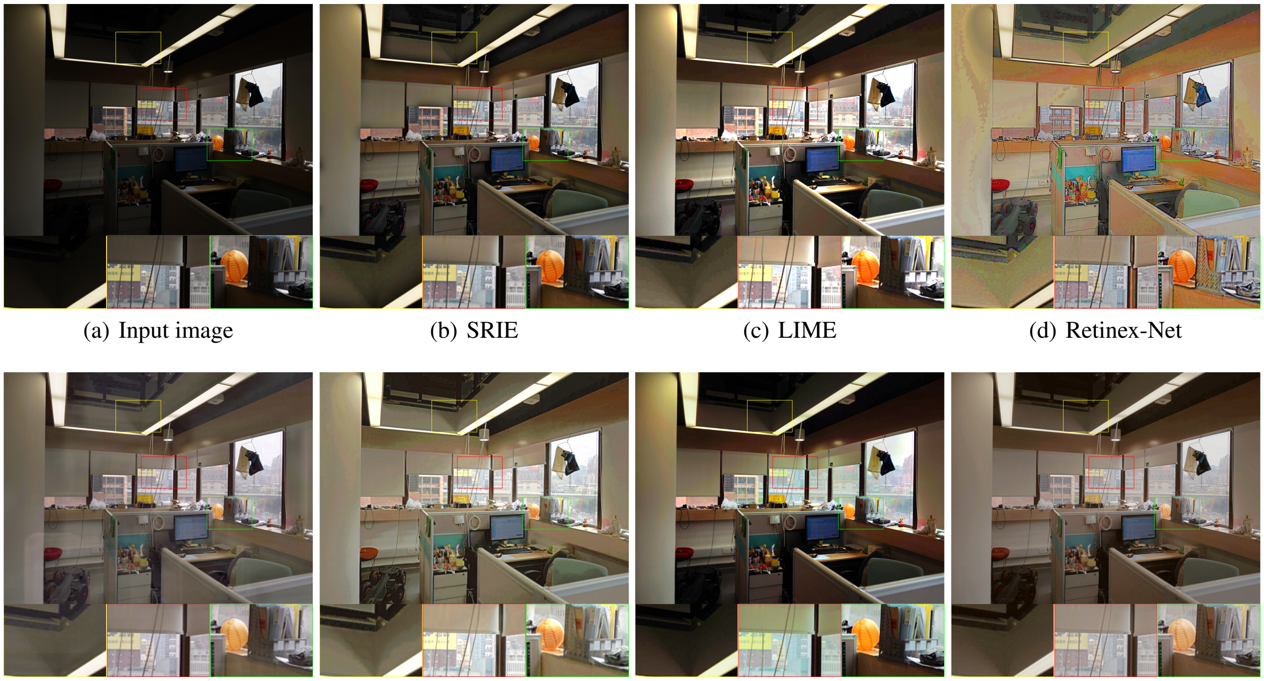

(a) Input image
(b) SRIE
(c) LIME
(d) Retinex-Net
(e) DHN
(f) SICE-Net
(g) UPE
(h) Ours

Figure 7: Visual comparison with state-of-the-art methods on under-exposure image from (Cai, Gu, and Zhang 2018). (a) Under-exposure image, (b) SRIE (Fu et al. 2016), (c) LIME (Guo, Li, and Ling 2017), (d) RetiNex-Net (Wei et al. 2018), (e) DHN (Ren et al. 2019), (f) SICE-Net (Cai, Gu, and Zhang 2018), (g) UPE (Wang et al. 2019), and (h) ours.

Comparison with Multi-Exposure Fusion

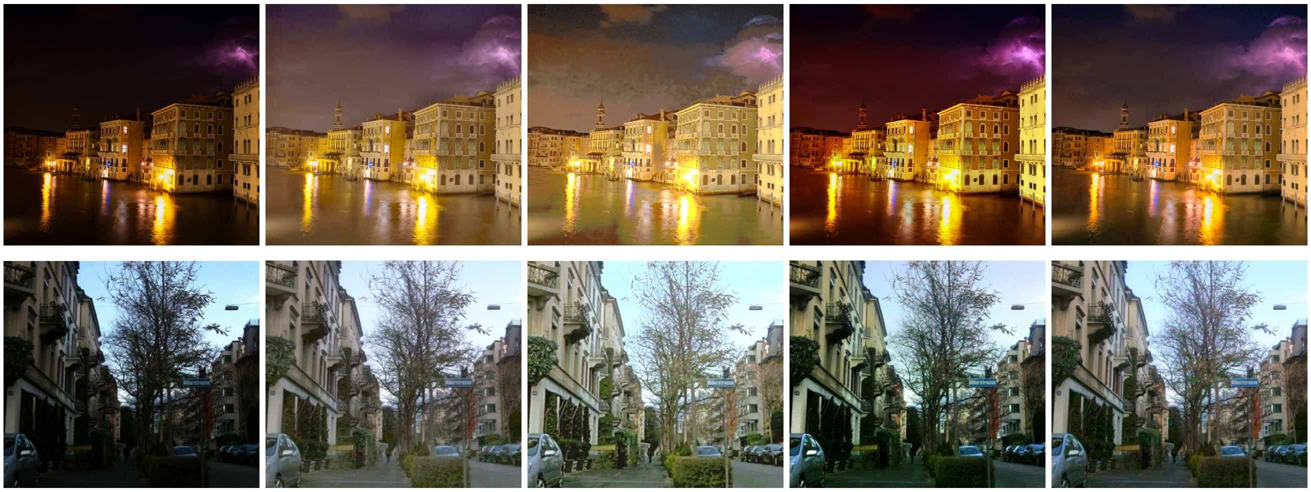
The structure tensor has been used to train the multi-exposure image fusion in (Jung et al. 2020), where the network generates a fused image which contains the overall contrast from multiple inputs. Given a stacked multi-exposure sequence $\mathbf{I}$, we trained the networks $\mathcal{D}_{CNN}^{fusion}(\mathbf{I})$ using the structure tensor loss with $M = 4$ and $N = 7$. Fig. 6 shows the results of $\mathcal{D}_{CNN}^{fusion}(\mathbf{I})$ and $\mathcal{D}_{CNN}(\mathbf{f})$. Fig. 6(a-b) are two examples of seven multi-exposure images. For the static scene with no moving objects (the first row of Fig. 6), $\mathcal{D}_{CNN}^{fusion}(\mathbf{I})$ outputs a fused image with abundant color and contrast. However, when the sequence is not perfectly aligned due to the movement of camera and objects, the fused image suffers from ghost and blurry artifacts as shown in the second row of Fig. 6(c). Different from multi-exposure fusion, our single image contrast enhancement is free from the misalignment problem (Fig. 6(d)). Even though several multi-exposure sequence in our training dataset is not perfectly aligned, there is no significant effect on the performance of contrast enhancement due to large portion of correct alignment region. This coincides with the observation in the literature of weakly- and self-supervised semantic segmentation (Lin et al. 2016; Pathak et al. 2017).

Qualitative Evaluation

Results on under-exposure benchmarks We compare our method with the following six state-of-the-art methods: SRIE (Fu et al. 2016), LIME (Guo, Li, and Ling 2017), Retinex-Net (Wei et al. 2018), DHN (Wang et al. 2019), SICE-Net (Cai, Gu, and Zhang 2018), and UPE (Wang et al. 2019). The first two methods are Retinex-based algorithm, and the others are the recent CNN-based methods. The results for the comparison methods are obtained from author's source codes. For fair comparison, we further re-trained the CNN-based methods on our training dataset.

Fig. 7 shows the visual result of our method with the comparison methods on under-exposure image from (Cai, Gu, and Zhang 2018). SRIE (Fu et al. 2016) tends to generate darker results compared to other methods. The results of LIME (Guo, Li, and Ling 2017) suffer from contrast distortion (ceiling in Fig. 7) and saturation (curtain in Fig. 7). Retinex-Net (Wei et al. 2018) severely amplifies the noise, and contain unnatural color information. DHN (Wang et al. 2019) and SICE-Net (Cai, Gu, and Zhang 2018) suffer from wrong details. UPE (Wang et al. 2019) brings visually better results, but still contains weak details in dark region and color distortion (near the window in Fig. 7). Compared to the comparison methods, the proposed method results with abundant details, better contrast, and natural colors, while avoiding saturation artifacts.

Results on various low-contrast benchmarks We apply the proposed method, which is trained on under-exposed inputs, on different low-contrast benchmarks to show the robustness. Fig. 8 shows the results of recent CNN-based methods on low-light (Guo, Li, and Ling 2017) (first row)



(a) Input (b) DHN (c) SICE-Net (d) UPE (e) Ours

Figure 8: Visual comparison on various low-contrast benchmarks. From first to second rows, the inputs lost details by low-light (Guo, Li, and Ling 2017) and low-end device (Ignatov et al. 2017).

Method	Under-exposure	Under-exposure	Low-light	DPED
	(Cai, Gu, and Zhang 2018)	(Ma et al. 2017)	(Guo, Li, and Ling 2017)	(Ignatov et al. 2017)
SRIE	4.147 / 2.220	4.766 / 2.896	**3.447** / 2.713	5.312 / 2.939
LIME	4.477 / 2.426	4.978 / 2.544	4.424 / 3.290	5.692 / 3.107
Retinex-Net	5.513 / 1.895	6.006 / 2.259	7.385 / 1.978	6.370 / 2.107
DHN	3.345 / 3.138	2.880 / 3.262	3.855 / 2.832	2.819 / 3.190
SICE-Net	2.528 / 3.291	3.133 / 3.349	4.008 / 3.153	2.585 / 3.118
UPE	2.164 / 3.280	2.689 / 3.408	3.833 / 2.898	2.505 / 3.350
Ours	**1.987 / 3.612**	**2.482 / 3.516**	3.613 / **3.642**	**2.246 / 3.420**

Table 1: Average NIQE (Mittal, Soundararajan, and Bovik 2013) and CDQE (Fang et al. 2015) metrics on four benchmarks (Cai, Gu, and Zhang 2018; Ma et al. 2017; Guo, Li, and Ling 2017; Ignatov et al. 2017) for contrast enhancement. The lower/higher values of NIQE/CDQE, the better image quality. The best results for each benchmark are highlighted in bold.

and smartphone (Ignatov et al. 2017) (second row) images. DHN (Wang et al. 2019) does not contain vivid and natural color. The results of SICE-Net (Cai, Gu, and Zhang 2018) contains distorted contrast (for low-light image) and over-smoothing (for smartphone image). UPE (Wang et al. 2019) generates naturalistic enhanced images, but tend to be low-contrast and dark compared to other methods. We observed that our method works well to enhance the low-contrast images from various benchmarks regardless of the difference between test and training examples.

Quantitative Evaluation

Since the ground-truths for real scenarios are not usually available, we use two no-reference quality metrics, i.e., NIQE (Mittal, Soundararajan, and Bovik 2013) and CDQE (Fang et al. 2015), for objective evaluation. NIQE (Mittal, Soundararajan, and Bovik 2013) measures the naturalness of an image by comparing with a corpus of pristine naturalistic images, and CDQE (Fang et al. 2015) measures the contrast distortion of enhanced results based on natural scene statistics (NSS). We constructed four different testsets from two under-exposure (Cai, Gu, and Zhang 2018; Ma et al. 2017), low-light (Guo, Li, and Ling 2017), and low-end device (Ignatov et al. 2017) benchmarks. Each dataset consists of 118, 24, 10, and 30 examples, respectively. Table 1 summarizes the quantitative evaluations of the comparison methods on the benchmarks. We observed that the proposed method outperforms other methods in terms of the image naturalness and contrast distortion.

Conclusion

We introduced the new deep learning approach for single image contrast enhancement motivated by existing variational fusion methods. The key idea of our approach is to utilize the structure tensor representations as the loss function with various multi-exposure image sequences. The proposed loss function enables the network to generate an enhanced image approximating the abundant contrast from the sequences. Rather than penalizing per-pixel intensity (or high-level feature) difference, our direct contrast mapping with various exposure levels is appropriate for enhancing low-contrast image. The experimental results demonstrate the superiority of our method in various low-contrast benchmarks.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00056, To create AI systems that act appropriately and effectively in novel situations that occur in open worlds).

References

Arici, T.; Dikbas, S.; and Altunbasak, Y. 2009. A histogram modification framework and its application for image contrast enhancement. *IEEE Trans. Image Process.* 18(9): 1921–1935.

Bychkovsky, V.; Paris, S.; Chan, E.; and Durand, F. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. *In: CVPR* .

Cai, J.; Gu, S.; and Zhang, L. 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Trans. Image Process.* 27(4): 2049–2062.

Celik, T.; and Tjahjadi, T. 2011. Contextual and variational contrast enhancement. *IEEE Trans. Image Process.* 20(12): 3431–3441.

Connah, D.; Drew, M. S.; and Finlayson, G. 2014. Spectral edge image fusion: Theory and applications. *In: ECCV* .

Di Zenzo, S. 1986. A note on the gradient of a multi-image. *CVGIP* 33(1): 116–125.

Dong, C.; Loy, C.-C.; He, K.; and Tang, X. 2015. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(2): 295–307.

Dosovitskiy, A.; and Brox, T. 2016. Generating images with perceptual similarity metrics based on deep networks. *In: NIPS* .

Fang, Y.; Ma, K.; Wang, Z.; Lin, W.; Fang, Z.; and Zhai, G. 2015. No-reference quality assessment of contrast-distortion images based on natural scene statistics. *IEEE Trans. Signal Process. Lett.* 22(7): 838–842.

Finlayson; G. D., Connah, D.; and Drew, M. S. 2011. Lookup-table-based gradient field reconstruction. *IEEE Trans. Image Process.* 20(10): 2827–2836.

Fu, X.; Zeng, D.; Huang, Y.; Zhang, X.-P.; and Ding, X. 2016. A weighted variational model for simultaneous reflectance and illumination estimation. *In: CVPR* .

Gu, B.; Li, W.; Wong, J.; Zhu, M.; and Wang, M. 2012. Gradient field multi-exposure images fusion for high dynamic range image visualization. *VCIR* 23(4): 604–610.

Guo, X.; Li, Y.; and Ling, H. 2017. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Trans. Image Process.* 26(2): 982–993.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *In: ICCV* .

Ignatov, A.; Kobyshev; N., Timofte, R.; Vanhoey, K.; and Van Gool, L. 2017. DSLR-quality photos on mobile devices with deep convolutional networks. *In: ICCV* .

Jung, H.; Kim, Y.; Jang, H.; Ha, N.; and Sohn, K. 2020. Unsupervised Deep Image Fusion With Structure Tensor Representations. *IEEE Transactions on Image Processing* 29: 3845–3858.

Kim, Y.; Jung, H.; Min, D.; and Sohn, K. 2017. Deeply aggregated alternating minimization for image restoration. *In: CVPR* .

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980* .

Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; and Matas, J. 2018. Deblurgan: Blind motion deblurring using conditional adversarial networks. *In: CVPR* .

Land, E. H. 1977. The retinex theory of color vision. *Scientific American* 237(6): 108–129.

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. *In: CVPR* .

Ma, K.; Li, H.; Yong, H.; Wang, Z.; Meng, D.; and Zhang, L. 2017. Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Trans. Image Process.* 26(5): 2519–2532.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2013. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* 20(3): 209–212.

Mozer, M. C. 1989. A focused back-propagation algorithm for temporal pattern recognition. *Complex Systems* 3(4).

Noh, H.; Hong, S.; and Han, B. 2015. Learning deconvolution network for semantic segmentation. *In: ICCV* .

Pathak, D.; Girshick, R.; Dollár, P.; Darrell, T.; and Hariharan, B. 2017. Learning features by watching objects moving. *In: CVPR* .

Ren, W.; Liu, S.; Ma, L.; Xu, Q.; Xu, X.; Cao, X.; Du, J.; and Yang, M. H. 2019. Low-Light Image Enhancement via a Deep Hybrid Network. *IEEE Trans. Image Process.* .

Ren, Y.; Ying, Z.; Li, T.; and Li, G. 2018. LECARM: low-light image enhancement using the camera response model. *CSVT : IEEE Trans. Circuits Syst. Video Technol.* 29(4): 968–981.

Socolinsky, D. A.; and Wolff, L. B. 2002. Multispectral image visualization through first-order fusion. *IEEE Trans. Image Process.* 11(8): 923–931.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019. Underexposed Photo Enhancement Using Deep Illumination Estimation. *In: CVPR* .

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. *In: BMVC* .

Ying, Z.; Li, G.; and Gao, W. 2017. A bio-inspired multi-exposure fusion framework for low-light image enhancement. *arXiv:1711.00591* .

Yu, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122* .

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.* 26(7): 3142–3155.

Zhang, Q.; Nie, Y.; and Zheng, W. 2019. Illumination estimation for robust exposure correction. *Comput. Graph. Forum* 39(7): 243–252.

Zhao, H.; Gallo, O.; Frosio, I.; and Kautz, J. 2016. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imag.* 3(1): 47–57.