Deep Perceptual Image Enhancement Network for Exposure Restoration

Karen Panetta, *Fellow, IEEE*, Shreyas Kamath K. M., *Student Member, IEEE*, Shishir Paramathma Rao, *Student Member, IEEE*, and Sos S. Agaian, *Fellow, IEEE*

***Abstract*—Image restoration techniques process degraded images to highlight obscure details or enhance the scene with good contrast and vivid color for the best possible visibility. Poor illumination condition causes issues, such as high-level noise, unlikely color or texture distortions, nonuniform exposure, halo artifacts, and lack of sharpness in the images. This article presents a novel end-to-end trainable deep convolutional neural network called the deep perceptual image enhancement network (DPIENet) to address these challenges. The novel contributions of the proposed work are: 1) a framework to synthesize multiple exposures from a single image and utilizing the exposure variation to restore the image and 2) a loss function based on the approximation of the logarithmic response of the human eye. Extensive computer simulations on the benchmark MIT-Adobe FiveK and user studies performed using Google high dynamic range, DIV2K, and low light image datasets show that DPIENet has clear advantages over state-of-the-art techniques. It has the potential to be useful for many everyday applications such as modernizing traditional camera technologies that currently capture images/videos with under/overexposed regions due to their sensors limitations, to be used in consumer photography to help the users capture appealing images, or for a variety of intelligent systems, including automated driving and video surveillance applications.**

***Index Terms*—Channel attention network, deep convolutional neural networks, dilated residual network, human vision system, image enhancement, logarithmic exposure transformation (LXT), multiscale human color vision (MHCV) loss.**

I. Introduction

IMAGES and videos capture a vast amount of rich and detailed information about the scene. Intelligent systems use these captured images for various computer vision tasks, such as image enhancement, object detection, classification and recognition, segmentation, 3-D scene understanding, and modeling [1]. These tasks form the building block for real-world applications, such as autonomous driving, security surveillance systems, search and rescue operations, and virtual and augmented reality environments. The quality of images becomes extremely important for these applications, and the systems' performance might be affected negatively by low-quality inputs.

Acquiring a high or optimum quality image is ideal but sometimes impractical. Specifically, smartphone cameras have considerably small apertures, limiting the amount of light captured, leading to noisy images in a low-lit environment [5]. The imaging sensor's linear characteristic fails to replicate the complex and nonlinear mapping achieved by human vision. Another issue that commonly restricts the performance of computer vision algorithms is nonuniform illumination. When the lighting source is not perfectly aligned and normal to the viewing surface, or if the surface is not planar, then the resulting image may have nonuniform illumination artifacts [6]. Another critical requirement for efficient image processing is global uniformity [6]. Similar objects or structures should appear the same within an image or in a series of images. This implies that the color content and the illumination must be stable for images acquired under varying conditions. Illuminations that cast strong shadows also cause problems. The edges and boundaries in an image need to be well defined and accurately located, implying that the image's high-frequency content needs to be preserved to have high local sensitivity. Vignetting is another common pitfall in many photos [7]. While it might be a desirable effect in some cases such as portrait mode photography, it is not ideal for various other use cases that require high accuracy and details. Furthermore, the compression algorithms used to store the images may cause some artifacts [8]. These factors affect the pleasantness of viewing the image and affect the usability of the images for computer vision algorithms and their ability to analyze them.

Traditionally, automatic image quality enhancement methods can be broadly classified into global enhancements and local enhancements. Global enhancement algorithms perform the same operation on every single image pixel, such as linear contrast amplification. Such a simple technique will lead to saturated pixels in high exposure regions. To avoid this effect, nonlinear monotonic functions, such as mu-law, power-law, logarithmic processing [9], [10], gamma functions, and piecewise-linear transformation functions, are used to perform enhancements [11].

One extensively used method to avoid saturation while improving the contrast is histogram equalization (HE) [12].

Manuscript received July 6, 2021; revised December 15, 2021; accepted December 31, 2021. This work was supported by the U.S. Army Combat Capabilities Development Command under Grant W911QY-15-2-0001. This article was recommended by Associate Editor P. Shi. *(Corresponding author: Shreyas Kamath K. M.)*

Karen Panetta, Shreyas Kamath K. M., and Shishir Paramathma Rao are with the Department of Electrical and Computer Engineering, Tufts University, Medford, MA 02155 USA (e-mail: karen@ece.tufts.edu; skamat04@tufts.edu; shishir.rao@tufts.edu).

Sos S. Agaian is with the Department of Computer Science, College of Staten Island and the Graduate Center, City University of New York, New York, NY 10017 USA (e-mail: sos.agaian@csi.cuny.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3140202.

Digital Object Identifier 10.1109/TCYB.2021.3140202

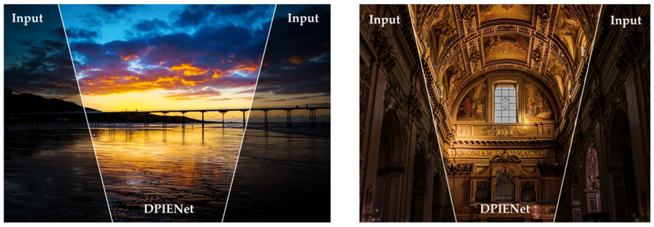


Fig. 1. Demonstration of the proposed DPIENet for a given ill exposed input. This system sets a new SOTA benchmark in terms of measures, such as PSNR, SSIM [2], GSSIM [3], and UQI [4].

Another local image enhancement technique is based on the Retinex theory [13], which assumes that the amount of light reaching the observer can be decomposed into two parts: 1) scene reflectance and 2) illumination components. These algorithms achieve better results than global methods by making use of the local spatial information directly and have become the forerunners for image enhancement. While methods based on Retinex such as MSR-CR [14] can effectively improve the sharpness of the image and increase the local contrast, they introduce the halation phenomenon at high contrast and amplified noise regions [15].

More recently, deep learning-based image enhancement methods have been used to mitigate these problems [16], [17]. These techniques allow for automatic parameter selection and training and have highly scalable architectures. They have been shown to outperform state-of-the-art (SOTA) methods in computer vision tasks, such as object detection, object recognition, segmentation, super-resolution, and enhancement. However, most of the deep learning networks are trained explicitly for either standard exposure images or low exposure images. Thus, they fail to achieve global uniformity for varying exposure inputs of the same scene.

This article proposes a deep learning-based perceptual image enhancement network (DPIENet) to address these issues. This network has a U-shaped structure similar to the U-Net architecture [18]. It consists of two stages: 1) a feature condense network (FeCN) that aims to acquire compact feature representation of the spatial context of the image and 2) a feature enhance network (FeEN) that performs nonlinear upsampling of the input feature maps to reconstruct an enhanced image. The architecture is equipped with skip connections between these two networks to use high-resolution image details during the reconstruction. An example of the result obtained using the network is illustrated in Fig. 1.

Some of the notable contributions of DPIENet include the following.

1) A unified network that can ensure global uniformity by generating perceptually similar enhanced images for input images of both standard and low exposure setting by utilizing dilated convolutions to preserve spatial resolution in convolutional networks and improve spatial image understanding. Furthermore, it incorporates a channel attention mechanism that aims to adaptively rescaling channelwise features by extracting the channel statistics to enhance the network's discriminative ability.
2) A combination of a classical log-based synthetic multiexposure image generation technique—logarithmic exposure transformation (LXT) that employs trainable parameters to improve the performance of the network.
3) A novel loss function—"multiscale human color vision (MHCV) loss." This loss aims at improving the quality of the reconstruction by considering human perception. This loss function promotes the model to learn complicated mappings and effectively reduces the undesired artifacts, such as noise, unrealistic color or texture distortions, and halo effects.

The remainder of this article is organized as follows. In Section II, related recent literature is reviewed. A detailed description of the DPIENet architecture and its analysis is provided in Section III. In Section IV, a brief description of the proposed MHCV loss is provided. Section V presents the training details and an ablation study with quantitative and visual experimental results. Section VI discusses the user study performed to measure human perceptual preferences. This section is followed by the computation complexity, application, and conclusion in Sections VII–IX, respectively.

II. Related Work

Various methods have been adopted in the literature for enhancing the quality of the images. Some of the early techniques include gray level slicing, contrast expansion, linear and nonlinear contrast stretching, and various histogram processing [19]. Many extensions to HE-based methods, such as adaptive HE [20], contrast-limited AHE [21], and dynamic HE [22], impose additional constraints while redistributing the luminous intensity of histogram. However, such global enhancement methods may suffer from loss of details in some local areas because of the inherently nonuniformity present in the image.

Most Retinex-based methods, including MSR-CR [14], SSR [23], and HECUP [24], recover the reflectance and illumination component and typically employ varying amounts of the illumination component for enhancing images while preserving naturalness. There exists multiple variations and extensions of the Retinex-based approach, such as AMSR [25], which uses an adaptive weighting strategy, LIME [26], which only estimates the illumination component for low light image enhancement, and NPE [27], which balances the enhancement by utilizing the bio-inspired multi-image fusion framework for image enhancement. Other fusion-based frameworks [28], [29] have also been proposed.

Recently, deep learning-based methods have introduced powerful tools, such as end-to-end trainable networks, generative adversarial networks (GANs) [30], and deep autoencoders [31], to perform image enhancement tasks. In [32], an end-to-end deep learning-based method for photo adjustment was proposed. Ignatov *et al.* [33] created a dataset of images captured by smartphone cameras and a DSLR camera and used the GAN model to learn the mapping between the two images. In [34] and [35], deep learning was used to approximate existing filters using a fully convolutional network (FCNs). While the methods mentioned above are all supervised learning, meaning they need paired images to learn the mapping, in [36], an unpaired deep learning model for image enhancement was proposed. This model uses an

TABLE I
LITERATURE REVIEW OF THE STATE-OF-THE-ART TECHNIQUES FOR IMAGE ENHANCEMENT

Author	Method	Explanation
Aubry, Mathieu, et al. (2014) [37]	Fast Local Laplacian Filters	This method utilizes the Laplacian filters for halo and other artifacts-free multi-scale manipulations for image enhancements. The execution time is high for these filters.
Fu, Xueyang, et al. (2016) [38]	Weighted variational model	This method utilizes Retinex theory [13] and simultaneously estimates highly detailed reflectance and illumination components using a weighted variational model while suppressing noise content.
Ying, Zhenqiang, Ge Li, and Wen Gao (2017) [39]	Multi-exposure fusion	This method is specially designed for low light images and utilizes the human visual system-inspired multi-exposure fusion framework to enhance the under-exposed regions of the image.
Wang, Shuhang, et al. (2017) [40]	Content adaptive histogram equalization	This method enhances the contrast of outdoor images by using content-aware histogram equalization and contrast-dependent color saturation adjustments while maintaining the naturalness of the images.
Chen, Qifeng, Jia Xu, and Vladlen Koltun (2017) [34]	Deep learning	This method utilizes fully convolutional neural networks to learn classical handcrafted image processing operations such as multiple variational models, tone and detail manipulation, photorealistic stylization in order to be able to perform these operations in constant time.
Ignatov, Andrey, et al. (2017) [33]	Deep learning	This method employs an end-to-end trainable network to translate a low-quality smartphone image into a DSLR-quality image. It utilizes residual CNNs to obtain high-quality color rendition and improve image sharpness.
Chen, Yu-Sheng, et al. (2018) [36]	Deep learning	This method uses GAN and Wasserstein GAN, which have an adaptive weighting scheme to enhance images.
Wei, Chen, et al. (2018) [41]	Deep learning	This method utilizes a deep Retinex model for low light image enhancement and combines two networks, Decom-Net and Enhance-Net, for decomposition and enhancement, respectively.
Wang, Wenjing, et al. (2018) [42]	Deep learning	This method proposes to use global illumination prior and detail-preserving reconstruction to enhance low light images.
Cai, J., Gu, S., & Zhang, L. (2018) [43]	Deep learning	This single image contrast enhancement method utilizes CNNs to reveal details in the low-light regions of the image. It uses a multi-exposure image dataset to train the model.
Lv, Feifan, et al. (2018) [44]	Deep learning	This method uses a multi-branch approach, including feature extraction, enhancement, and fusion branches for low light image enhancement while suppressing noise and other artifacts in low light regions.
Jiang, Yifan, et al. (2019) [45]	Deep learning	This method uses a GAN network for unsupervised or unpaired training for low light image enhancement and has a good real-world generalization. It also incorporates a global-local discriminator structure that can handle complex spatially varying lighting conditions.
Wang, Ruixing, et al. (2019) [46]	Deep learning	This method uses an intermediate illumination learning component that differs from the conventional image-to-image translation approaches to learning complex human-like adjustments to the photographs.

adaptive weighting scheme extension of Wasserstein GAN for faster convergence, a global U-net model for the generator, and individual batch normalization (BN) for high-quality sharpened image enhancements. Other CNN-based methods, such as LLNet [31], utilize autoencoders, to extract features from low-light images. They adaptively adjust the image brightness without overamplification or saturation artifacts, thus achieving both image enhancement and denoising.

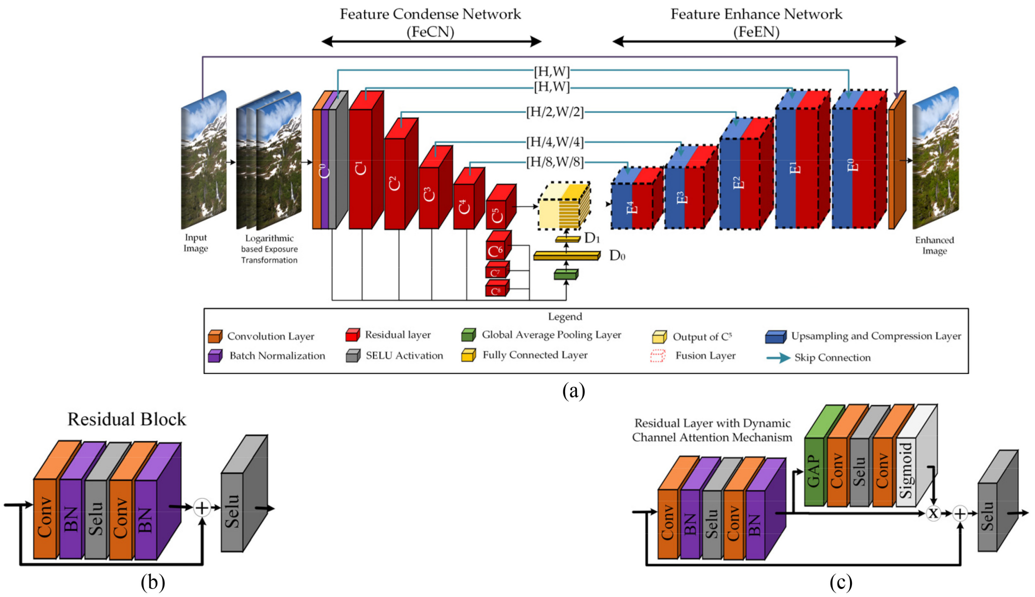
Furthermore, a few inverse tone mapping techniques utilize deep learning to improve the image's perceptual quality. Eilertsen *et al.* [47] used the U-Net structure operating in the logarithmic domain to generate a high dynamic range (HDR) output. Endo *et al.* [48] utilized UNet-based autoencoders to synthesize a set of LDR images with varying exposures to mimic exposure bracketing. These LDR images are then fused using a classical method to generate the HDR output. Table I provides a chronological list of various other image enhancement methods, along with a brief explanation for each method.

III. PROPOSED METHOD

A brief description of the proposed deep perceptual image enhancement network (DPIENet) is provided in this section. A basic flow diagram of the proposed system is outlined in Fig. 2. The goal of this article is to construct a function f developed specifically to obtain an enhanced image $f(I)$, where I is an input image of any arbitrary size (m, n). This network addresses the image-to-image translation problem, which transforms an input image with color rendition, ill exposure, and unrealistic color issues to an enhanced output image with desired characteristics. In accordance with this, DPIENet comprises of three main components: 1) logarithmic-based exposure transformation; 2) joint local and multiblock global feature extraction; and 3) dynamic channel attention (DCA) blocks. These components are tightly coupled and trained in an end-to-end fashion. For training, a novel loss is designed to obtain $f(I)$. This loss aims at enhancing the desired characteristics by using reflectance and illumination components. Additional details of these components are provided in further sections.

A. Logarithmic Exposure Transformation

To represent the wide range of luminance present in a natural scene, such as bright and direct sunlight to dark and faint shadows, the exposure range of the image needs to be adjusted. An ideal enhanced image would preserve high-quality details in the shadows while retaining a good



(a)

(b)

(c)

Fig. 2. Network architecture of the proposed deep perceptual image enhancement (DPIENet): (a) provides an overall structure of DPIENet with the FeCN that aims at acquiring information of the spatial context of the image and FeEN that focuses on reconstructing a perceptually enhanced image; (b) visualizes the standard residual network proposed by He *et al.* [49]; and (c) visualizes the residual network with a DCA mechanism to emphasize more on significant features.

contrast in the bright regions. On the contrary, an image with nonuniform scene luminance will have a tradeoff between the bright and dark regions due to the limited exposure and results in the loss of data in those regions. Various SOTA systems have been developed such as HDR imaging, which aim at combining multiple exposures to create an image with a greater dynamic range of light. The main constraint with such system is the requirement of multiple images across time with varying exposures. Inspired by the multiexposure mechanism from the HDR imaging systems, a synthetic simulation of changes in exposures to generate a perceptually enhanced image from a single image is explored. Specifically, the synthetic images need to have under and overexposed images. The underexposed images have bright regions, which are well defined with proper contrast and overexposed images where the finer details in the dark and shadow areas are highlighted

$$I' = \left.\frac{\log\left\{1+\alpha_x * \hat{I}_{x_{\max}} * \left(\hat{I}_x/\hat{I}_{x_{\max}}\right)^{\gamma_x}\right\}}{\log\{1+\alpha\}}\right| x=\{O,U\}$$
$$I_x = \begin{cases} I', & x = O \\ 1-I', & x = U \end{cases} \tag{1}$$

where $\hat{I}_{x_{\max}}$ is the maximum intensity; α is initialized to 2; $\gamma_U = 1.75$; $\gamma_O = 0.75$; $\hat{I}_U = \hat{I}_{U_{\max}} - \hat{I}$; $\hat{I}_O = \hat{I}$.

Consider an input image $\hat{I}$ of any arbitrary size (m, n), then the LXT of that image is generated by employing (1). This transform is derived using companding functions, such as μ-law and the power law, and it produces underexposed (U) and overexposed images (O). In (1), α is a learnable parameter and γ_x value is empirically set to 1.75 and 0.75 for underexposed and overexposed, respectively, based on the tradeoff between the expansion of underexposed regions and the amount of details in the overexposed areas. To simulate the overexposed image I'_O, LXT maps the low-intensity values to a broader range of values while compressing the range of higher intensity values

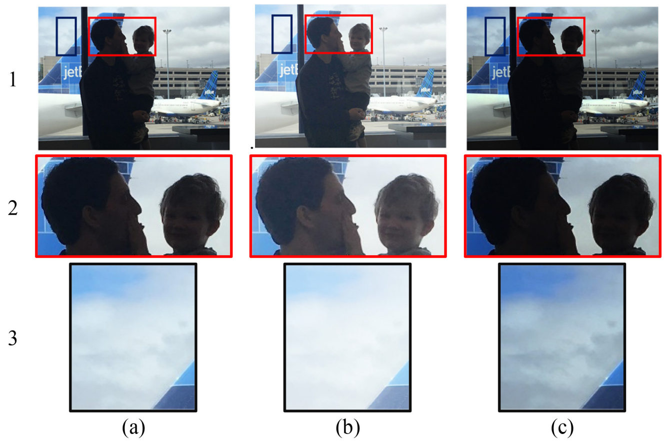
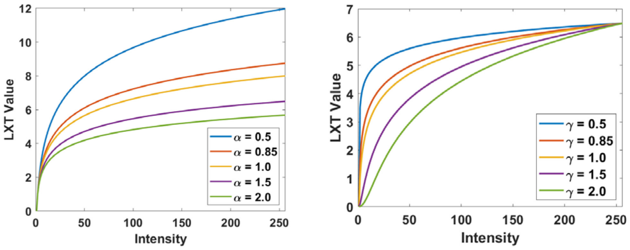


Fig. 3. Example of the LXT operation applied on an image: Row 1 is a visualization of the complete image, and rows 2 and 3 are the zoomed section of the image. Column (a) is the original image; (b) is the simulation of an over-exposed image wherein the darker regions are enhanced appropriately; and (c) is a simulation of an under-exposed image wherein the brighter regions are well defined.

Conversely, to obtain the underexposed images I'_U, the LXT function expands the higher intensity regions and compresses the range of lower intensities. Fig. 3 shows the result of the operation for various values of α. Fig. 3(b) visualizes an overexposed image with $\alpha_O = 2$ and $\gamma_O = 0.75$, and Fig. 3(c) demonstrates an underexposed image with $\alpha_U = 0.5$ and $\gamma_U = 1.75$. As seen in Fig. 3(b), the details of the image in darker regions are much clearer, while in Fig. 3(c), the details in highlights are more pronounced. Fig. 4 shows the result of the companding operation for various values of α and γ. As seen in the figure, increasing α decreases the limit of higher intensity values and vice versa. Similarly, increasing γ decreases the expansion of lower intensity values.



(a) (b)

Fig. 4. LXT curves for various values of alpha and gamma. (a) Resulting LXT values for the variations in alpha. (b) Resulting LXT curve for changes in gamma.

B. Joint Fusion of Multiblock Global and Local Features

A novel approach to extract and fuse global and local features is provided in this section. Local features define a portion of information about the image in a specific region or single point [41]. In distinction, global features describe the entire image by considering all pixels in the image [42]. The global features provide information regarding the context of the entire image that can be integrated with local features to obtain visually pleasing results with lower artifacts [50]. For image enhancement, the global features could determine the type of scene, subjects in the scene, and lighting conditions to aid local adjustments in the image. In contrast, local features represent the local texture or object at a given location.

The extraction technique is inspired by the UNet architecture that is developed specifically for biomedical image segmentation [18] and ColorNet architecture that was utilized to colorize grayscale images automatically [51]. Both these architectures encompass an end-to-end encoder–decoder network. The UNet architecture focuses mainly on local features, thereby degrading the performance of image enhancement tasks that highly rely on global features [36]. On the contrary, ColorNet utilizes both local and global features; however, the network requires explicit scene labels for training purposes and requires an extra supervised network to compute global features. Both these networks utilize FCN to perform their respective tasks. Even though these networks perform reasonably well, the model efficiency and performance can be enhanced by incorporating a residual layer instead of the FCN block.

The proposed DPIENet comprises of a novel FeCN and a novel FeEN. FeCN aims at producing local and global features. The local features are obtained through a series of layers, while the global features are extracted from every layer of the condense network rather than just the final layer. FeEN aims at reconstructing the enhanced image by exploiting skip connections from FeCN. A flow diagram of DPIENet with FeCN and FeEN can be visualized in Fig. 2.

1) Feature Condense Network: The condense network comprises of feature group, which can be denoted as C_l^g where group $g = 1, 2, \ldots, 8$, and l indicates the number of the residual layer in that particular group and ranges from $1, 2, \ldots, n$. For simplicity, the first feature extraction section is denoted by C^0 and it consists of a convolutional (CONV) layer followed by BN [52] and SELU activation layer [53]. This layer extracts features from the image domain. The CONV layer employs a 3 × 3 kernel and produces 16 feature maps.

The basic structure of the residual layer used in C^{1-8} in the FeCN can be seen in Fig. 2(b) and is formulated in

$$\Theta_{l+1} = S(I(\Theta_l) + \Omega(\omega_l * \Theta_l + b_l)) \,\Big|\, \left\{\omega_l = \left[\varpi_{l,k} : k = 1 \le k \le K\right]\right\} \tag{2}$$

where Θ_l is the input feature map for the lth residual layer, ω_l and b_l are the associated set of weights and biases, respectively, Ω denotes the combination of layers such as CONV→BN→SELU→CONV→BN, S denotes the SELU activation function, and I is the identity map. In groups C^{2-7}, the first layer performs downsampling by striding instead of max pooling since max pool layers lead to high amplitude, high-frequency activations in the subsequent layers, which might increase gridding artifacts [54]. For image enhancement techniques, downsampling may cause loss of spatial information; however, it is required to understand the scenes and reconstruct the image with finer details. Eliminating downsampling may increase resolution; however, it affects the receptive field in subsequent layers, thereby increasing context loss. To overcome this, dilated convolution is employed to adjust receptive fields of feature points without decreasing the resolution of feature maps [55]. It is used in all the layers in the group C^{5-7} instead of traditional convolution, as suggested by Yu *et al.* [54].

Furthermore, to increase the representative power of the global features in the network, the output of the last layer (κ) of each condense group from C^{0-8} is connected to a global average pooling (GAP) layer. The GAP layer compresses the information of the residual layers making it more robust to the spatial translation. The outputs from each layer are concatenated, as shown in

$$y_{\text{fuse}} = \left[C_n^0; C_n^1; C_n^2; \cdots; C_n^8\right]. \tag{3}$$

These features generate a total of $[\sum_{i=0}^{8} \varsigma(C_\kappa^i) \times 1 \times 1]$ where ς is the number of channels/feature maps. The stacked feature maps are then fed into a dense layer D_0, which produces $[\{2 \times \varsigma(C_\kappa^8)\} \times 1 \times 1]$ output, followed by a SELU activation layer and another dense layer D_1 that produces $[\{\varsigma(C_\kappa^8)\} \times 1 \times 1]$ global features. These are replicated to match the dimensions of C_κ^5. Thus, the dimensions of the replicated features are [128 × 32 × 32] (see Table II). The joint fusion comprises stacking the global features from D_1 and the local features from C_κ^5. This aids in incorporating global features into local features. Due to this way of concatenation, the network is independent of any input image resolution restrictions.

2) Feature Enhance Network: Once the local and global feature maps are concatenated, they are fed to the enhance network. The enhance network comprises of feature group, which can be denoted as E_l^g, where group $g = 0, 1, \ldots, 4$ and l indicates the number of the residual layer in that particular group and ranges from $1, 2, \ldots, n$. The feature layers of the condense and enhance network are symmetric across the fusion block, as shown in Fig. 2(a). If the condense group C^2 contains two residual layers, then E^2 also consists of two residual layers.

In the case of the condense layer C^0, E^0consists of just one residual layer. Each enhance group in E^g mainly consists

TABLE II
ARCHITECTURE DETAILS OF THE FECN

Group name		E^1	E^2	E^3	E^4	E^5	E^6	E^7	E^8
Output size considering 512x512 input image		512x512	256x256	128x128	64x64	32x32	16x16	8x8	8x8
[filter size, # feature maps, stride, dilation]	E_0	[3,16,1,1]	[3,32,2,1]	[3,64,2,1]	[3,128,2,1]	[3,128,2,2]	[3,128,2,4]	[3,128,2,2]	[3,128,1,1]
	E_{1-n}	[3,16,1,1]	[3,32,1,1]	[3,64,1,1]	[3,128,1,1]	[3,128,1,2]	[3,128,1,4]	[3,128,1,2]	[3,128,1,1]

of upsampling layers, compression layers, and residual layers. The input to each enhance group is the fusion of feature maps from the previous enhance group and the output of the corresponding condense group. This helps in propagating context information to higher resolution layers. The upsampling layer consists of transposed convolutions with the kernel size 2 × 2 and stride 2 × 2. This aids in increasing the resolution of the feature maps by a factor of 2. The compressing layer consists of CONV→BN→SELU, wherein the kernel size of CONV is 1 × 1. This is used to compress the feature dimensions by a factor of 2. The compressed feature maps are then fed to the residual layers for further processing. Finally, the output of the group E^0 is connected to a CONV layer with kernel size 3 × 3, and residual learning is adopted by adding the input image to this layer.

3) Dynamic Channel Attention Mechanism: Most of the deep learning-based image enhancement techniques consider all the feature maps equally, which may not be correct in many real-world cases. Among the residual layers' generated feature maps, few of the features might contribute more when compared to the rest. Moreover, the learned filters in the residual layers have a local receptive field, and each filter output exploits the contextual information outside of the subregion very poorly. Thus, a mechanism is required to recalibrate features such that more emphasis is provided for the feature maps with better mapping compared to the less essential feature maps. Researchers have offered tentative work to apply attention in deep neural networks [56]–[58], which ranges from localization and understanding in images [59] to sequence-based networks [60]. However, these attention mechanisms are not yet mature for low-level vision tasks such as image enhancement.

This mechanism's main objective is to assign different values to various channels according to their interdependencies in each convolution layer. Thus, to increase each channel's sensitivity, an intuitive way is to access the global spatial information by using average pooling over the entire feature map. The channel attention mechanism can be formulated, as shown in

$$\Theta = \sigma\left(W_{\uparrow}\left(S\left(W_{\downarrow}\left(1/(H \times W)\sum_{m=0}^{H-1}\sum_{n=0}^{W-1}(\Phi)\right)\right) + b_{\downarrow}\right) + b_{\uparrow}\right) \tag{4}$$

where $\Phi = [\Phi_1, \Phi_2, \ldots, \Phi_{\varsigma}]$ is the input feature map with ς number of channels/feature maps and $H \times W$ dimensions, $W_{\downarrow}[b_{\downarrow}]$ denotes weight [bias] of the compression convolution, which reduces the dimension by a factor of r, $W_{\uparrow}[b_{\uparrow}]$ denotes weight [bias] of the expansion convolution, which increases the dimension by a factor of r, S denotes the SELU activation function, and σ is the sigmoid activation function. The GAP output can be realized as the fusion of local descriptors whose statistics express the entire feature map [56].

The channel attention mechanism comprises of the convolutions with kernel size 1 × 1 along with the sigmoid activation. This aids in learning the nonlinear interaction between the channels and ensures multiple channels with informative maps are emphasized more [56]. As the number of channels/feature maps ς in the condense and enhance network keeps varying, the gating mechanism needs to be adjusted to accommodate these changes. The factor r is a hyperparameter, which varies the capacity of the gating mechanism. The ratio r was formulated as $r = \varsigma^i/4$ where ς^i denotes the number of channels/feature maps at the input of the GAP layer.

IV. MULTISCALE HUMAN COLOR VISION LOSS

Several loss functions, such as L1, L2, cosine similarity measures [61], and perceptual and adversarial losses [36], have been investigated for various computer vision tasks. These perform reasonably well, but losses based on dense pixelwise image differences lead to poor perceptual quality [33]. In [47], an HDR cost function that treats illumination and reflectance separately was proposed. However, the method utilized only the information around the predicted image's saturated areas to compute the loss. This pixelwise blending-based cost function will be ineffective for image enhancement tasks that require global and local adjustments. Thus, in this article, a multiscale loss function that works on the principle of the Retinex theory is proposed. According to this, the low-frequency information of the image represents the global naturalness, and the high-frequency information represents the local details. By decomposing the image into a low-frequency luminance component and a high-frequency detail component, the loss function incorporates both the local and global information. This loss is driven by the close to the logarithmic response of the human visual system (HVS) in large luminance range areas, which follows Weber-Fechner's law [62].

The loss is constructed under the assumption that the image can be decomposed into illuminance and reflectance components. The illumination component $\mathcal{L}$ defines the global deviations in an image, while the reflectance $\mathcal{R}$ represents the details and colors. In combination, these components modulate the reconstruction of a perceptually enhanced image $P_e = \mathcal{L} \times \mathcal{R}$. For the simplicity of exposition, consider the case in which the loss function consists of a single scale: the extension to multiple scales is straightforward. Consider a predicted image I and ground-truth image T of any arbitrary size (m, n). The log-based illumination component is

TABLE III
PERFORMANCE COMPARISON BETWEEN PROPOSED ARCHITECTURES ON THE MIT-ADOBE 5K DATASET. THESE ARE AN AVERAGE OF 500 IMAGES* FROM THE TEST DATASET WITH DIFFERENT EXPOSURE SETTINGS. AS SEEN IN THE MHCV LOSS WITH LXT AND DCA SHOWS THE BEST PERFORMANCE

Losses – DPIENet	LXT	DCA	PSNR (dB) (↑)	SSIM (↑)	GSSIM (↑)	UQI (↑)
L1	✓	✓	23.2145	0.8869	0.8315	0.8727
MSE	✓	✓	22.6766	0.8802	0.8099	0.8688
Cosine [61]	✓	✓	23.3323	0.8791	0.8204	0.8747
SSIM	✓	✓	23.6444	0.8979	0.8510	0.8758
Single Scale HCV - $\sigma = 0.5$	✓	✓	24.0237	0.8955	0.8470	0.8835
Single Scale HCV - $\sigma = 8$	✓	✓	24.1297	0.8985	0.8489	0.8836
MHCV	X	X	21.8469	0.8670	0.7951	0.8588
MHCV	✓	X	23.3174	0.8733	0.8050	0.8568
MHCV	✓	✓	**24.2102**	**0.9013**	**0.8494**	**0.8847**

* The 500 images in the test dataset consist of 250 low exposure and 250 standard exposure images.

constructed by employing a center/surround algorithm, particularly, a Gaussian filter $\mathcal{G}_\sigma$, which can be formulated, as shown in

$$\mathcal{L}_\sigma^\Psi = \log\left(\mathcal{G}_\sigma \otimes \Psi^2\right)\sigma\epsilon\{0.5, 1, 2, 4, 8\}$$
$$\text{where} \quad \mathcal{G}_\sigma = \frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5}$$

where $\otimes$ denotes convolution and for the illumination component of predicted image, Ψ takes the value of I and $\Psi = T$ for ground-truth image. The value of σ cannot be theoretically modeled and determined [63]. The choice of right scale σ for the surround filter is crucial for single scale retinex. These can be overcome by utilizing the multiscale retinex, which seems to afford an acceptable tradeoff between a good local dynamic range and a good color rendition. Thus, empirically, σ values were set to 0.5, 1, 2, 4, and 8. The *log*-based reflectance component is constructed by taking the difference between the image and illumination component. This can be formulated, as shown in (6). The resulting MHCV loss function using these two components can be defined, as shown in (7), as follows:

$$\mathcal{R}_\sigma^\Psi = \log\left(\Psi^2\right) - \mathcal{L}_\sigma \tag{6}$$

$$\text{MHCV} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\alpha}{n}\sum_{j=1}^{n}\left(\mathcal{L}_{\sigma_i,j}^{\mathrm{T}} - \mathcal{L}_{\sigma_i,j}^{I}\right)^2 + \frac{1-\alpha}{n}\sum_{j=1}^{n}\left(\mathcal{R}_{\sigma_i,j}^{\mathrm{T}} - \mathcal{R}_{\sigma_i,j}^{I}\right)^2\right]$$
$$N = \dim(\sigma);\ \ \alpha = 0.5. \tag{7}$$

Equal weight is provided to both illumination and reflectance components as both global variations of illuminance and local colors, and details are very important for the successful reconstruction of enhanced images.

V. EXPERIMENTAL RESULTS

This section provides the performance evaluation of the DPIENet. After outlining the experimental settings, chosen datasets, and training details, the performance comparisons with SOTA methods are provided to demonstrate the effectiveness and generality of the DPIENet.

A. Dataset

For training, validation, and testing purposes, the MIT-Adobe FiveK dataset [64] is employed. This dataset contains 5000 photographs taken with SLR cameras by various photographers. These photographs covered a broad range of scenes, objects, subjects, and lighting conditions. Each image was retouched by five well-trained photographers using global and local adjustments. Among these retouchers, the result of photographer C was selected as ground truth because the photographs received a high rank in the user study [64]. The untouched images were considered as input images. This consisted of images with standard exposure (Λ_S), which comprises of images captured with default camera settings and low exposure (Λ_L) involves simulated low exposure settings. The dataset was split into three partitions: 4000 images for training, and 500 images (250 low + 250 std exposure) for validation and testing. All the images from this dataset were downsized to 512 along the long side for training, validation, and testing purposes.

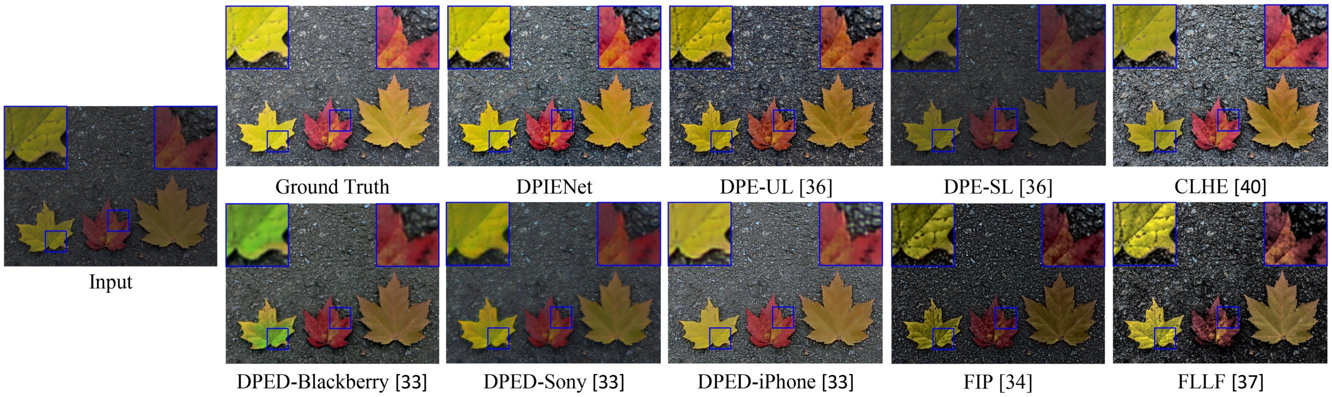
B. Training Details

For training, RGB input patches of size 256×256 along with the corresponding ground truth were considered. The training data were augmented using random horizontal, vertical, and 90° rotations along the center of the image. According to [53], the ideal initialization for SELU is mean 0 and standard deviation $\sqrt{1/n}$. However, this unequivocally causes the gradients to explode. To stabilize the network, the standard deviation was set to $\sqrt{0.1/n}$. For training the model, the AdaBound optimizer [65] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1 \times 10^{-8}$, and $\gamma = 1 \times 10^{-3}$ was employed. The batch size was set to 20. The learning rate was initialized as $1e^{-3}$ and the final learning rate was initialized as 0.1. The network was trained for a total of 2.85×10^6 updates and multistep learning rate scheduler was used to decrease the learning rate by 0.1 at 9.5×10^5, 1.9×10^6, and 2.375×10^6 iterations. For training, the proposed multiscale human vision loss was employed instead of L1 and L2 loss. Minimizing L2 is generally preferred as it maximizes the PSNR. However, based on a series of experiments conducted, MHCV loss provides better convergence than L1 or L2 loss. The evaluation of this comparison is provided in the next section.

TABLE IV
Quantitative Evaluation of DPIENet With SOTA on MIT-Adobe FiveK Dataset for Standard (Λ_S) and Low Exposure (Λ_L) Inputs. These Are an Average of 250 Images From the Test Dataset. *Red* Text Indicates the Best and *Blue* Text Indicates the Second-Best Performance for Respective Input Settings. This Demonstrates That the Proposed DPIENet Performs Significantly Better Than SOTA Techniques

Methods	Type	Input Type	Params (K)	PSNR (dB) (↑)	SSIM (↑)	GSSIM (↑)	UQI (↑)
CLHE [40]	Classical	Λ_S	-	16.5091	0.7503	0.7054	0.7828
FLLF [37]	Classical		-	17.0479	0.6680	0.5744	0.7581
FIP [34]	Deep Learning		37.2	16.7871	0.7425	0.6587	0.7556
DPED - blackberry [33]	Deep Learning		401.5	**20.6017**	0.7621	0.5991	**0.8627**
DPED - iPhone [33]	Deep Learning			19.2550	0.7343	0.5425	0.8254
DPED - Sony [33]	Deep Learning			19.7129	0.7353	0.5554	0.8555
DPE supervised [36]	Deep Learning		3335.4	18.0833	0.7914	**0.7580**	0.8125
DPE unsupervised [36]	Deep Learning			18.5277	**0.7990**	0.7405	0.8211
DPIENet	Deep Learning		4268.6	**23.4189**	**0.8974**	**0.8475**	**0.8803**
BIMEF [39]	Classical	Λ_L	-	17.4640	0.7809	0.7548	0.8037
LIME [26]	Classical		-	13.2663	0.7350	0.6900	0.692
SRIE [38]	Classical		-	18.2237	0.8203	0.7886	0.8255
MBLLEN [44]	Deep Learning		450.2	18.8487	0.8033	0.6986	0.7771
GLADNet [42]	Deep Learning		931.5	16.5113	0.7466	0.6505	0.7636
RetinexNet [41]	Deep Learning		444.6	12.4149	0.6767	0.5540	0.6940
EnlightenGAN [45]	Deep Learning		8636.6	15.6517	0.7686	0.7117	0.7570
DEEPUPE [46]	Deep Learning		998.8	**22.0973**	**0.8569**	**0.8181**	**0.8574**
DPIENet	Deep Learning		4268.6	**25.0017**	**0.9052**	**0.8512**	**0.8890**

* The 500 images in the test dataset consists of 250 low exposure and 250 standard exposure images.


Input

Ground Truth

DPIENet

DPE-UL [36]

DPE-SL [36]

CLHE [40]

DPED-Blackberry [33]

DPED-Sony [33]

DPED-iPhone [33]

FIP [34]

FLLF [37]

Fig. 5. Visual comparisons with respect to the ground truth. Zoom-in regions are used to illustrate the visual difference. DPIENet not only restores the details but also avoids discoloration. The SOTA techniques tend to exhibit few artifacts, such as variation in color (for example, DPE-UL tends to shift the color toward orange from red, DPED-Blackberry introduced green color), over enhancement (for example, FLLF and FIP over enhance the detail which look dark), and blurriness (for instance, DPED-Sony image look smoothened). Note: UL stands for unsupervised learning, and SL stands for supervised learning.

C. Benchmark Results

DPIENet is compared with other SOTA algorithms using measures, such as PSNR, SSIM [2], GSSIM [3], and UQI [4]. These measures are applied to all the RGB channels of the image. All these measures access the image quality based on the given reference benchmark image that is assumed to have the desired quality [66]. Higher quality value depicts how close the enhanced images are to the ground truth.

The ablation tests comprise of experiments exploring different designs and exposure settings. The quantitative performance of different models is provided in Table III. When the LXT and DCA mechanism is removed from the network, the performance is relatively low. For example, in terms of PSNR, DPIENet without LXT and DCA reaches 21.84 dB; when LXT is added, it increases to 23.31 dB. When both LXT and DCA are combined, it reaches 24.21 dB. This indicates that the proposed LXT+DCA mechanism, along with stacking, is much more powerful than the residual block-stacking method and gives a boost in performance roughly by a factor of 2.3 dB.

Furthermore, to show the effectiveness of DPIENet with MHCV loss, a comparison with existing losses, such as L1, L2, SSIM, Cosine, and single scale HCV loss, is also provided in Table III. This was obtained by applying PSNR on 500 images (a combination of both low and standard exposure) from the validation set. It can be inferred that MHCV loss outperforms with a higher margin of improvements when compared to L1 and L2 loss. The single scale HCV loss performs fairly; however, PSNR fluctuates for each scale; for example, when $\sigma = 0.5$, PSNR is 24.02 and when $\sigma = 0.5$, PSNR is 24.12. To overcome this variation, multiple sigma levels in MHCV are utilized and it performs slightly better than the single scale HCV loss.

The proposed network is compared with SOTA methods for standard and low exposure settings. For standard

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

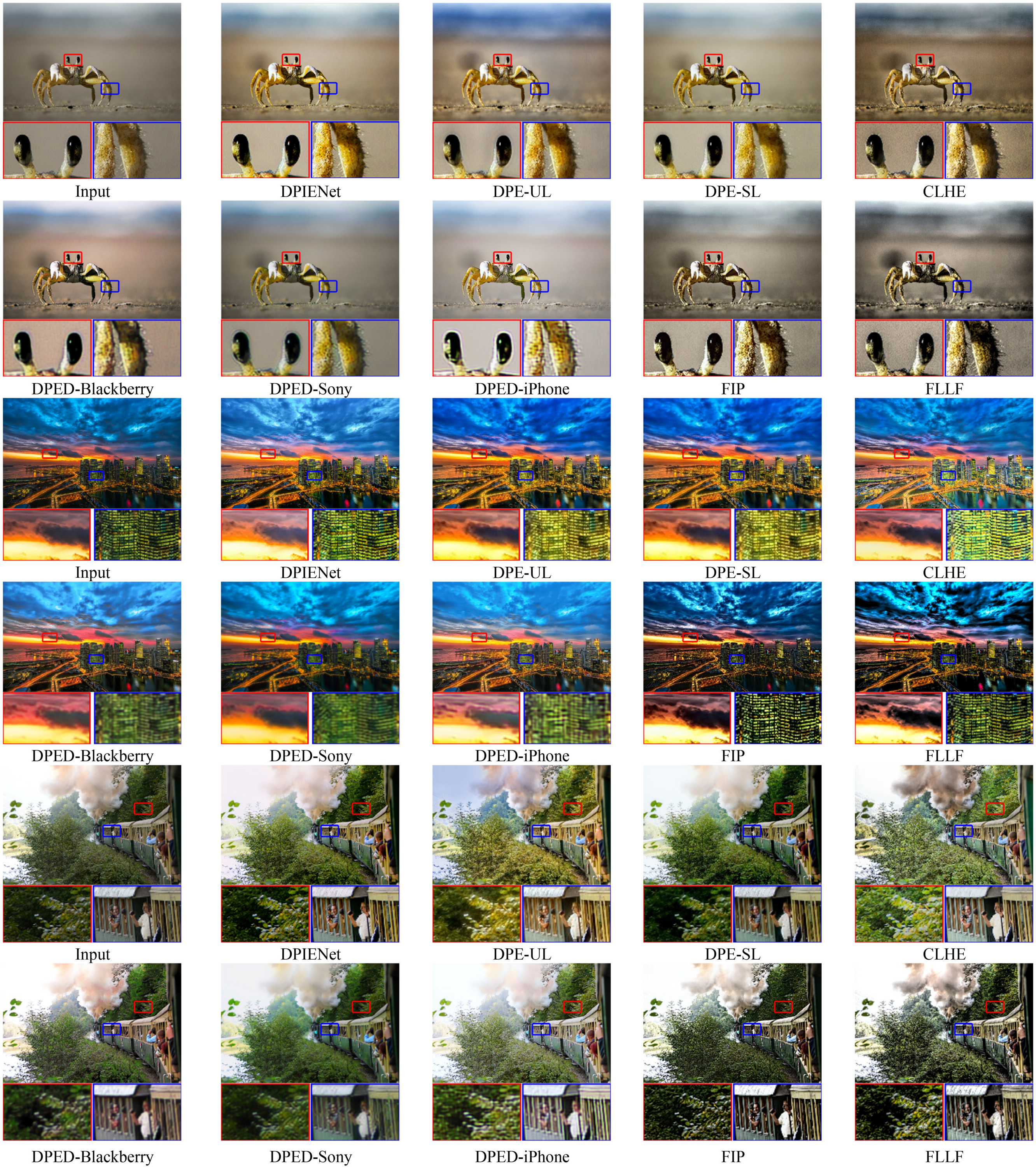
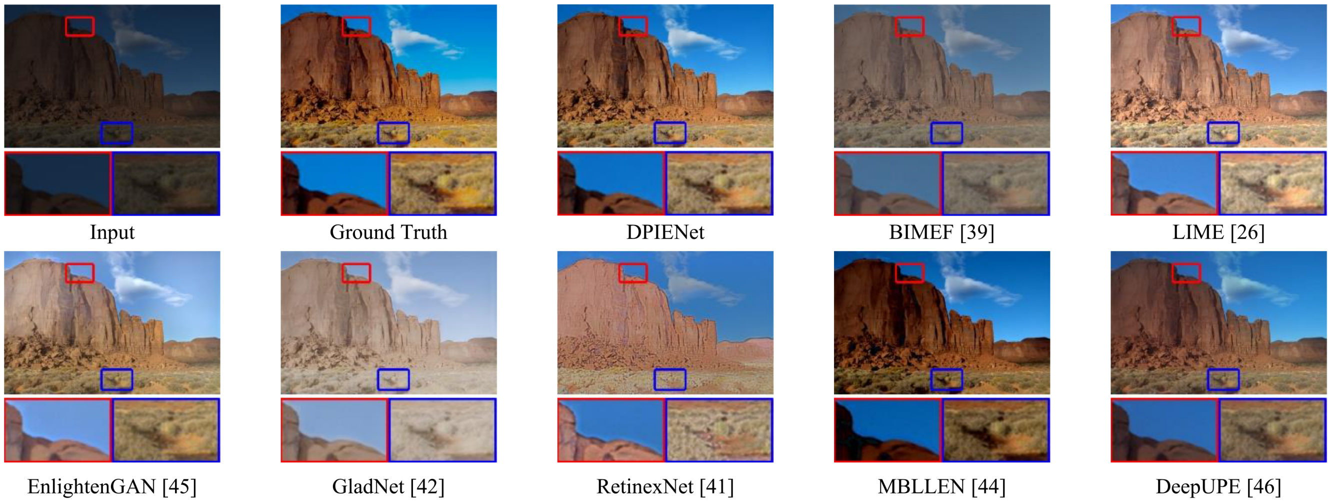


Fig. 6. Real-world visual comparisons of DPIENet with the SOTA models. Zoom-in regions are used to illustrate the visual difference. In the first example, DPIENet successfully suppresses the noise, which is visible in CLHE, FIP, and FLLF. Furthermore, it does not have halo artifacts that are introduced by DPE-UL and DPED. In the second example, the structural details of the building are preserved when compared to DPE-UL and CLHE. In the third example, the color of the leaves is preserved when compared to the other techniques. DPE-UL has introduced blue sky, which is not present in the input, and the leaves are yellow. In all the examples, DPED introduces blurring, FIP, and FLLF generate underexposed/darker images.

exposure input setting, several recent competing methods, such as CLHE [40], FLLF [37], DPE supervised and unsupervised [36], DPED trained with Blackberry, iPhone, and Sony images [33], and FIP [34], were considered. Table IV demonstrates that DPIENet performs significantly better when compared to the other methods. The visual comparison is provided in Figs. 5 and 6. Fig. 5 illustrates that the enhanced colors of the DPIENet are very similar to the



Input

Ground Truth

DPIENet

BIMEF [39]

LIME [26]

EnlightenGAN [45]

GladNet [42]

RetinexNet [41]

MBLLEN [44]

DeepUPE [46]

Fig. 7. Demonstration of low exposure image performance using various models. Zoom-in regions are used to illustrate the visual difference. DeepUPE generates an image with a soft haze effect; MBLLEN produces dark images. EnlightenGan and GladNet introduce a foggy effect. However, DPIENet restores details but also avoids various artifacts and provides results similar to the ground truth.

ground truth, while Fig. 6. provides results of a few real-world examples. For real-world images, NASA dataset [67], Google HDR [5], DIV2K dataset [68], and a database provided in [45] were utilized. The zoomed regions in both these images demonstrate the color and edge-preserving property of DPIENet when compared to the SOTA techniques, which tend to oversaturate, introduce variations in color, and induce blurriness.

The quantitative results for low exposure settings are provided in Table IV. This indicates that the images are restored with superior quantitative performance. The visual comparison of this setting is illustrated in Fig. 7 (with ground truth) and Fig. 8 (real world). The network reconstructed a visually pleasing image close to the ground truth and mimic human perception while retaining natural color rendition. In comparison, the SOTA techniques contain exposure artifacts, and the colors are less perceptually similar when compared to the ground truth.

Furthermore, the model is compared with the most recent deep learning-based competing low light IE techniques, such as MBLLEN [34], EnlightenGAN [35], DEEPUPE [36], GLADNet [32], and RetinexNet [30]. The proposed network reconstructs perceptually improved images with a higher correlation with the ground truth when compared to the other models.

The merged images from the Google HDR [5] dataset were utilized to show the effectiveness of DPIENet on real-world images. This dataset contains 153 sets of images—each set comprises of a merged image and a final reconstructed image along with a reference frame. As DPIENet aims at exposure correction, the merged images were used as inputs to the systems. To compute the quality, no reference-based quality measure, such as CRME [70], Brisque [71], and Divine [72], were utilized. Comparative results are provided in Table V. Due to the supervised training of DPIENet, it has to be noted that it tries to enhance the image so that it is close to the reference image, and thus, it is not optimized for the no-reference-based measure. This is indicated by the marginally better results obtained by DPIENet in comparison to other methods.

TABLE V
PERFORMANCE COMPARISON BETWEEN PROPOSED ARCHITECTURES ON THE GOOGLE HDR DATASET. THESE ARE AN AVERAGE OF 153 IMAGES. *Highlighted* TEXT INDICATES THE TOP THREE PERFORMANCE

Methods	CRME [71] (↑)	Brisque [72] (↓)	Divine [73](↓)
CLHE [40]	**1.2348**	**20.1794**	13.1394
FIP [34]	**1.2415**	22.6112	**11.4606**
DPED [33]	1.0614	28.6947	28.4124
DPE [36]	1.0335	24.0821	20.2975
BIMEF [39]	1.0862	29.2549	14.8724
LIME [26]	1.1825	31.8902	**11.0948**
SRIE [38]	1.1038	30.4923	15.9013
DEEPUPE [46]	1.1072	**22.0752**	14.2188
MBLLEN [44]	1.1143	32.0196	17.6881
GLADNet [42]	1.1037	33.6988	17.7910
RetinexNet [41]	1.1574	37.2537	13.9779
EnlightenGAN [45]	1.1026	27.8058	15.3681
DPIENet	**1.2574**	**19.3760**	**11.6013**

VI. USER STUDY

The user study conducted follows the practice provided in [72]. A paired comparison is adopted to assess the perceptual quality using Qualtrics [73]. For each test, each user was asked to select the preferred one from a pair of images. Using this setup, relative scores and standard exposure input images show minimal perceptual differences between the proposed DPIENet and the SOTA methods, such as CLHE, FLLF, DPED-iPhone, and DPE-unsupervised, for standard exposure methods, and MBLLEN, GLADNet, RetinexNet, EnlightenGAN, and DEEPUPE for low exposure methods are obtained.

For this study, five images per comparison were picked randomly from the Adobe FiveK dataset (testing and validation

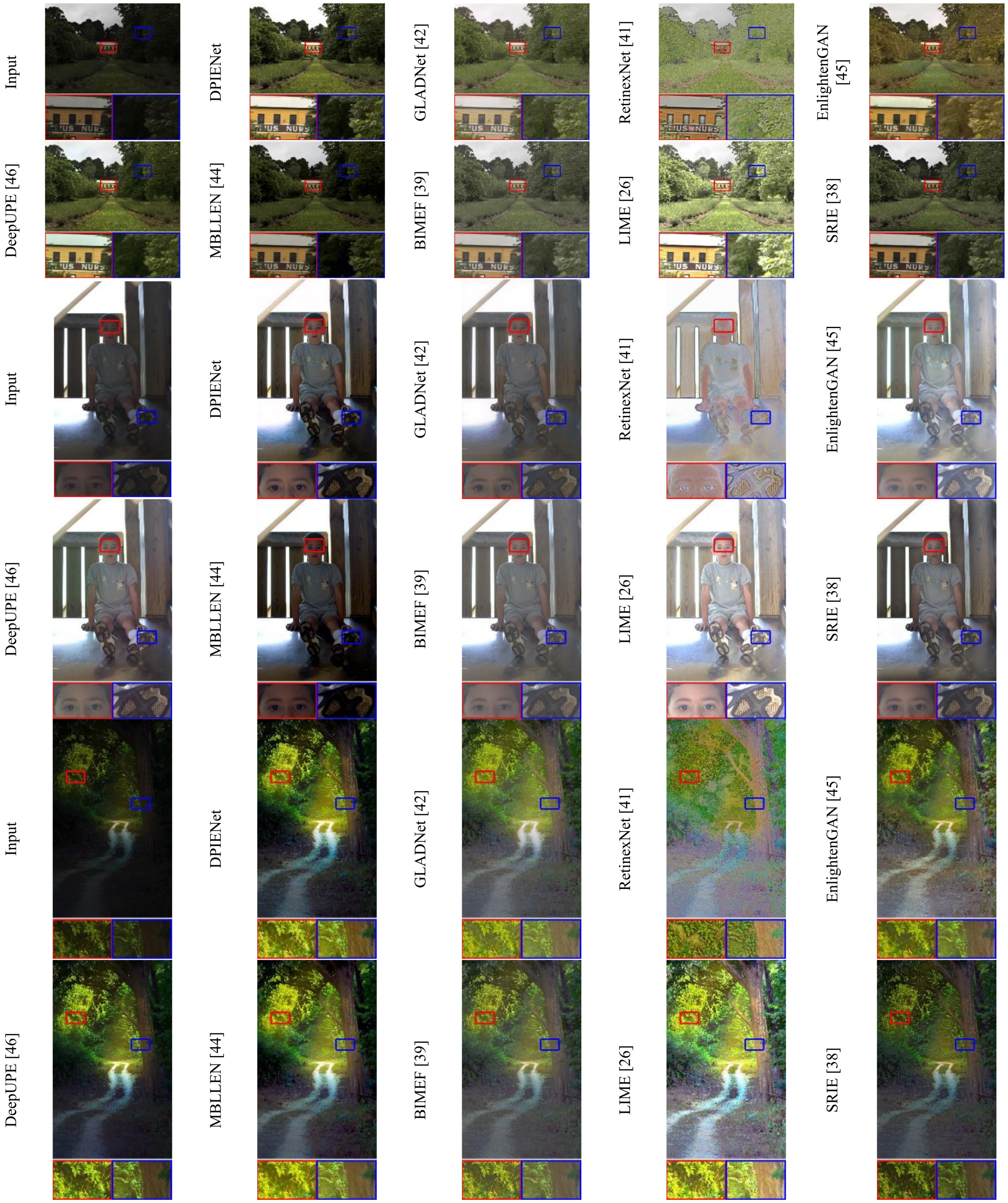


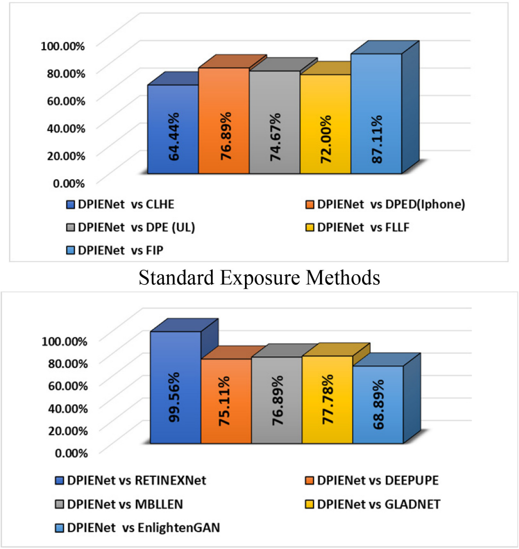
Fig. 8. Real-world visual comparison of DPIENet with SOTA low exposure methods. Zoom-in regions are used to illustrate the visual difference. The first example, DPIENet, produces visually pleasing realistic colors. DeepUPE and MBLLEN do produce realistic colors; however, they introduce exposure artifacts. The second example, DPIENet, produces images with better details (see zoomed shoe). The third example, DPIENet, provides better visible details and color, as seen in the zoomed regions. Overall, EnlightenGAN and RetinexNet tend to produce unrealistic colors. GLADNet introduces a hazy effect, and DEEPUPE and MBLLEN suffer from exposure-related artifacts.

images) [64], NASA dataset [67], Google HDR [5], DIV2K dataset [68], and a database provided in [45]. Each participant was asked to compare 50 pairs of images. The users were instructed to consider the following aspects: 1) visible noise; 2) over or underexposure artifacts; 3) overenhancement; and 4) unrealistic color or texture distortions. For detailed analysis,

TABLE VI
BT SCORES FOR IMAGE ENHANCEMENT IN THE USER STUDY. THE PROPOSED DPIENET PERFORMS FAVORABLY AGAINST OTHER SOTA COMPARISONS

Methods	Scores
Λ_S	
DPIENet	0.9554
CLHE	0.3607
DPED	-0.2465
DPE	-0.1254
FLLF	0.0110
FIP	-0.9553

Methods	Scores
Λ_L	
DPIENet	1.6277
RetinexNet	-3.7840
DeepUPE	0.5231
MBLLEN	0.4256
GladNet	0.3749
EnlightenGAN	0.8327



Standard Exposure Methods

Low Exposure Methods

Fig. 9. Analysis of user study. The bar plot provides the percentage number of times the users selected DPIENet versus the SOTA method. The DPIENet was preferred by an average of 76% and 79% of users on the standard and low exposure settings, respectively. Note: 76% and 79% are obtained by averaging all the bars on the graph.

the results from 45 participants were considered. The percentage that users chose DPIENet over the SOTA methods for both low and standard exposure images is provided in Fig. 9. The bar plot provides the number of times the user preferred DPIENet versus the SOTA method. For example, DPIENet was chosen 64.44% of the time when compared to CLHE under standard exposure methods. On average, the proposed DPIENet is preferred by 76% and 79% of users for standard and low exposure settings, respectively. These averages are obtained by taking the mean of the graph bars of Fig. 9. The runner-up was CLHE for Λ_S and EnlightenGAN for Λ_L methods. For further analysis, the global score was obtained by fitting the results of paired comparisons to the Bradley–Terry (BT) model [74]. The normalized zero mean BT score for both exposures is quantized in Table VI. These scores, along with the user study, shows that the results of the proposed method have higher perceptual quality than existing SOTA methods.

VII. CONCLUSION

In this work, a novel deep learning-based image enhancement for exposure restoration is presented. The method is built on multiexposure simulation using LXT. The proposed DPIENet, which is an end-to-end mapping approach, comprises of a condense and enhance network, which leverages the idea of residual learning to reach a larger depth. Furthermore, the skip connection between these networks aids in recovering spatial information while upsampling. In addition, to improve the network's ability to realize the context of the image, global features are exploited from each group in the condense network. A DCA mechanism to adaptively rescale channelwise features is employed to boost the network's channel interdependencies further. To obtain realistic images that correlate to human vision, a novel multiscale human vision loss is presented—these aid in accounting for the global variation in illumination, details, and colors. Extensive quantitative, qualitative, and user study evaluations conducted on the presented technique demonstrate DPIENet's performance surpasses the existing methods and achieves SOTA results. Furthermore, DPIENet overcomes artifacts, such as halo effects, noise amplification in dark regions, and artificial color generation, which occur in a few existing techniques. As a part of the future work, the authors intend to test the accuracy of the system for various low-level computer vision tasks, such as super-resolution, image recoloring, and image denoising.

ACKNOWLEDGMENT

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Combat Capabilities Development Command, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, Feb. 2018, Art. no. 7068349, doi: 10.1155/2018/7068349.

[2] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[3] S. Nercessian, S. S. Agaian, and K. A. Panetta, "An image similarity measure using enhanced human visual system characteristics," in *Proc. SPIE Defense Security Sens.*, 2011, Art. no. 806310.

[4] W. Zhou and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002, doi: 10.1109/97.995823.

[5] S. W. Hasinoff *et al.*, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Trans. Graph.*, vol. 35, no. 6, p. 192, 2016.

[6] J. C. Russ and F. B. Neal, *The Image Processing Handbook*. Boca Raton, FL, USA: CRC Press, 2018.

[7] W. Yu, "Practical anti-vignetting methods for digital cameras," *IEEE Trans. Consum. Electron.*, vol. 50, no. 4, pp. 975–983, Nov. 2004.

[8] M. J. Nadenau, J. Reichel, and M. Kunt, "Wavelet-based color image compression: Exploiting the contrast sensitivity function," *IEEE Trans. Image Process.*, vol. 12, pp. 58–70, 2003.

[9] K. Panetta, S. Agaian, Y. Zhou, and E. J. Wharton, "Parameterized logarithmic framework for image enhancement," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 2, pp. 460–473, Apr. 2011.

[10] K. A. Panetta, E. J. Wharton, and S. S. Agaian, "Human visual system-based image enhancement and logarithmic contrast measure," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 1, pp. 174–188, Feb. 2008.

[11] R. C. Gonzales and R. E. Woods, *Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice Hall, 2002.

[12] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice Hall, 1989.

[13] E. H. Land and J. J. McCann, "Lightness and retinex theory," *J. Opt. Soc. America*, vol. 61, no. 1, pp. 1–11, 1971.

[14] Z.-u. Rahman, D. J. Jobson, and G. A. Woodell, "Multi-scale retinex for color image enhancement," in *Proc. 3rd IEEE Int. Conf. Image Process.*, vol. 3, 1996, pp. 1003–1006.

[15] J. Hu, H. Gao, Z. Zhang, G. Lin, H. Wang, and W. Liu, "A novel image enhancement method based on variational retinex approach," in *Proc. IOP Conf. Ser. Mater. Sci. Eng.*, vol. 452, Dec. 2018, Art. no. 042202, doi: 10.1088/1757-899x/452/4/042202.

[16] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3291–3300.

[17] H. Jiang and Y. Zheng, "Learning to see moving objects in the dark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7324–7333.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2015, pp. 234–241.

[19] H. Sawant and M. Deore, "A comprehensive review of image enhancement techniques," *Int. J. Comput. Technol. Electron. Eng.*, vol. 1, no. 2, pp. 39–44, 2010.

[20] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Comput. Vis. Graph. Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.

[21] S. M. Pizer, "Contrast-limited adaptive histogram equalization: Speed and effectiveness," presented at the Proc. 1st Conf. Visualization Biomed. Comput., Atlanta, GA, USA, May 1990.

[22] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 593–600, May 2007.

[23] D. J. Jobson, Z.-U. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Trans. Image Process.*, vol. 6, pp. 451–462, 1997.

[24] Q. Zhang, G. Yuan, C. Xiao, L. Zhu, and W.-S. Zheng, "High-quality exposure correction of underexposed photos," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 582–590.

[25] C.-H. Lee, J.-L. Shih, C.-C. Lien, and C.-C. Han, "Adaptive multiscale retinex for image contrast enhancement," in *Proc. Int. Conf. Signal-Image Technol. Internet-Based Syst.*, 2013, pp. 43–50.

[26] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, pp. 982–993, 2017.

[27] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Trans. Image Process.*, vol. 22, pp. 3538–3548, 2013.

[28] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and J. Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal Process.*, vol. 129, pp. 82–96, Dec. 2016.

[29] Q.-C. Tian and L. D. Cohen, "A variational-based fusion model for non-uniform illumination image enhancement via contrast optimization and color correction," *Signal Process.*, vol. 153, pp. 210–220, Dec. 2018.

[30] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2014, pp. 2672–2680.

[31] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.

[32] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, "Automatic photo adjustment using deep neural networks," *ACM Trans. Graph.*, vol. 35, no. 2, pp. 1–15, 2016.

[33] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, "DSLR-quality photos on mobile devices with deep convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3277–3285.

[34] Q. Chen, J. Xu, and V. Koltun, "Fast image processing with fully-convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2497–2506.

[35] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, 2017.

[36] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 6306–6314. [Online]. Available: https://ieeexplore.ieee.org/document/8578758/

[37] M. Aubry, S. Paris, S. W. Hasinoff, J. Kautz, and F. Durand, "Fast local laplacian filters: Theory and applications," *ACM Trans. Graph.*, vol. 33, no. 5, p. 167, 2014.

[38] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2782–2790.

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

[40] S. Wang, W. Cho, J. Jang, M. A. Abidi, and J. Paik, "Contrast-dependent saturation adjustment for outdoor image enhancement," *J. Opt. Soc. America A*, vol. 34, no. 1, pp. 7–17, 2017.

[41] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.

[42] W. Wang, C. Wei, W. Yang, and J. Liu, "GLADNet: Low-light enhancement network with global awareness," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2018, pp. 751–755.

[43] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, pp. 2049–2062, 2018.

[44] F. Lv, F. Lu, J. Wu, and C. Lim, "MBLLEN: Low-light image/video enhancement using CNNs," in *Proc. BMVC*, 2018, p. 220.

[45] Y. Jiang *et al.*, "EnlightenGAN: Deep light enhancement without paired supervision," 2019, *arXiv:1906.06972*.

[46] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6849–6857.

[47] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Trans. Graph.*, vol. 36, no. 6, p. 178, 2017.

[48] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 177:1–177:10, 2017.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[50] D. Nie, L. Wang, E. Adeli, C. Lao, W. Lin, and D. Shen, "3-D fully convolutional networks for multimodal isointense infant brain image segmentation," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1123–1136, Mar. 2019.

[51] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color! Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, p. 110, 2016.

[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[53] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2017, pp. 971–980.

[54] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. comput. Vis. Pattern Recognit.*, 2017, pp. 472–480.

[55] L. Cui *et al.*, "Context-aware block net for small object detection," *IEEE Trans. Cybern.*, early access, Jul. 28, 2020, doi: 10.1109/TCYB.2020.3004636.

[56] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[57] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3156–3164.

[58] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: A fast and lightweight network for single-image super resolution," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1443–1453, Mar. 2021.

[59] C. Cao *et al.*, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2956–2964.

[60] T. Bluche, "Joint line segmentation and transcription for end-to-end handwritten paragraph recognition," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2016, pp. 838–846.

[61] D. Marnerides, T. Bashford-Rogers, J. Hatchett, and K. Debattista, "ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content," *Comput. Graph. Forum*, vol. 37, no. 2, pp. 37–49, 2018.

[62] G. T. Fechner, D. H. Howes, and E. G. Boring, *Elements Psychophysics*. New York, NY, USA: Holt, Rinehart Winston, 1966.

[63] A. B. Petro, C. Sbert, and J.-M. Morel, "Multiscale retinex," *Image Process. On-Line*, vol. 4, pp. 71–88, Apr. 2014.

[64] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *Proc. CVPR*, 2011, pp. 97–104.

[65] L. Luo, Y. Xiong, Y. Liu, and X. Sun, "Adaptive gradient methods with dynamic bound of learning rate," 2019, *arXiv:1902.09843*.

[66] S. Wang, C. Deng, W. Lin, G.-B. Huang, and B. Zhao, "NMF-based image quality assessment using extreme learning machine," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 232–243, Jan. 2017.

[67] "Retinex Image Processing." [Online]. Available: https://dragon.larc.nasa.gov/retinex/pao/news/ (Accessed: 2018).

[68] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 126–135.

[69] K. Panetta, C. Gao, and S. Agaian, "No reference color image contrast and quality measures," *IEEE Trans. Consum. Electron.*, vol. 59, no. 3, pp. 643–651, Aug. 2013.

[70] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, pp. 4695–4708, 2012.

[71] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, pp. 3350–3364, Dec. 2011.

[72] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019.

[73] "Qualtrics." [Online]. Available: https://www.qualtrics.com/ (Accessed: 2020).

[74] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, nos. 3–4, pp. 324–345, 1952.

Karen Panetta (Fellow, IEEE) received the B.S. degree in computer engineering from Boston University, Boston, MA, USA, and the M.S. and Ph.D. degrees in electrical engineering from Northeastern University, Boston.

She is currently the Dean of Graduate Engineering Education, a Professor with the Department of Electrical and Computer Engineering, and an Adjunct Professor of Computer Science with Tufts University, Medford, MA, USA, and the Director of the Dr. Panetta's Vision and Sensing System Laboratory. Her research focuses on developing efficient algorithms for simulation, modeling, signal, and image processing for biomedical and security applications.

Prof. Panetta was a recipient of the 2012 IEEE Ethical Practices Award and the Harriet B. Rigas Award for Outstanding Educator. In 2011, she was awarded the Presidential Award for Engineering and Science Education and Mentoring by U.S. President Obama. She was inducted into the National Academy of Inventors in 2021. She was the President of the IEEE-HKN—2019. She is the Editor-in-Chief of the *IEEE Women in Engineering magazine*. She was the IEEE-USA Vice-President of Communications and Public Affairs. From 2007 to 2009, she served as the world-Wide Director for IEEE Women in Engineering, overseeing the world's largest professional organization supporting women in engineering and science.

Shreyas Kamath K. M. (Student Member, IEEE) received the B.E. degree in electronics and communication engineering from Visvesvaraya Technological University, Belgaum, India, the M.S. degree in electronic and computer engineering from the University of Texas at San Antonio, San Antonio, TX, USA, and the Ph.D. degree in electrical and computer engineering from Tufts University, Medford, MA, USA.

He is working as a Graduate Research Assistant with the Visual and Sensing Lab, Tufts. His main areas of research interests include signal/image processing, deep learning, computer vision, 3-D scanning, and automated biometric technologies particularly focusing on fingerprints and their applications.

Shishir Paramathma Rao (Student Member, IEEE) received the B.E. degree in electronics and communication from Visvesvaraya Technological University, Belgaum, India, the M.S. degree in electrical and computer engineering from the University of Texas at San Antonio, San Antonio, TX, USA, and the Ph.D. degree in electrical and computer engineering from Tufts University, Medford, MA, USA.

His research interests include 3-D photography, image-based modeling, multiview stereovision, image and video analytics, machine-learning and neural networks, signal/image processing, and 3-D sensors.

Sos S. Agaian (Fellow, IEEE) received the M.S. degree in mathematics and mechanics (*summa cum laude*) from Yerevan State University, Yerevan, Armenia, the Ph.D. degree in mathematics and physics from the Steklov Institute of Mathematics, Russian Academy of Sciences (RAS), Moscow, Russia, and the Doctor of Engineering Sciences degree from the Institute of Control Systems, RAS.

He is currently a Distinguished Professor with The City University of New York/CSI, New York, NY, USA. He is also listed as a co-inventor on 44 patents/disclosures. The technologies that he invented have been adopted by multiple institutions, including the U.S. government, and commercialized by industry. His research interests include computational vision and machine learning, large scale data analytic analytics, multimodal data fusion, biologically inspired signal/image processing modeling, multimodal biometric and digital forensics, 3-D imaging sensors, information processing and security, and biomedical and health informatics. He has authored more than 650 technical articles and ten books in these areas.

Dr. Agaian received the Distinguished Research Award at the University of Texas at San Antonio. He received MAEStro Educator of the Year, sponsored by the Society of Mexican American Engineers. He was a recipient of the Innovator of the Year Award (2014), the Tech Flash Titans-Top Researcher-Award (San Antonio Business Journal, 2014), the Entrepreneurship Award (UTSA-2013 and 2016), and the Excellence in Teaching Award (2015). He is an Editorial Board Member for the *Pattern Recognition and Image Analysis* and an Associate Editor for several journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS, *Journal of Electrical and Computer Engineering* (Hindawi Publishing Corporation), *International Journal of Digital Multimedia Broadcasting* (Hindawi Publishing Corporation), and *Journal of Electronic Imaging* (SPIE, IS&T). He also serves as a Foreign Member of the Armenian National Academy. He is a Fellow of the SPIE, a Fellow of the IS&T, and a Fellow of the AAAS.