IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 4, 2018, accepted August 28, 2018, date of publication September 6, 2018, date of current version September 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2868733

Learning Based Image Transformation Using Convolutional Neural Networks

XIANXU HOU[1], YUANHAO GONG[1], BOZHI LIU[1], KE SUN[2], JINGXIN LIU[1], BOLEI XU[1], JIANG DUAN[3], AND GUOPING QIU[1,4]

[1]College of Information Engineering, Shenzhen University, Shenzhen, China
[2]Key Laboratory of Spatial Information Smarting Sensing and Services, Shenzhen University, Shenzhen, China
[3]School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, China
[4]School of Computer Science, University of Nottingham, Nottingham, U.K.

Corresponding author: Guoping Qiu (qiu@szu.edu.cn)

ABSTRACT We have developed a learning-based image transformation framework and successfully applied it to three common image transformation operations: downscaling, decolorization, and high dynamic range image tone mapping. We use a convolutional neural network (CNN) as a non-linear mapping function to transform an input image to a desired output. A separate CNN network trained for a very large image classification task is used as a feature extractor to construct the training loss function of the image transformation CNN. Unlike similar applications in the related literature such as image super-resolution, none of the problems addressed in this paper have a known ground truth or target. For each problem, we reason about a suitable learning objective function and develop an effective solution. This is the first work that uses deep learning to solve and unify these three common image processing tasks. We present experimental results to demonstrate the effectiveness of the new technique and its state-of-the-art performances.

INDEX TERMS Deep learning, image downscaling, image decolorization, HDR image tone mapping.

I. INTRODUCTION

Many classic image processing tasks (Fig. 1) can be framed as image transformation, where an input image is transformed to an output image based on a given criterion. In this paper, we consider three image transformation tasks: image downscaling, image decolorization (color to grayscale conversion), and high dynamic range (HDR) image tone mapping. Downscaling image operations are widely used today to allow users to view a reduced resolution image that preserves perceptually important details of its original megapixel version. Decolorization aims to convert a color image to a grayscale image which will preserve the visual contrasts of the original color version. Another image processing task is HDR image tone mapping. HDR images contain a much higher bit depth than standard image formats and can represent a dynamic range closer to that of human vision. The goal of HDR tone mapping is trying to faithfully reproduce the appearance of the high dynamic range image in display devices with limited displayable range.

These three seemingly disparate image processing tasks have a similar objective of outputting a reduced information version which will maximally convey important perceptual details of the original image. All these tasks face similar technical challenges. There are no known targets, an image can be transformed to an arbitrary number of plausible outputs. The transformation criterion, e.g., to preserve the perceptual details and contrasts of the original image, etc., are qualitative and subjective. There is no well-defined mathematical objective function to describe the transformation criterion and this makes it difficult to find a canonical computational solution to these problems.

In this paper, we take advantage of recent developments in deep convolutional neural networks (CNNs) and have developed a deep feature consistent deep image transformation (DFC-DIT) framework in which we train a deep CNN to transform an input image to an output image by keeping the deep features of the input and output consistent through another pre-trained (and fixed) deep CNN. We show that common traditional image processing tasks such as downscaling, decolorization and HDR tone mapping can be unified under the DFC-DIT framework and produce state-of-the-art results. To the best knowledge of the authors, this is the first work that successfully uses deep learning to solve downscaling, decolorization and HDR tone mapping problems in a unified framework.

The new DFC-DIT framework is built on two crucial insights, one into the visual appearance of an image and the other into the properties of the deep convolutional

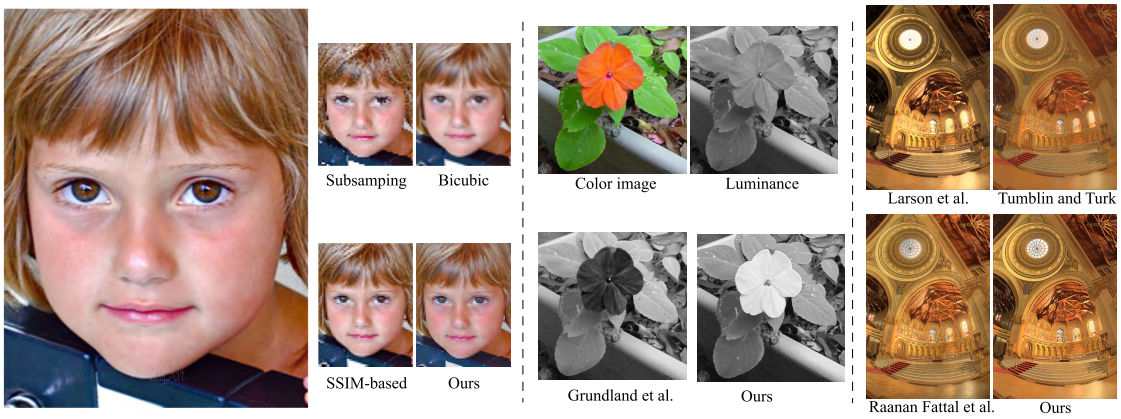


FIGURE 1. Examples of classic image transformation tasks. Image downscaling (left) where we show results of our method, two traditional methods (subsampling and bicubic) and a state-of-the-art SSIM-based method [1]. Decolorization (middle) where we show results of our method, the Luminance channel and a state-of-the-art method [2]. HDR image tone mapping (right) where we show results of our method and 3 methods from the literature [3]–[5].

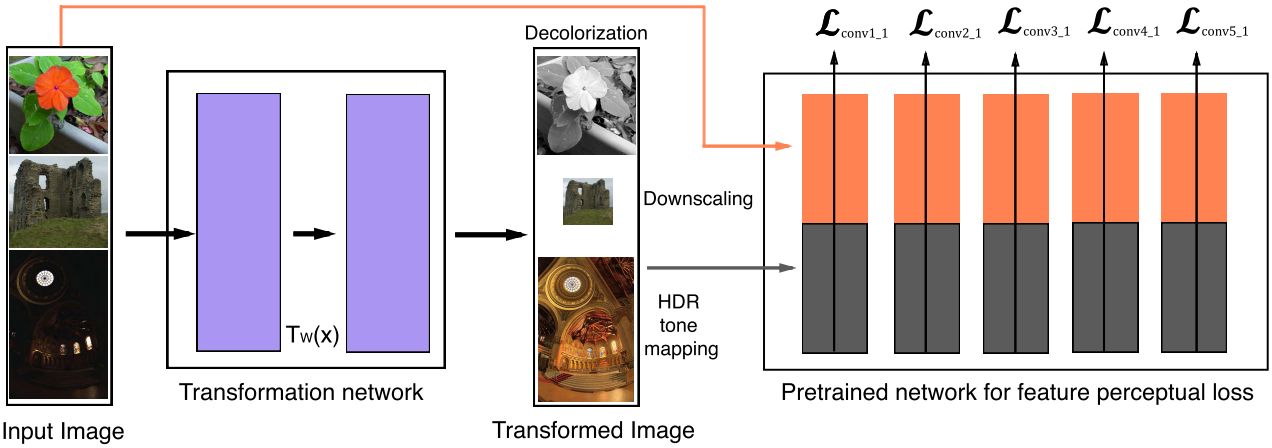


FIGURE 2. The deep feature consistent deep image transformation (DFC-DIT) framework. A convolutional neural network transforms an input to an output. A pretrained deep CNN is used to compute feature perceptual loss for the training of the transformation network.

neural networks. From image quality measurement literature, it is known that the change of spatial correlation is a major factor affecting the visual integrity of an image [6]. Research in deep learning has shown that the hidden layers of a convolutional neural network can capture a variety of spatial correlation properties of the input image [7]. As one of the most important objectives of many image processing tasks such as the three studied in this paper is to maximally preserve the visual integrity of the input, it is therefore crucial to keep the spatial correlations of the output consistent with those of the input. As the deep features (i.e., hidden layers' outputs) of a CNN capture the spatial correlations of the input image, we can therefore employ a (pre-trained and fixed) CNN and use its deep features to measure the spatial correlations of an image. Therefore, the goal of preserving the visual integrity of the input is equivalent to keeping the spatial correlations of the output consistent with that of the input, which in turn is equivalent to keeping their deep features consistent. Based on these key insights, we have successfully developed the DFC-DIT image processing framework (see Fig. 2).

The rest of the paper is organized as follows. We briefly review related literature in section II. Section III presents the DFC-DIT framework and its application to three important image processing tasks, i.e., spatial downscaling, decolorization and high dynamic range image tone mapping. Section IV presents experimental results which show that DFC-DIT stands out as a state-of-the-art technique. Finally we present a discussion to conclude the paper.

II. RELATED WORK

A. IMAGE DOWNSCALING

Classical image downscaling techniques usually involve processing the input images by applying a spatially constant low-pass filter, subsampling, and reconstructing the result to prevent aliasing in the reconstructed signal. Approximations to the theoretically optimum sinc filter such

as the Lanczos filter, and other filters (e.g., bilinear and bicubic) have been developed and used in practice. However the filtering kernels of these methods do not adapt to the image content. A recent content-adaptive technique [8] is proposed to overcome the above shortcoming by adapting the shape and location of every kernel to the local image content and demonstrates better downscaling results. A better depiction of the input image was proposed [1] by formulating image downscaling as an optimization problem with the structural similarity (SSIM) [9] as the perceptual image quality metric. In addition convolutional filters [10] are used to preserve visually important details in downscaled images.

B. DECOLORIZATION

Decolorization aims to convert color images into grayscale images while preserving structures and contrasts as much as possible. The baseline method is to extract the luminance channel of a given color image from the RGB channels. However it could fail to express salient structures of the color image because of the fixed weights to combine RGB channels. Other more advanced techniques are proposed to obtain better results by either focusing on local contrasts or global contrasts. Local contrasts [11], [12] use different mapping functions in different local regions of the image, while global contrasts [13]–[16] are designed to produce one mapping function for the whole image. Reference [17] takes into account multi-scale contrast preservation in both spatial and range domain and uses bilateral filtering to mimic human contrast perception. Reference [18] used a bimodal objective function to alleviate the restrictive order constraint for color mapping. Image fusion based strategy [19] is proposed for image and video decolorization. In addition, color-to-gray structural similarity (C2G-SSIM) index [20] is designed to quantitatively evaluate the luminance, contrasts and structure similarities between the reference color image and the corresponding grayscale image.

C. HDR IMAGE TONE MAPPING

HDR image tone mapping aims to reproduce high dynamic range radiance maps in low dynamic range reproduction devices. Tone mapping operators can be classified as global operators and local operators. Global operators [21]–[23] usually employ the same mapping function for all pixels and can preserve the intensity orders of the original scenes to avoid ''halo'' artifacts, however the global operators will generally cause loss of details in the mapped image. In contrast, local operators [4], [24], [25] use mapping functions which vary spatially across the image. Most local operators employ a pipeline to decompose an image into different layers or scales and then recompose the mapped results from various scales after contrast reduction. However, the major shortcoming of local operators is the presence of haloing artifacts. In addition, global operator is used in the local regions to reproduce local contrast and ensure better quality [26]. In order to quantitatively evaluate HDR tone mapping algorithm or multi-exposure image fusion, several objective quality assessment algorithms [27]–[29] are proposed recently. What's more, an up-to-date, detailed guide on the theory and practice of high dynamic range imaging is included in the book [30], which also provides MATLAB code for common tone mapping operators and TMQI index. In this paper, we use their code to reproduce previous methods.

D. IMAGE QUALITY METRICS

The choice of image quality metric is essential for image transformation tasks. Standard pixel-by-pixel measurement like mean square error is problematic and the resultant images are often of low quality. This is because the measurement is poorly correlated with human perception and can not capture the perceptual difference and spatial correlation between two images. Better metrics have been proposed for image quality assessment in recent years. Structural similarity (SSIM) index [9] is one of the most popular metrics, which computes a matching score between two images by local luminance, contrast, and structure comparisons. It has been successfully used for image downscaling [1] and super-resolution [31]. As mentioned in the previous sections C2G-SSIM index [20] and TMQI index [27] are commonly used objective quality assessment for decolorization and HDR image tone mapping algorithm.

E. RELEVANT DEEP LEARNING/CNN LITERATURE

Recently, there has been an explosion of publications on deep learning/CNN, we here briefly review the most closely related publications to our current work. A number of papers have successfully generated high-quality images based on the high-level features extracted from pretrained deep convolutional neural networks. By optimizing individual deep features [7], [32]–[34], better visual quality images can be generated, which in turn can help understand the learned representations of deep networks. Additionally [35] have achieved style transfer by minimizing content and style reconstruction loss which are also based on features extracted from deep networks. Other works try to train a feed-forward network for real-time style transfer and super-resolution [36]. Different loss functions are compared for image restoration with neural networks [37]. In addition image-to-image translation framework [38] are proposed to generate high quality images based on adversarial training.

It is worth noting that the downscaling problem studied in this paper has the opposite goal to super-resolution. Deep CNN based super-resolution training data has a unique corresponding target for a given input image. The downscaling operation, however, there is no known target in the training data for a given input. Therefore, existing end to end super-resolution learning [36], [39], [40] and other similar CNN based image processing techniques such as colorization [41], [42] cannot be directly applied to the problems studied in this paper.

III. METHOD

We seek to train a simple convolutional neural network as a non-linear mapper to transform an input image to an output

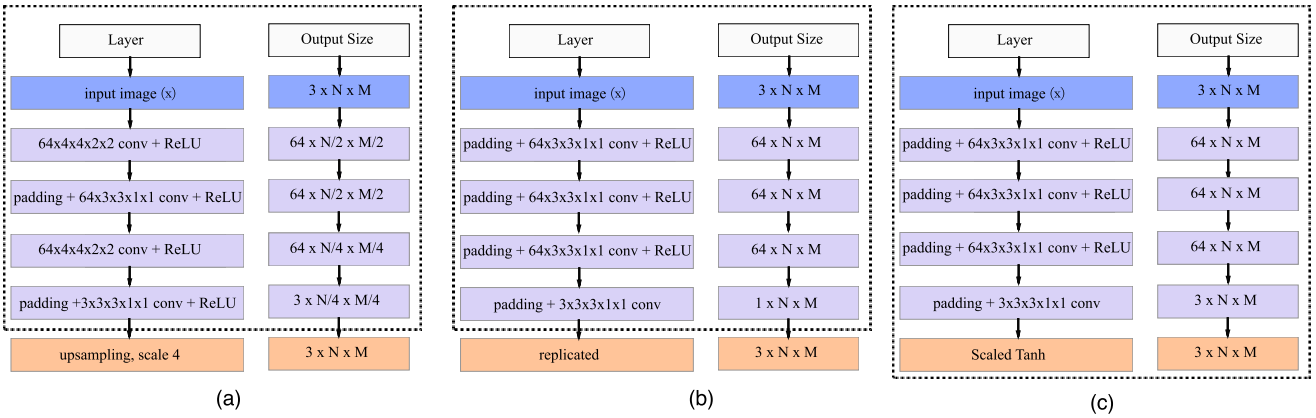


FIGURE 3. Transformation neural network architecture for image downscaling, decolorization and HDR image tone mapping. (a) Image downscaling. (b) Image decolorization. (c) HDR image tone mapping.

image following what we call the deep feature consistent principle. The schematic is illustrated in Fig. 2. Our system consists of two components: a transformation network $T_W(x)$ and a loss network $\Phi(x)$. The transformation network is a convolutional neural network parameterized by weights W, which transforms an input image x to an output image $\hat{x}$, i.e. $\hat{x} = T_W(x)$. The other component is the loss network Φ which is a pretrained deep convolutional neural network to help define the feature perceptual loss function for training $T_W(x)$. We feed both the original image x and the transformed image $\hat{x}$ to Φ and compute the feature perceptual loss $\mathcal{L}(x, \hat{x})$. Training $T_W(x)$ is to find the weights W that minimize $\mathcal{L}(x, \hat{x})$, i.e.

$$W^* = \arg\min_W E_x[\mathcal{L}(x, T_W(x))] \tag{1}$$

Equation (1) can be seen as an extension of the concept of perceptual loss, e.g. [36] and others. However, the three new applications we consider here are very different from those studied by others. These extensions are non-trivial and non-obvious; each requires in-depth understanding of the problem and ingenuity that cannot be readily derived from existing works. Unlike previous applications, none of our problems has a known ground truth or target for a supervised learning network. Instead, we have to reason about the suitable target and develop solutions to construct the perceptual loss for each application accordingly. In downscaling, we created a perceptual loss to match two images with different shapes (sizes). In decolorization, we constructed a perceptual loss to match two images with different number of color channels. In HDR tone mapping, we introduced a perceptual loss to match two images with different dynamic ranges.

A. DEEP FEATURE BASED FEATURE PERCEPTUAL LOSS

As alluded to earlier, the spatial correlation of an image is a major determining factor of the visual integrity of an image. The goal of image transformation in Fig. 2 and the tasks in Fig. 1 is to ensure $\hat{x}$ preserves the visual integrity of x. This can be alternatively stated as making the spatial correlations in $\hat{x}$ consistent with those in x. Instead of using handcrafted functions to describe an image's spatial correlations, we make use of a pretrained deep CNN. The hidden layers outputs, which we call deep features, capture the spatial correlations of the input image.

Specifically, let $\Phi_i(x)$ represent the i^{th} hidden activations when feeding the image x to Φ. If the i^{th} is a convolutional or ReLU layer, $\Phi_i(x)$ is a feature map of shape $[C_i, W_i, H_i]$, where C_i is the number of filter for the i^{th} convolutional layer, H_i and W_i are the height and width of the given feature map respectively. The feature perceptual loss $\mathcal{L}_i(x, \hat{x})$ for a given layer of two images x and $\hat{x} = T_W(x)$ is defined as the normalized Euclidean distance between the corresponding 3D feature maps. The final loss $\mathcal{L}_i(x, T_W(x))$ is the total loss of different layers as follows.

$$\mathcal{L}_i(x, T_W(x)) = \frac{1}{C_i W_i H_i} \sum_{c=1}^{C_i} \sum_{w=1}^{W_i} \sum_{h=1}^{H_i} (\Phi_i(x)_{c,w,h} - \Phi_i(T_W(x))_{c,w,h})^2 \tag{2}$$

$$\mathcal{L}(x, T_W(x)) = \sum_i \mathcal{L}_i(x, T_W(x)) \tag{3}$$

It is worth noting that Φ is pre-trained and fixed during the training of $T_W(x)$, it is used as convolutional filters to capture the spatial correlations of the images.

B. TRANSFORMATION NETWORKS ARCHITECTURE

The transformation networks are convolutional neural networks based on the architecture guidelines from VGGNet [43] and DCGAN [44], and the details of the architecture vary with different image transformation tasks (Fig. 3).

1) IMAGE DOWNSCALING

For image downscaling we use strided convolutions to construct the networks with 4×4 kernels. The stride is fixed to be 2×2 to achieve in-network downsampling instead of deterministic spatial functions such as max pooling and average pooling. The ReLU layer is used after the first convolutional

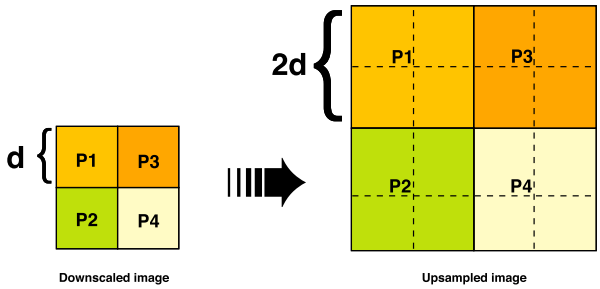


FIGURE 4. Nearest neighbor upsampling for the transformed image. The upsampled image contains the same amount of information as the downscaled image and is the same size as the original input image.

layer as non-linear activation function. Thus after two strided convolutions, the size of the input image can be downscaled to 1/4. In order to compute the feature perceptual loss we need to make sure that the transformed image and the original image have the same shape by a predefined upsampling method. In our experiments we apply a 2D upsampling of a factor of 4 over every channel of the transformed output (see Fig. 4), thus upscaling the downscaled image back to the same size as the original input. The simplest nearest neighbor upsampler is chosen to ensure the upsampled image has the same information as the downscaled image. Thus we can feed the upscaled version and the original image into the loss network to compute the feature perceptual loss.

2) IMAGE DECOLORIZATION

The image decolorization transformation only affects the color of the input images, and there is no need to incorporate downsampling architecture in the network. We use 3×3 kernels with 1×1 stride for all the convolutions. In addition, each feature map of a given input is padded by 1 pixel with the replication of the input boundary before the convolution operation. Thus the convolutional layers do not change the size of the input image. Like the image downscaling network we use ReLU layer after the first convolutional layer, but only a single filter for the last convolution to represent the transformed grayscale image. What we desired is the deep feature consistency of the decolorized output and the original image. We replicate the single channel of the decolorized output to a 3 channel color image (3 channels are identical), which is then fed to the loss network $\Phi(x)$ to calculate the feature perceptual loss with the original input. This is designed to ensure the replicated 3 channel color image have the same amount of information as the decolorized output.

3) HDR IMAGE TONE MAPPING

The network architecture for HDR image tone mapping is similar to the one used in image decolorization above. We use replication to pad the input boundary, and all the convolutions are 3×3 kernels with 1×1 stride. The difference is that 3 filters are needed for the last convolutional layer for reproducing a color image. The output layer is a scaled Tanh layer, restricting the pixel value of the transformed image to the displayable range [0, 255] from a high dynamic range. During the training we seek the deep feature consistency of the tone mapped and the original high dynamic range image. Specific implementation details of each of the applications are described in the experiments section.

IV. EXPERIMENTS

We present experimental results on three image transformation tasks: image downscaling, image decolorization and HDR image tone mapping to demonstrate the effectiveness of our method. We also investigate how the feature perceptual loss constructed with different hidden layers of the loss network affects the performances.

A. TRAINING DETAILS

Our image downscaling and decolorization transformation CNNs are trained offline using Microsoft COCO dataset released in 2014 [45], which is a large-scale dataset containing 82,783 training images. We resize all the image to 256×256 as the final training data, and train our models with a batch size of 16 for 10 epochs over all the training images. Once the transformation CNN is trained, it can be used to perform downscaling or decolorization.

For HDR image tone mapping, the transformation CNN is trained online, i.e., an HDR image is compressed using the transformation CNN trained with its own data. The practical consideration is that it is difficult to collect large enough training dataset. With large enough collection of training data, the model can also be trained offline.

For training, Adam [46] method is used for stochastic optimization with a learning rate of 0.0002. A pretrained 19-layer VGGNet [43] is used as loss CNN Φ to compute feature perceptual loss which is fixed during the training of the transformation CNN. When constructing the feature perceptual loss for a pretrained network, the first step is to decide which layer (layers) should be used. Unlike image generation works [7], [36] using ReLU layers, we use convolution layers for feature extraction. This is because the ReLU activation is just the corresponding convolutions output thresholded at 0, the convolutions could contain more subtle information when compared with ReLU output. Specifically we experiment feature perceptual loss by using convolutional layer conv1_1, conv2_1, conv3_1, conv4_1, conv5_1 and conv123_1 for comparison. The conv$i$_1 ($i = 1, 2, 3, 4, 5$) and conv123_1 represent the five convolutional layers of VGGNet and the combination of the first 3 layers respectively. Our implementation is built on open source machine learning framework Torch [47].

B. IMAGE DOWNSCALING

Image downscaling is trying to transform a high-resolution input image to a low-resolution output image. In our experiments we focus on the $\times$ 1/4 image downscaling similar to previous works [1], [8]. This seemingly simple routine image operation is actually a technically challenging task because it is very difficult to define the correct low-resolution image. Based on our DFC-DIT framework, we ensure that the

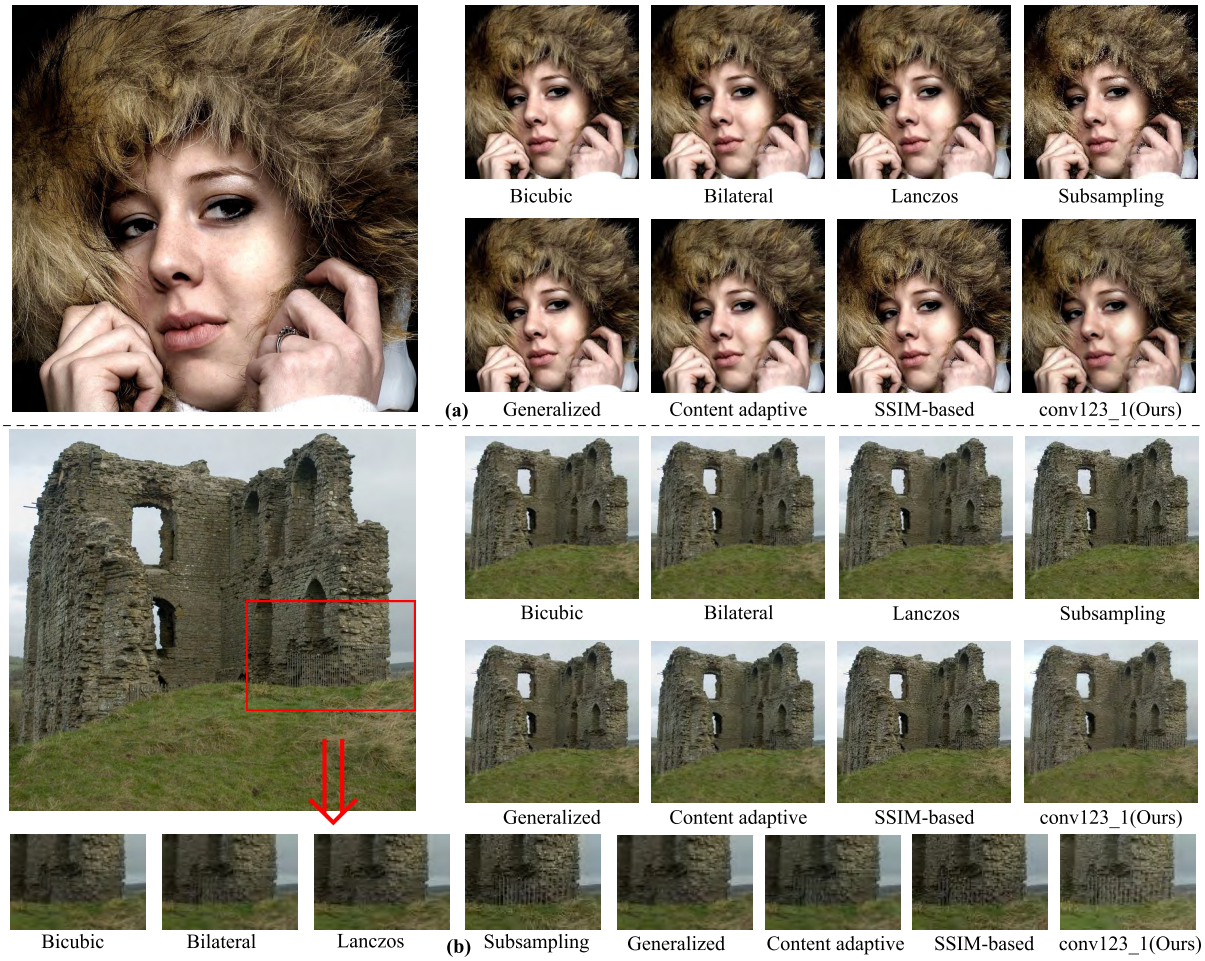


FIGURE 5. A comparison of natural images downscaled by different methods. The results are downscaled by a factor of × 1/4 while the original inputs are resized for better display. For each image, results of common filters such as Bicubic, Bilateral, Lanczos and Subsampling are shown in the first row. Results of recent methods, generalized sampling [48], content-adaptive [8] and SSIM-based downsampling [1] and ours are shown in the second row. Our conv123_1 results are produced by a model trained with a combined loss of convolutional layers conv1_1, conv2_1, and conv3_1. The bottom row of the second image shows a local region of the downscaled image by different methods. All the images are courtesy of [1]. The results are best viewed in native resolution electronically.

downsampled image and the original image will have similar deep features which means that the output will maintain the spatial correlations of the original image thus keeping the visual integrity of the original image.

1) QUALITATIVE RESULTS

Although our network is trained on images of shape 256 × 256, it can be adapted to any image sizes because of its fully convolutional architecture. After training, we evaluate our method on the testing images from [1]. We first show the qualitative examples and compare our results with other state-of-the-art methods. We then evaluate how perceptual losses constructed at different convolutional layers affect the performances.

Fig. 5 shows qualitative examples of our results, other common techniques and state-of-the-art methods. We only show results of downscaling by a factor of × 1/4, the original images are resized for better display. We can see that bicubic filter is known to lead to oversmoothing results and cannot preserve fine structures such as the fence area highlighted by the red rectangle (Fig. 5(b)). Other filters such as bilateral filter and Lanczos filter achieve sharper downscaled results, however these filters are also problematic. Bilateral filter can lead to ringing artifacts (the hair in Fig. 5(a)), and Lanczos filter could not preserve small-scale features such as the fence area in Fig. 5(b). More advanced methods such as generalized sampling [48], and content-adaptive downscaling [8] and SSIM-based downscaling [1] could produce better results, but still cannot preserve all perceptually important details. In contrast our method trained by a feature perceptual loss constructed using layer conv1_1, conv2_1 and conv3_1 deep features can capture important fine textures and produce better transformed results, visually closer to the original high-resolution inputs. From Fig. 5(b), the fine textures of the fence area can be seen clearly in the downscaled image.

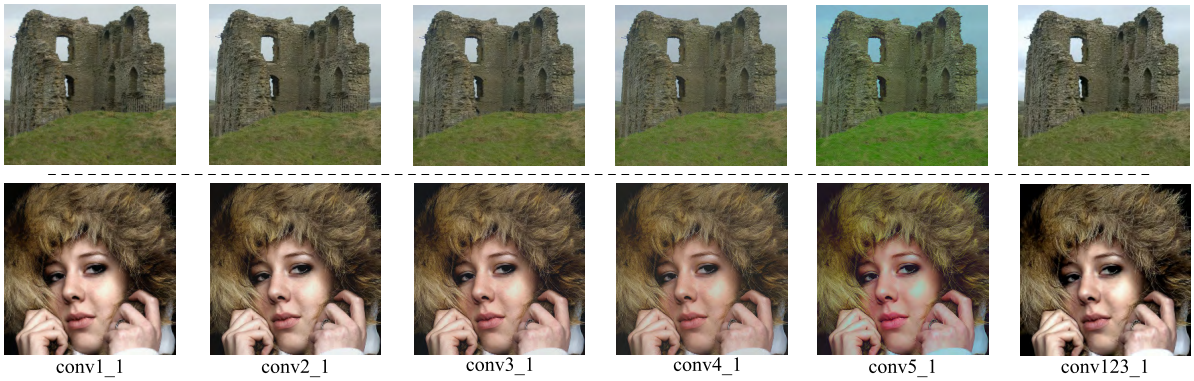


FIGURE 6. A comparison of natural images downscaled with the DFC-DIT framework with different levels of feature perceptual loss. The examples, from left to right, are × 1/4 downscaling results with perceptual losses computed with individual hidden layers of VGGNet (from layer 1 to layer 5). The last column is the results based on a perceptual loss combining the first 3 layers.

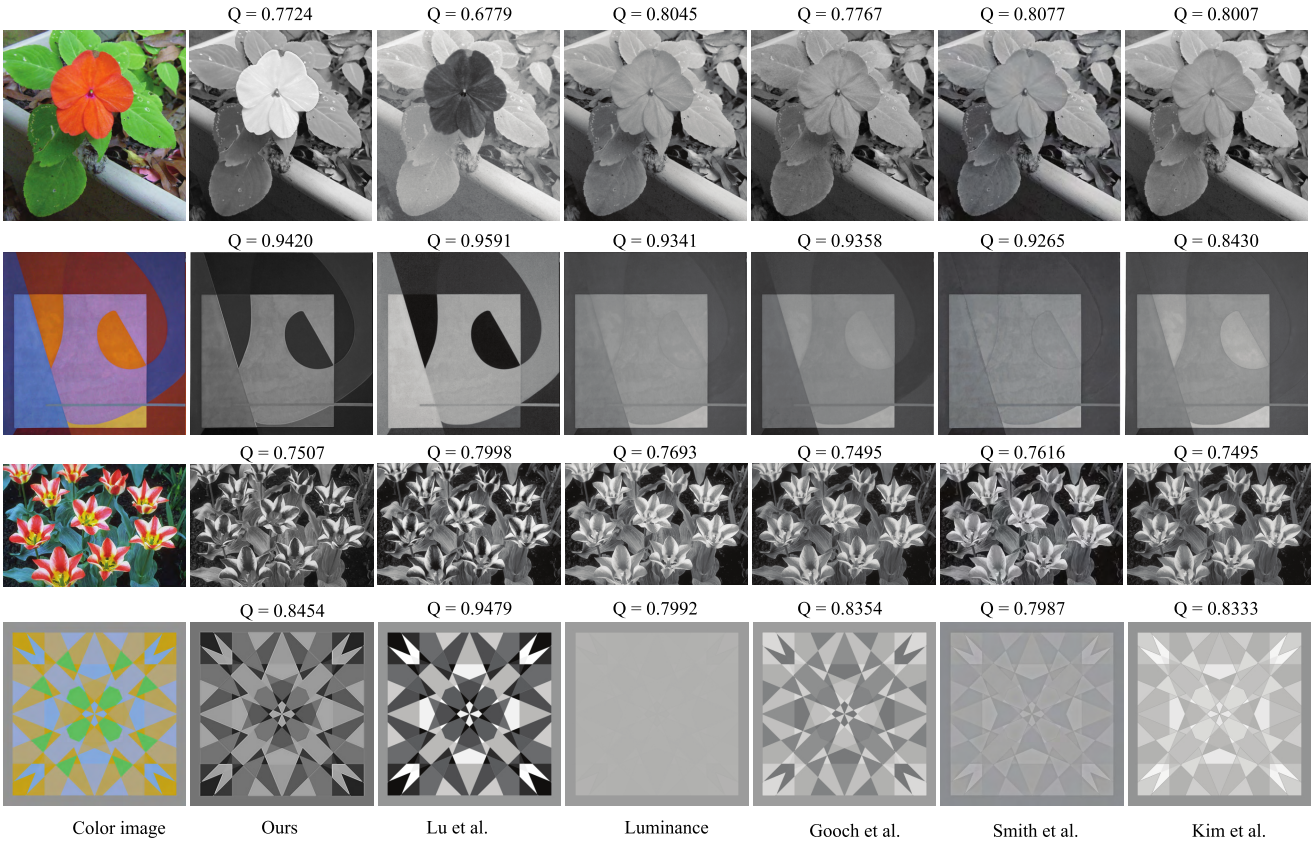


FIGURE 7. A comparison of decolorized images by different methods. We compare our method trained with conv4_1 layer with standard luminance and other recent methods [11], [12], [18], [49]. The C2G-SSIM index value (Q) [20] is also shown for each decolorized image. The results are best viewed electronically.

Although simple (nearest neighbor) subsampling can also achieve sharper images, the results are sometimes noisy and suffer from aliasing (see the hair in Fig. 5(a)). Our algorithm avoids both oversmoothing and aliasing problems and produces a crisp and noise-free image. These results demonstrate that by keeping the deep features of the downscaled image consistent with those of the original can indeed preserve the visual integrity of the input.

2) DEEP FEATURE CONSISTENCY AT DIFFERENT LAYERS

Fig. 6 shows results of DFC-DIT downscaled images using perceptual losses computed using conv1_1, conv2_1, conv3_1, conv4_1 and conv5_1 layer of the VGGNet individually. We find that keeping the deep feature consistent at these individual layers can in general preserve the original texture or content well. However for the high level layers, the downscaled images could lose detailed pixel information

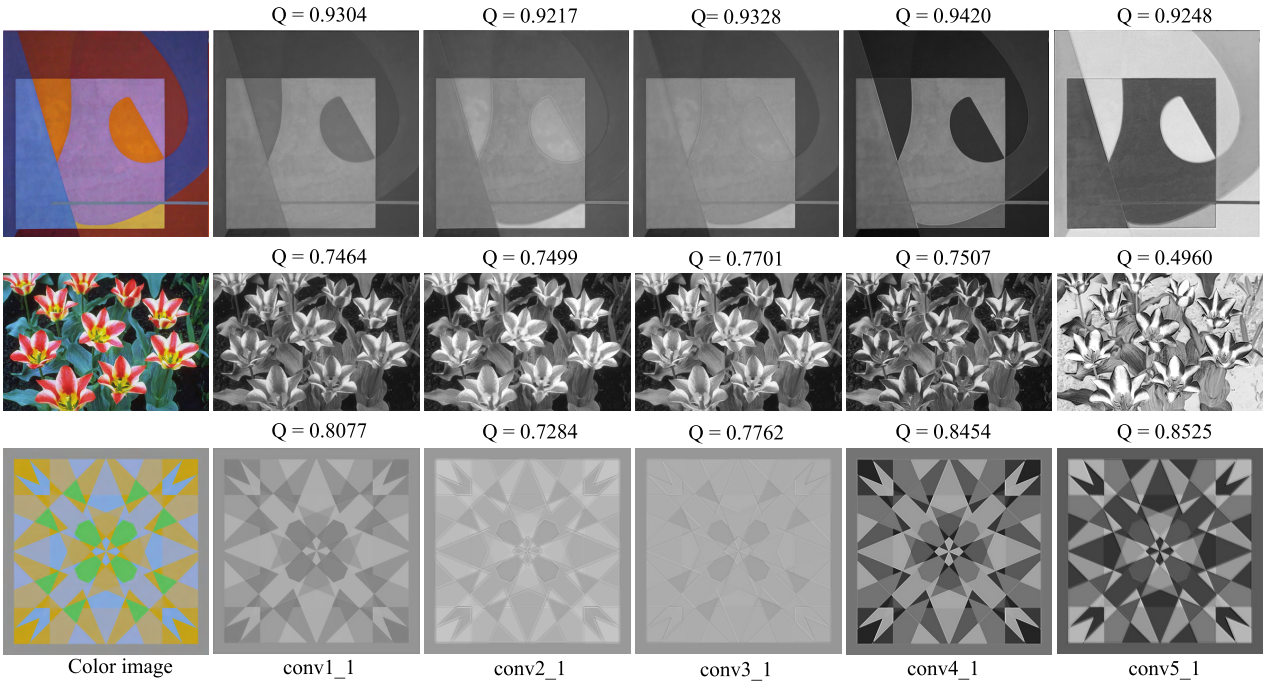


FIGURE 8. A comparison of decolorization by our methods trained with different level feature perceptual loss. The examples are trained from low level to high level layers in VGGNet. The C2G-SSIM index value (Q) [20] is also shown for each decolorized image. The results are best viewed electronically.

such as pixel color. For example, results of conv4_1 and conv5_1 in Fig. 6 have higher color contrasts. We also found that by combining the first three layer deep features in general works very well.

C. IMAGE DECOLORIZATION

Like image downscaling we also train a four-layer convolutional network to transform color images into grayscale images using the DFC-DIT framework. One of the major problems in traditional approach to this task is that in iso-luminant areas the color contrasts will disappear in the grayscale image because even though the pixels have different colors their luminance levels are the same. In our neural network based nonlinear mapping framework, we enforce deep feature consistency which means that the spatial correlations of the color images are preserved in the grayscale image. Thus even in iso-luminant regions, the color contrasts will be preserved as grayscale contrasts.

1) EXPERIMENTAL RESULTS

Again, our fully convolutional neural network architecture can be applied to process images of any sizes even though the training images have a fixed size. Fig. 7 shows several comparative results against standard luminance and recent color to grayscale methods [11], [12], [18], [49]. Our training-based approach can preserve the color contrasts of the original images, the grayscale images appear sharp and fine details are well protected. In addition, we adopt C2G-SSIM index [20] as the objective quality metric for quantitative evaluation. C2G-SSIM index is designed to evaluate the luminance, contrast and structure similarities between the reference color image and the decolorized image. The quality index values are shown in Fig. 7 for each decolorized image. Furthermore, a visual demonstration is shown in Fig. 9 where the brightness indicates the magnitude of the local C2G-SSIM index values. As it can be seen, the decolorized images converted by our algorithm and [18] can better preserve the contrast and structure in flat area, but stronger penalty (dark pixel in the map) is given at several color edges.

It is interesting to note that unlike previous methods, we did not explicitly compute color contrasts and grayscale contrasts, instead we only enforce deep feature consistency of the color and the decolorized images. From these examples, we have shown convincingly that our DFC-DIT framework is an effective decolorization method.

2) DEEP FEATURE CONSISTENCY AT DIFFERENT LAYERS

We also conduct experiments to evaluate how deep feature consistency at different hidden layers of the loss network affects the decolorization results. Converted grayscale images produced by models trained with perceptual loss of different hidden layers are shown in Fig. 8 and the C2G-SSIM index are calculated for each images. Again we can see that all the transformed images are able to reconstruct the content of the original color image and preserve the contrasts. Compared to lower layers, the decolorized images from higher layers generally have higher C2G-SSIM values and do a better job at reconstructing fine details, especially the

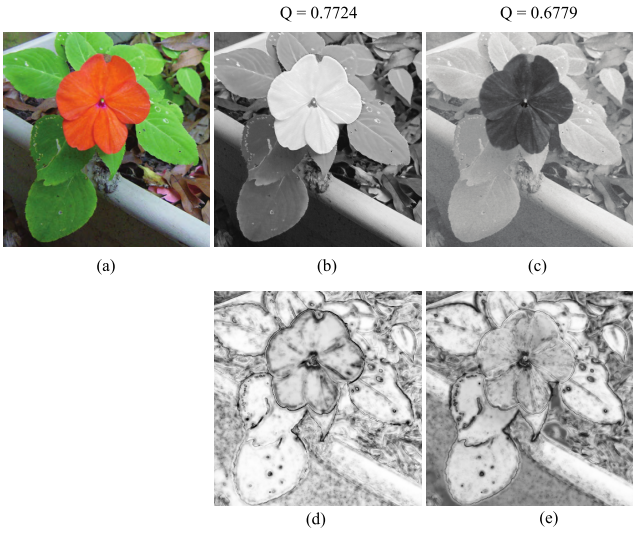


FIGURE 9. Decolorized images and their corresponding quality C2G-SSIM index [20] maps. (a) is the reference color image. (b) and (c) are the decolorized images produced by our algorithm and [lu2014contrast]. (d) and (e) are the corresponding C2G-SSIM index maps of (b) and (c) respectively. For C2G-SSIM index maps, brighter indicates better quality.

contrast preservation that is desired. Specifically the results from lowest layers, i.e., conv1_1 are similar to the luminance channel (Fig. 7), isoluminant regions are mapped onto the same output intensity and global appearance is not well preserved. Constructing feature perceptual loss from higher layer is better for contrast preserving. However when using the highest conv5_1 layer (Fig. 8), the contrast of the outputs is too high that makes the decolorized images look unnatural. Our best model is trained by using conv4_1 layer.

D. HDR IMAGE TONE MAPPING

Unlike image downscaling and decolorization where a single model is trained offline using a large collection of training images and used to process all testing images, we adapt one network to a single HDR image due to the lack of large HDR dataset available for training. This can be seen as an online process where we use an HDR image's own data to optimize its own transformation function. It is important to note that this approach is realistic in practice as the process only needs the HDR input to produce its tone mapped output and there is no need to use any other extra information. The only slight disadvantage is that it requires online training the neural network using an HDR image's own data before outputting the final tone mapped image. Comparing with training the model offline using a large collection of training images, this online approach will be slower because it needs to adapt the neural network to the current testing image before producing the output tone mapped image. In our implementation on a machine with an Intel Core i7-4790K CPU and a Nvidia Tesla K40 GPU, it takes around 20 seconds to tone map a 768×512 HDR image.

It is a common practice to process the HDR radiance map in the logarithmic domain, we feed the logarithm of the radiance signal to the transformation CNN. Dynamic range compression is achieved by a *Tanh* function in the last layer of the transformation network (Fig. 3(c)). In practice, the dynamic range of the input HDR radiance signal is compressed to the displayable range [0, 255]. Following the principle of DFC-DIT, the HDR tone mapping transformation network is optimized by enforcing deep feature consistency between the transformation output image and the original HDR radiance map.

1) RENDERING DISPLAY IMAGE

The output of the transformation network will have the correct dynamic range suitable for display, however, its color may not be correct due to the nonlinear mapping operations of the transformation CNN. We therefore need to render the output of the transformation network to have the correct color for display. As in other tone mapping method [23], the final tone mapped image is rendered as

$$R_{out} = \left(\frac{R_{in}}{L_{in}}\right)^{\gamma} L_{out} \tag{4}$$

$$G_{out} = \left(\frac{G_{in}}{L_{in}}\right)^{\gamma} L_{out} \tag{5}$$

$$B_{out} = \left(\frac{B_{in}}{L_{in}}\right)^{\gamma} L_{out} \tag{6}$$

where R_{out}, G_{out} and B_{out} are the final tone-mapped RGB channels, R_{in}, G_{in} and B_{in} are the original radiance values in the corresponding HDR channels, and γ can be used to render the correct display color. L_{in} and L_{out} are respectively the luminance value of the HDR radiance map and the luminance value of the transformation image by the transformation CNN. According to the literature, γ should be set between 0.4 and 0.6 and we set it to 0.5 in all our results.

2) EXPERIMENTAL RESULTS

Fig. 10 and Fig. 11 display examples of tone mapping results of some HDR radiance maps of real scenes that are widely used in the literature, i.e., "Stanford Memorial Church" and "Vine Sunset". We compare our results with some of the best known and latest methods in the literature including Larson *et al.* [3], Expoblend [50], Lischinski *et al.* [51], Reinhard *et al.* [25], gradient domain [5], fast bilateral filtering [24] and Kim and Kautz [52]. In addition, TMQI (tone-mapped image quality index) [27] is used for quantitative comparison, which is a widely-used objective quality assessment algorithm for tone-mapped images by combining a multi-scale signal fidelity measure on the basis of a modified SSIM [9] and a naturalness measure on the basis of intensity statistic of nature images. The corresponding TMQI index value is given for each tone mapped image in Fig. 10 and Fig. 11. We can see that our method is able to render the images with excellent visual appearances to keep tiny details and contrast of the radiance map and produce high TMQI index values, which are at least as good as those produced by the best methods.

FIGURE 10. Stanford Memorial Church displayed using different methods with corresponding TMQI index value [27]. We show those of Larson *et al.* [3], Expoblend [50], Lischinski *et al.* [51], Reinhard *et al.* [25], gradient domain [5], fast bilateral filtering [24] and Kim and Kautz [52]. Our results are based on feature perceptual loss of 3 layers conv1_1, conv2_1 and conv3_1.

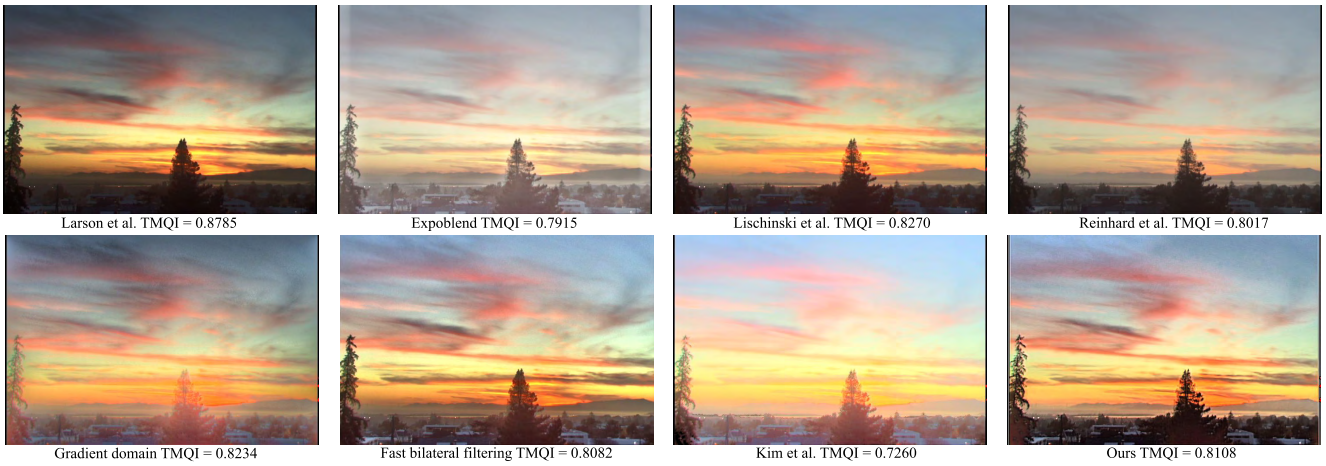


FIGURE 11. Sunset image displayed using different methods with corresponding TMQI index value [27]. We show those of [3], Expoblend [50], Lischinski *et al.* [51], Reinhard *et al.* [25], gradient domain [5], fast bilateral filtering [24] and Kim and Kautz [52]. Our results are based on feature perceptual loss of 3 layers conv1_1, conv2_1 and conv3_1. Our results are based on feature perceptual loss of 3 layers conv1_1, conv2_1 and conv3_1.

3) DEEP FEATURE CONSISTENCY AT DIFFERENT LAYERS

In Fig. 12 we show how feature perceptual loss from different hidden layers affect the tone mapped images of the DFC-DIT framework for HDR tone mapping. Overall the tone mapped images based on perceptual losses from the middle level (conv2_1 and conv3_1) have a good balance

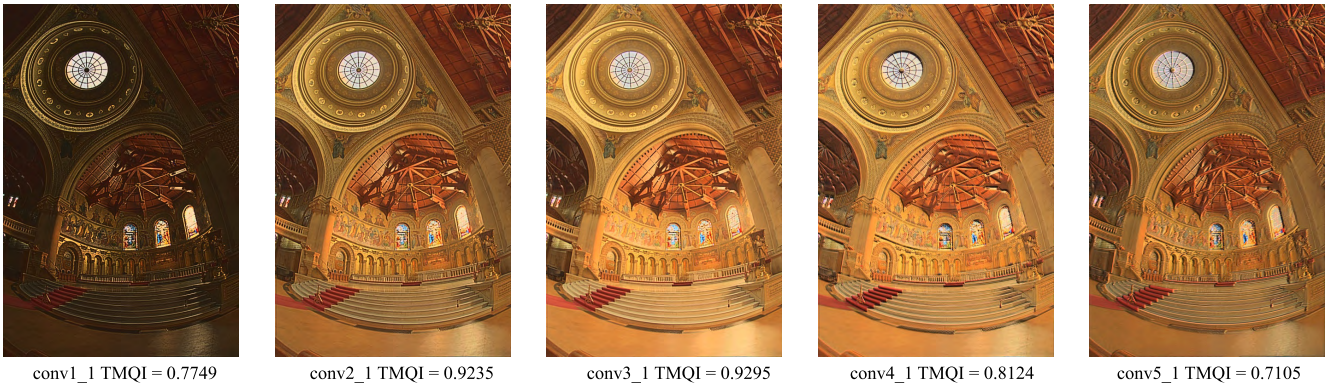


FIGURE 12. A comparison of HDR image tone mapping by our methods trained with different level feature perceptual loss. The corresponding TMQI index value [27] is given for each image. The results are best viewed electronically.

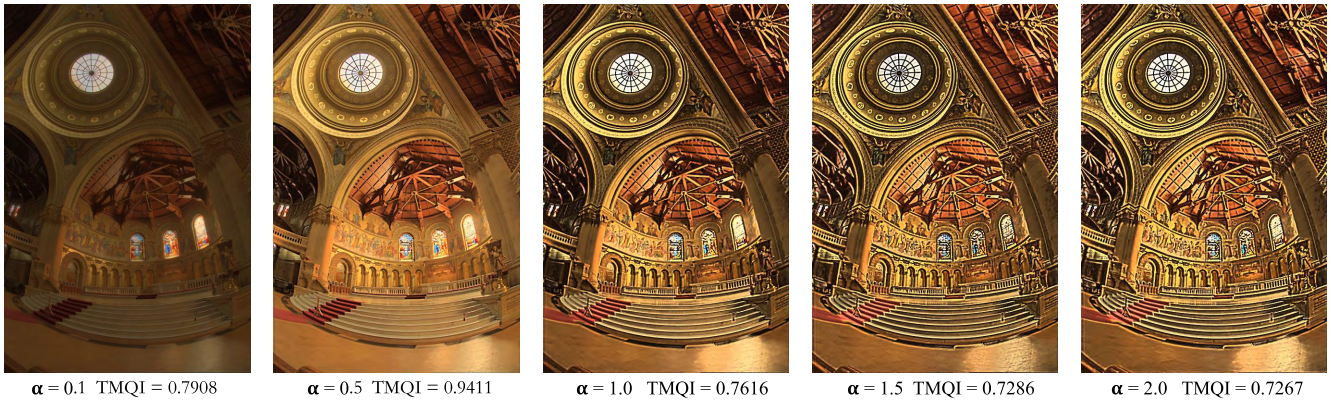


FIGURE 13. A demonstration of the effects of logarithmic compression based on feature perceptual loss of 3 layers conv1_1, conv2_1 and conv3_1. The corresponding TMQI index value [27] is given for each image.

between local and global contrasts and also produce highest TMQI index values. Combining the perceptual losses of first several layers together tend to produce somewhat better results than using a single layer. The tone mapped outputs based on higher layers (conv4_1 and conv5_1) have a relatively lower TMQI index values and appear slightly bumpy effect on different regions.

4) THE EFFECTS OF LOGARITHMIC COMPRESSION

As mentioned above, we first compress the HDR radiance map with the logarithmic functions and try to seek the deep feature consistency in the logarithmic domain. We can multiply the compressed radiance map with a factor α to control the logarithmic transformation. The tone mapping results and corresponding TMQI index values with different α are shown in Fig. 13. It can be seen that a higher α can lead to a more noticeable local contrast and crisp appearance of the tone mapped results. In our experiments, we find that a proper α is critical for our algorithm to produce high quality images. Specifically the tone-mapped results tend to be too dark and lack of contrast when α is too small, while a too big α could result in over-sharpen images and cause very unnatural and unattractive results, which can be also reflected by a relatively lower TMQI index values. This is because the compressed HDR radiance map with a higher α retains a higher dynamic range in logarithmic domain with more local details, as a result, global naturalness of the outputs could be sacrificed when the algorithm puts more effort on the local details. In our experiments, it works well when α is around 0.5 and our method can extract exquisite details from high-contrast images.

E. SUBJECTIVE EVALUATION OF DFC-DIT FRAMEWORK

We have conducted a subjective evaluation of results of downscaling, decolorization and HDR tone mapping of the new DFC-DIT framework. For each transformation, we evaluate our technique against several best techniques in the literature. For downscaling, we use bicubic, bilateral, lanczos, subsampling, generalized sampling [48], content-adaptive [8] and SSIM based method [1] as the benchmarks. For decolorization, we use luminance the methods of Smith *et al.* [12], Kim and Kautz [49], Gooch *et al.* [11] and Lu *et al.* [18] as benchmarks. For HDR tone mapping we use Larson *et al.* [3], fast bilateral filtering [24], gradient domain [5], Expoblend [50], Kim and Kautz [52], Lischinski *et al.* [51]

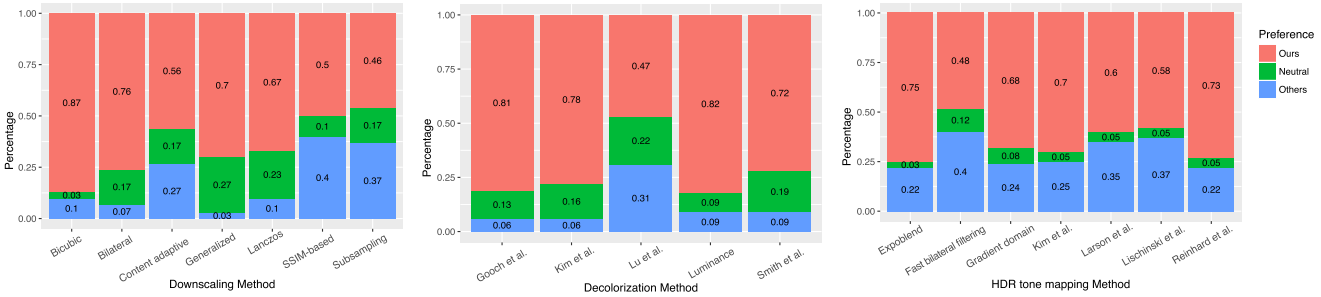


FIGURE 14. Subjective evaluation results. The red areas represent the percentage that our algorithm is selected, green areas for no preference and the blue ones for the other methods.

and Reinhard *et al.* [25] as benchmarks. For each image, we show the original input image (in the case of HDR tone mapping, the original radiance map cannot be shown), a version produced by our method and a version of one benchmark technique to subjects and ask which version they prefer or indicate no preference. 50 undergraduate science and engineering students from our university evaluated 10 pairs of images for image downscaling and 8 pairs of images for image decolorization and HDR tone mapping. Fig. 14 shows the voting results. We can see that there is an obvious preference for our method against all other methods for all the transformation tasks. These results demonstrate DFC-DIT framework is comparable to or better than state-of-the-art techniques. In image downscaling, subsampling and SSIM-based are two competing methods to produce sharp and crisp downscaled images, however subsampling sometimes suffer strong aliasing artifacts like the hair in Fig. 5. In image decolorization, the method of Lu *et al.* [18] is the best competing candidate that maximally preserves color contrast. However some participants prefer ours than theirs because the decolorized versions of Lu *et al.* [18] may show too strong contrast while the corresponding color images in fact have low contrasts. For HDR image tone mapping, fast bilateral filtering [24] is the best comparable tone mapping operator in our study.

V. CONCLUDING REMARKS

This paper has successfully introduced the DFC-DIT framework which unifies several common difficult image processing tasks. This is also the first time that deep learning has been successfully applied to image downscaling, decolorization and high dynamic range image tone mapping. Experimental results have demonstrated the effectiveness of the method and its state-of-the-art performances.

One fundamental problem for traditional image transformation tasks like image downscaling, image decolorization and HDR image tone mapping is that the problems are inherently ill-posed, because there is no unique correct ground truth. For image downscaling fine details should be preserved from visually ambiguous high-resolution inputs; for image decolorization the gray image should be semantically similar to the original color version and preserve the contrast as much as possible in spite of drastic loss of color information; for HDR image tone mapping we want to compress the scene radiance to displayable range while preserving details and color appearance to appreciate the original scene content. Therefore, success in these image transformation tasks requires semantic and perceptual reasoning about the input.

It is very difficult to design a numerical image quality metric to measure the perceptual similarity between the transformed outputs and the original input. Based on two crucial insights into the determining factors of visual quality of images and the properties of deep convolutional neural network, we have developed the deep feature consistent deep image transformation (DFC-DIT) framework which unifies common and ill-posed image processing tasks like downscaling, decolorization and HDR tone mapping. We have shown that the hidden layer outputs of a pretrained deep CNN can be used to compare perceptual similarities between the input and the output images. One possible explanation is that the hidden representations of a pre-trained CNN have captured essential visual quality details such as spatial correlation information and other higher level semantic information. Exactly which hidden layer represents what kind of essential visual quality details is not very clear and we have shown that perceptual losses constructed with different hidden layer features can affect the final results. Future researches are needed to understand better the kinds of visual semantics captured by the hidden features of the pre-trained CNN in the context of the DFC-DIT framework. A better way to combine (e.g. weighting) different level deep features may lead to better and more consistent results.

ACKNOWLEDGMENT

(Xianxu Hou and Yuanhao Gong contributed equally to this work.)

REFERENCES

[1] A. C. Öztireli and M. Gross, "Perceptually based downscaling of images," *ACM Trans. Graph.*, vol. 34, no. 4, 2015, Art. no. 77.
[2] M. Grundland and N. A. Dodgson, "Decolorize: Fast, contrast enhancing, color to grayscale conversion," *Pattern Recognit.*, vol. 40, no. 11, pp. 2891–2896, 2007.
[3] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," *IEEE Trans. Vis. Comput. Graphics*, vol. 3, no. 4, pp. 291–306, Oct./Dec. 1997.

[4] J. Tumblin and G. Turk, "LCIS: A boundary hierarchy for detail-preserving contrast reduction," in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn.* Reading, MA, USA: Addison-Wesley, 1999, pp. 83–90.

[5] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 249–256, Jul. 2002.

[6] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[7] X. Hou, L. Shen, K. Sun, and G. Qiu. (2016). "Deep feature consistent variational autoencoder." [Online]. Available: https://arxiv.org/abs/1610.00291

[8] J. Kopf, A. Shamir, and P. Peers, "Content-adaptive image downscaling," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 173:1–173:8, Nov. 2013.

[9] Z. Wang, 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.

[10] N. Weber, M. Waechter, S. C. Amend, S. Guthe, and M. Goesele, "Rapid, detail-preserving image downscaling," *ACM Trans. Graph.*, vol. 35, no. 6, Nov. 2016, Art. no. 205, doi: 10.1145/2980179.2980239.

[11] A. A. Gooch, S. C. Olsen, J. Tumblin, and B. Gooch, "Color2Gray: Salience-preserving color removal," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 634–639, 2005.

[12] K. Smith, P.-E. Landes, J. Thollot, and K. Myszkowski, "Apparent greyscale: A simple and fast conversion to perceptually accurate images and video," *Comput. Graph. Forum*, vol. 27, no. 2, pp. 193–200, 2008.

[13] K. Rasche, R. Geist, and J. Westall, "Re-coloring images for gamuts of lower dimension," *Comput. Graph. Forum*, vol. 24, no. 3, pp. 423–432, 2005.

[14] C. O. Ancuti, C. Ancuti, and P. Bekaert, "Enhancing by saliency-guided decolorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 257–264.

[15] C. Lu, L. Xu, and J. Jia, "Contrast preserving decolorization," in *Proc. IEEE Int. Conf. Comput. Photogr. (ICCP)*, Apr. 2012, pp. 1–7.

[16] M. Qiu, G. D. Finlayson, and G. Qiu, "Contrast maximizing and brightness preserving color to grayscale image conversion," in *Proc. Conf. Colour Graph., Imag., Vis.*, 2008, pp. 347–351.

[17] Y. Song, L. Bao, X. Xu, and Q. Yang, "Decolorization: is rgb2gray() out?" in *Proc. SIGGRAPH Asia Tech. Briefs*, 2013, Art. no. 15.

[18] C. Lu, L. Xu, and J. Jia, "Contrast preserving decolorization with perception-based quality metrics," *Int. J. Comput. Vis.*, vol. 110, no. 2, pp. 222–239, 2014.

[19] C. O. Ancuti, C. Ancuti, C. Hermans, and P. Bekaert, "Image and video decolorization by fusion," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 79–92.

[20] K. Ma, T. Zhao, K. Zeng, and Z. Wang, "Objective quality assessment for color-to-gray image conversion," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4673–4685, Dec. 2015.

[21] J. Tumblin and H. Rushmeier, "Tone reproduction for realistic images," *IEEE Comput. Graph. Appl.*, vol. 13, no. 6, pp. 42–48, Nov. 1993.

[22] G. Ward, "A contrast-based scalefactor for luminance display," in *Graphics Gems IV*. 1994, pp. 415–421.

[23] G. Qiu, J. Duan, and G. D. Finlayson, "Learning to display high dynamic range images," *Pattern Recognit.*, vol. 40, no. 10, pp. 2641–2655, 2007.

[24] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257–266, 2002.

[25] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, Jul. 2002.

[26] J. Duan, M. Bressan, C. Dance, and G. Qiu, "Tone-mapping high dynamic range images by novel histogram adjustment," *Pattern Recognit.*, vol. 43, no. 5, pp. 1847–1862, 2010.

[27] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 657–667, Feb. 2013.

[28] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.

[29] K. Ma, H. Yeganeh, K. Zeng, and Z. Wang, "High dynamic range image compression by optimizing tone mapped image quality index," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3086–3097, Oct. 2015.

[30] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging: Theory and Practice*. Natick, MA, USA: AK Peters, 2011.

[31] F. Zhou and Q. Liao, "Single-frame image super-resolution inspired by perceptual criteria," *IET Image Process.*, vol. 9, no. 1, pp. 1–11, 2015.

[32] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. (2015). "Understanding neural networks through deep visualization." [Online]. Available: https://arxiv.org/abs/1506.06579

[33] C. Szegedy *et al.* (2013). "Intriguing properties of neural networks." [Online]. Available: https://arxiv.org/abs/1312.6199

[34] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 427–436.

[35] L. A. Gatys, A. S. Ecker, and M. Bethge. (2015). "A neural algorithm of artistic style." [Online]. Available: https://arxiv.org/abs/1508.06576

[36] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 694–711.

[37] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. (2016). "Image-to-image translation with conditional adversarial networks." [Online]. Available: https://arxiv.org/abs/1611.07004

[39] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 184–199.

[40] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[41] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 649–666.

[42] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 577–593.

[43] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[44] A. Radford, L. Metz, and S. Chintala. (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1511.06434

[45] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.

[46] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980

[47] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *Proc. BigLearn NIPS Workshop*, 2011, Paper EPFL-CONF-192376.

[48] D. Nehab and H. Hoppe, "Generalized sampling in computer graphics," Tech. Rep., Feb. 2011.

[49] Y. Kim, C. Jang, J. Demouth, and S. Lee, "Robust color-to-gray via nonlinear global mapping," *ACM Trans. Graph.*, vol. 28, no. 5, 2009, Art. no. 161.

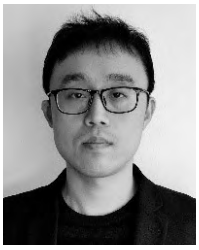
[50] N. D. Bruce, "Expoblend: Information preserving exposure blending based on normalized log-domain entropy," *Comput. Graph.*, vol. 39, pp. 12–23, Apr. 2014.

[51] D. Lischinski, Z. Farbman, M. Uyttendaele, and R. Szeliski, "Interactive local adjustment of tonal values," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 646–653, 2006.

[52] M. H. Kim and J. Kautz, "Consistent tone reproduction," in *Proc. 10th IASTED Int. Conf. Comput. Graph. Imag. (CGIM)*. Anaheim, CA, USA: ACTA Press, 2008, pp. 152–159. [Online]. Available: http://dl.acm.org/citation.cfm?id=1722302.1722332

XIANXU HOU received the B.S. and M.S. degrees from the China University of Mining and Technology, Beijing, in 2011 and 2014, respectively, and the Ph.D. degree in computer science from the University of Nottingham in 2018. He currently holds a post-doctoral position with the College of Information Engineering, Shenzhen University, China. His research interests include deep learning, computer vision, and natural language processing.

YUANHAO GONG received the B.Sc. degree in mathematics from Tsinghua University, Beijing, China, in 2007, and the Ph.D. degree in computer science from ETH Zürich, Switzerland, in 2015. He held a post-doctoral position at the Computer Vision Laboratory, ETH Zürich. Since 2018, he has been an Assistant Professor with the College of Information Engineering, Shenzhen University, China. He received several awards, including the Best Paper Award at the IEEE International Symposium on Biomedical Imaging 2012.



BOZHI LIU received the B.S. degree from the Nanjing University of Science and Technology, and the master's and Ph.D. degrees from the University of Nottingham, U.K. He is currently a Research Fellow with the College of Information Engineering, Shenzhen University, China. His research interests include color constancy, image processing, and video enhancement.

KE SUN received the B.S. degree from Donghua University, Shanghai, in 2009, and the M.S. degree from the University of St. Andrews, U.K., in 2013. He is currently pursuing the Ph.D. degree with the University of Nottingham Ningbo China. He is also a Research Assistant with the College of Civil Engineering, Shenzhen University, China. His research interests include artificial intelligence, computer vision, and natural language processing.

JINGXIN LIU received the master's degree in signal processing and communications from The University of Edinburgh and the Ph.D. degree from the University of Nottingham, U.K. He is currently a Post-Doctoral Researcher with the College of Information Engineering, Shenzhen University. His main interests include medical image processing, computer-aided diagnosis, and computer-vision-related areas.

BOLEI XU received the Ph.D. degree from the University of Nottingham, U.K. He is currently a Post-Doctoral Researcher with the College of Information Engineering, Shenzhen University.

JIANG DUAN received the degree from Southwest Jiaotong University, the master's degree in color science from Derby University, and the Ph.D. degree from the University of Nottingham, U.K. He then did a post-doctoral research at Northwestern University, USA and was with the Xerox Research Center Europe. He is currently a Professor with the Southwestern University of Finance and Economics and also an Honorary Professor with the University of Nottingham, U.K.

GUOPING QIU is currently a Distinguished Professor with the College of Information Engineering, Shenzhen University, China, and a Professor of visual information processing with the School of Computer Science, University of Nottingham. He holds several U.S. and European patents. His general research interests include image processing, pattern recognition, computer vision, machine learning, data mining, and their real-life applications. His particular research focus and expertise is in high dynamic range imaging and video, multimedia content analysis and processing, content-based image indexing and retrieval, automatic image annotation, and medical image analysis (digital pathology). He received the Best Paper Award at the 18th International Conference on Pattern Recognition 2006.