Aesthetic-Driven Image Enhancement by Adversarial Learning

Yubin Deng Chen Change Loy Xiaoou Tang
Department of Information Engineering, The Chinese University of Hong Kong
{dy015, ccloy, xtang}@ie.cuhk.edu.hk

Abstract

We introduce EnhanceGAN, an adversarial learning based model that performs automatic image enhancement. Traditional image enhancement frameworks involve training separate models for automatic cropping or color enhancement in a fully-supervised manner, which requires expensive annotations in the form of image pairs. In contrast to these approaches, our proposed EnhanceGAN only requires weak supervision (binary labels on image aesthetic quality) and is able to learn enhancement parameters for tasks including image cropping and color enhancement. The full differentiability of our image enhancement modules enables training the proposed EnhanceGAN in an end-to-end manner. A novel stage-wise learning scheme is further proposed to stabilize the training of each enhancement task and facilitate the extensibility for other image enhancement techniques. Our weakly-supervised EnhanceGAN reports competitive quantitative results against supervised models in automatic image cropping using standard benchmarking datasets, and a user study confirms that the images enhancement results are on par with or even preferred over professional enhancement.

1. Introduction

Image enhancement is considered a skillful artwork that involves transforming or altering a photograph using various methods and techniques to improve the aesthetics of a photo. Examples of enhancement include color enhancement (such as adjusting color / contrast / white balance) and image cropping (removing unwanted elements and improving image composition), as shown in Fig. 1. This task is conventionally conducted manually through some professional tools such as *Adobe Photoshop*. Manual editing is time-consuming even for a professionally trained artist. While there are an increasing number of applications (*e.g.,* Fotor[1] and Instagram[2]) that allow casual users to choose a fixed set of filters and perform cropping in more convenient ways, human involvement is still inevitable. Given the increasing amount of digital photographs captured daily with mobile devices, it is desirable to perform image enhancement with minor human involvement or even without human supervision at all.

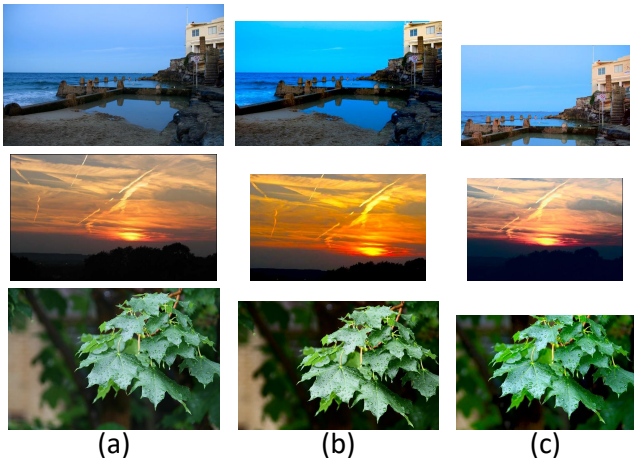

Figure 1. Examples of image enhancement given original input (a). Can you distinguish which figure is enhanced by professional and which by adversarial learning model? (Answer in footnote[4]; best viewed in color.)

Is it possible to directly learn a supervised model that allows full automatic image enhancement without human intervention? Previous research efforts have shown some success in automating some of the processes therein, including cropping [2, 3, 35, 37], color enhancement [19, 28, 36] and style transfer [17]. However, most of these models are prone to the overfitting problem as existing datasets are either small in the number of images or limited to a few style categories. For instance, we need to provide uncropped and cropped image pairs to learn the capability of cropping [3, 37]. Unfortunately, such data is scarce due to the expensive cost of obtaining professional annotations. To date, the largest publicly available cropping dataset only consists of ∼1000 training image pairs [3, 37], which is far

[1] www.fotor.com
[2] www.instagram.com

[4] Row 1: (b) adversarial learning model (c) professional
Row 2: (b) adversarial learning model (c) professional
Row 3: (b) professional (c) adversarial learning model

from enough to train or fine-tune a deep network without the risk of overfitting. Obtaining groundtruth labels for the task of image color enhancement is even more difficult, not mentioning the subjective nature when judging the image color. It is hence non-trivial to reach the goal of automatic image enhancement with a pure learning-based method.

In this study, we focus on technical image enhancement but not creative image enhancement [17]. We address this problem in a weakly-supervised manner based on the notion that an edited image should gain increased aesthetic quality compared to the original image. In particular, we attempt to learn image enhancement from images with only binary labels on aesthetic quality, *i.e.*, good or poor quality. As the edited image should be closer in the sense of aesthetic quality to the photo collections by professional photographers compared to the original image with poor aesthetic quality, this notion can be well formulated in an adversarial learning framework [8]. Specifically, we have a discriminator D that attempts to distinguish images of poor and good aesthetic quality. Such a network can be trained by an abundant amount of images from existing aesthetic datasets [22, 29]. A generator G, on the other hand, generates a set of manipulation parameters given an image with poor aesthetic quality. The task of G is to fool D so that D confuses the G's outputs (*i.e.*, enhanced images) as images with high quality.

We point out the most obvious differences of our model to existing General Adversarial Networks (GAN) [8] as follows. (1) Our method no longer generates *new natural images* conditioned on a random vector [1, 25]. In contrast, our framework specifically enhances the given user input that is of poor aesthetic quality. The generator G is tasked to generate a set of enhancement parameters conditioned on this image input. (2) We carefully design the enhancement modules to be fully-differentiable for end-to-end learning of image enhancement parameters in the adversarial learning framework. (3) Moreover, our generator network learns multiple forms of image enhancement, which can be regarded as a process of multiple-task learning. Ensuring effective learning of each task in G becomes a challenging problem. Specifically, D might overweight one of the tasks, *i.e.*, D may be easily fooled by good contrast although the generated images have poor composition. We alleviate this problem by stage-wise learning, *i.e.*, we train G-D to focus on a specific task at each stage before moving to the next. Extra precautions are introduced to constrain the inputs to D at different stages.

Our study is the first attempt that investigates weakly-supervised image enhancement by adversarial learning. Our main contribution is three-fold:

1. Our model leverages abundant images that are annotated only with good and poor quality labels. *No knowledge of the groundtruth enhancement action is given to the system.* Image enhancement parameters are learned through adversarial learning driven by aesthetic judgement.
2. Our framework permits multiple forms of image enhancement. Unlike conventional methods that focus on only a single task (*e.g.*, cropping or color enhancement), our model is capable of encapsulating various functions such as scaling, translation, contrast adjustment, and color enhancement in a unified network.
3. We discuss effective ways to incorporate different functionalities in a single network using stage-wise training. Our EnhanceGAN is hence extensible to include further image enhancement schemes (provided that the enhancement operations are fully-differentiable) without worrying whether we have additional labels.

Owing to the subjective nature of image enhancement, we show the effectiveness of our EnhanceGAN for image enhancement in two sets of evaluations. We quantitatively evaluate the performance of image cropping on two benchmark datasets, and we perform a blind user study to compare our method with human enhancement.

2. Related Work

Aesthetic Quality Assessment: Computationally understanding the aesthetic quality of an image is by itself a challenging open research problem. The task of aesthetic quality assessment is to distinguish high-quality photos from low-quality ones based on human perceived aesthetics. Previous successful attempts rely on a data-driven approach to train convolutional neural networks for binary classification of image quality [20, 21] or aesthetic score regression [18]. We refer readers to a comprehensive study [5] on the state-of-the-art models on image aesthetic assessment. Although the focus of image enhancement is not on assessing the quality of a given image, our work is closely related to this research domain in the sense that image enhancement aims at improving the aesthetic quality of the given input.

Automatic Image Enhancement: The majority of techniques for image manipulation with the goal to enhance the aesthetic quality of an image can be divided into two genres, namely (1) cropping and re-targeting, (2) color enhancement. Pixel-level manipulation and image restoration (*e.g.,* super resolution [6], de-haze [9] and de-artifacts [33]) are also closely related to image enhancement but is beyond the focus in this work.

Cropping and Re-targeting: Image cropping and re-targeting aim at finding the most visually significant region based on aesthetic value or human attention focus. Aesthetic-based approaches [2, 14, 35, 37] evaluate the crop window candidates based on handcrafted low-level features or learned aesthetic features, while attention-based approaches [4, 12, 16] rely on image saliency and produce

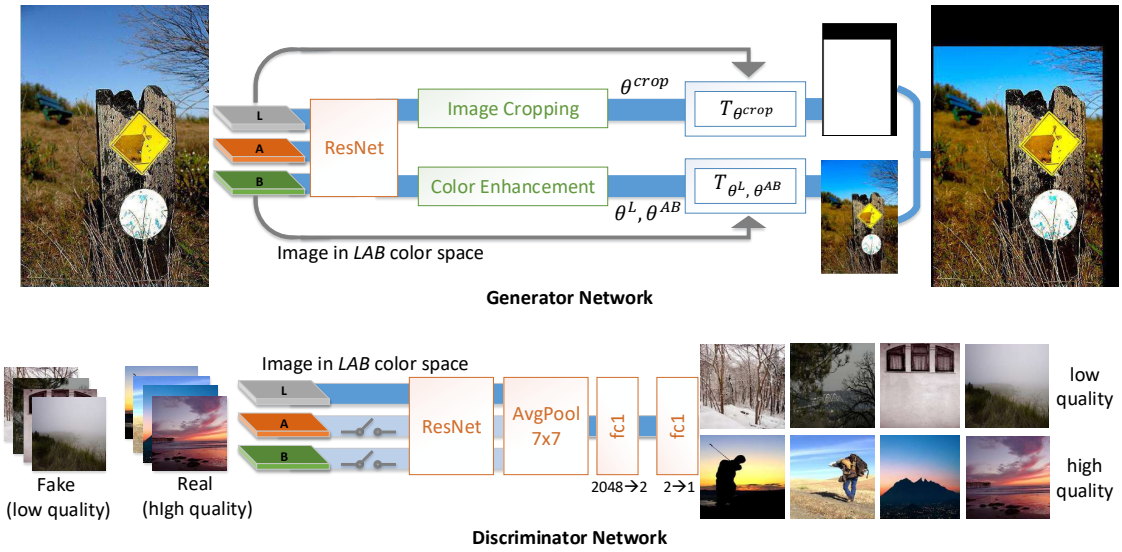

Figure 2. **The architecture of our proposed EnhanceGAN framework**. ResNet module is the feature extractor; in this work, we use the ResNet-101 [10] and removed the last average pooling layer and the final fc layer. The switch icons in the discriminator network represent zero-masking during stage-wise training as discussed in Section 3.4 (Best viewed in color.)

the cropping window encapsulating the most salient region. These systems for cropping and re-targeting are mostly based upon a limited amount of labeled cropping data (∼1000 training image pairs in total), where the cropping problem is modeled as window regression or window candidate classification in a supervised learning manner [2, 3]. Network fine-tuning with pre-trained convolutional neural network has obtained some success with extensive data augmentation [5], nevertheless, the notorious overfitting problem has yet to be solved.

Color Enhancement and Style Transfer: The visual quality of an image can be enhanced by color adjustment, where regression models and ranking models have been trained to map the input image to a corresponding enhanced groundtruth [7]. Such color mappings are learned [36, 38] from a small set of labeled data by professional editors. To alleviate the lack of sufficient labeled data, recent research efforts formulate the color enhancement problem as the color transfer problem [13, 26]. In particular, the popular exemplar-based color transfer approaches [19, 28] seek to retrieve the most suitable matching exemplar based on image content and perform color transfer onto the given input. However, they suffer from potential visual artifacts due to erroneous exemplars. Artistic style transfer [17, 31] is also closely related to color enhancement, but their focus is to transform an input image into an output that matches the artistic style of a given exemplar, typically a painting or a drawing. Unlike color/style transfer studies that re-style images based on a specific input style, we aim at enhancing an image through spatial and chrominance manipulations driven by content and aesthetic quality of the individual image. Thus each image would experience different manipulations. Unlike exemplar matching [19, 28], our work does not require exemplars for style transfer, and hence our model is not limited by the subset of exemplars.

3. Methodology

We formulate the problem of image enhancement in an adversarial learning framework [8]. Specifically, our proposed approach builds upon Wasserstein GAN (W-GAN) by Arjovsky *et al*. [1]. The full architecture of our proposed framework is shown in Fig. 2. In the following, we first give a brief review on general GAN learning framework before diving into the core design of our generator and discriminator.

3.1. Preliminary

Generative Adversarial Network (GAN) [8] has shown a powerful capability of generating realistic natural images. Typical GANs contain a generator G and a discriminator D, and it was proven [8] that the minimax game

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{I}\sim p_{data}}[\log D(\mathbf{I})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[\log(1 - D(G(\mathbf{z})))] \quad (1)$$

would reach a global optimum when p_g converges to the real data distribution p_{data}, where p_g is the distribution of the samples $G(\mathbf{z})$ obtained when $\mathbf{z} \sim p_{\mathbf{z}}$, and $\mathbf{z}$ is a random or encoded vector. At that point, the discriminator cannot distinguish the images $\mathbf{I} \sim p_{data}$ from the ones of $G(\mathbf{z})$. Alternative GAN frameworks have been proposed to stabilize the training process of GAN, such as DC-GAN [25] and W-GAN [1], which are less prone to mode collapse. In

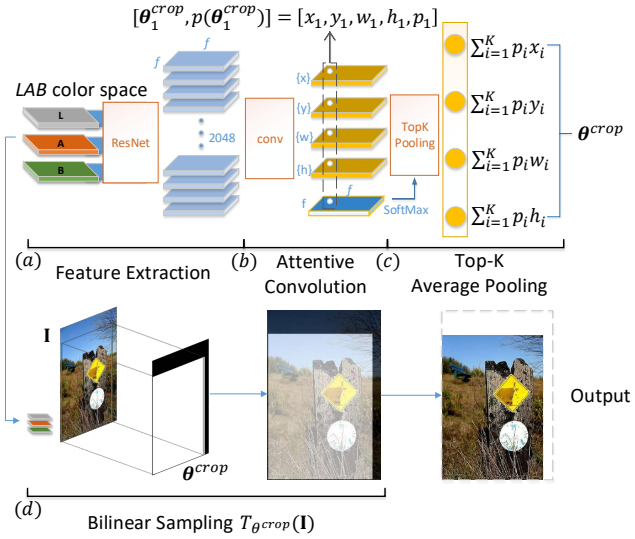

Figure 3. **Image cropping module**. We perform pooling from Top-K predicted cropping coordinates. The $\boldsymbol{\theta}_1^{crop}$ shows the first one of the total f^2 coordinates in the coordinate pool generated from a convolution layer. Top-K average pooling [32] is used to produce the final crop window for bilinear sub-sampling [15]. (Best viewed in color.)

this work we follow the practice in [1] and adopt the loss function based on Wasserstein distance as follows,

$$\begin{aligned} L &= \mathbb{E}_{\mathbf{I}\sim p_{data}}[f_W(\mathbf{I})] - \mathbb{E}_{\mathbf{I}\sim p_{gen}}[f_W(\mathbf{I})] \\ &\approx \mathbb{E}_{\mathbf{I}\sim p_{data}}[D(\mathbf{I})] - \mathbb{E}_{\mathbf{I}\sim p_{gen}}[D(\mathbf{I})] \end{aligned} \tag{2}$$

where $f_W(\cdot)$ is a K-Lipschitz function parameterized by W, which is approximated by our discriminator network D as detailed in Sec. 3.3.

3.2. Generator Network (Net-G)

Different from existing GAN frameworks, our generator does not generate images by itself. Instead, the generator G in our EnhanceGAN is responsible for learning a set of image enhancement parameters $\{\boldsymbol{\theta}\}$, with which a corresponding set of differentiable transformations $\{T_{\boldsymbol{\theta}}\}$ will apply sequentially on the given input image $\mathbf{I}$:

$$\mathbf{I}^{\text{output}} = (T_{\boldsymbol{\theta}^L, \boldsymbol{\theta}^{AB}} \circ T_{\boldsymbol{\theta}^{crop}})(\mathbf{I}), \tag{3}$$

$$\boldsymbol{\theta}^L, \boldsymbol{\theta}^{AB}, \boldsymbol{\theta}^{crop} \in \{\boldsymbol{\theta} | \boldsymbol{\theta} = G(\mathbf{I})\}, \tag{4}$$

where $\boldsymbol{\theta}^L$ and $\boldsymbol{\theta}^{AB}$ denotes the enhancement parameters for luminance and color adjustments, respectively. $\boldsymbol{\theta}^{crop}$ specifies the cropping coordinates given an input image. The base architecture of our generator network is a ResNet-101 [10] without the last fully-connected layer, and we further remove the last pooling layer to preserve spatial information in the feature maps. As such, this ResNet module acts as a fully-convolutional feature extractor given an input image (see Fig. 3*a*). The 2048 output feature maps produced by the ResNet module has a spatial size $f \times f$ and is

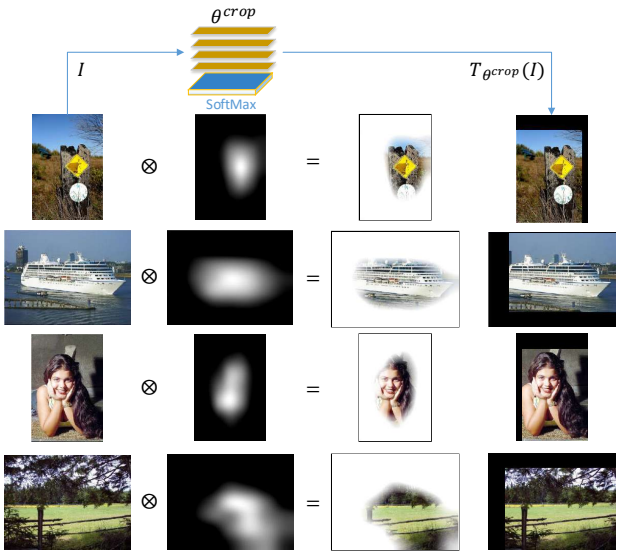

Figure 4. The focus of attention as revealed by overlaying the softmax feature map (as in Fig. 3) onto the input image. (Best viewed in color)

subsequently utilized in our enhancement parameter generation modules. In this work, we focus on learning two major image enhancement modules, namely image cropping and color enhancement.

3.2.1 Image Cropping Module

The goal of image cropping is to produce a set of cropping coordinates $\boldsymbol{\theta}^{crop} = [x, y, w, h]$ given an input image. Directly learning one single set of cropping coordinates may not be optimal as crop window prediction is multi-modal to some extents – an image can have several plausible cropping solutions (*e.g.*, a crop solution can keep the identical/similar aspect ratio, or change the image orientation from landscape to portrait or vice versa).

With inspirations drawn from attention models [34], our cropping module is built with a convolution layer ($2048 \rightarrow 5$) with kernel size 1×1. We call this process as *attentive convolution* (see Fig. 3*b*). In particular, the first 4 feature maps correspond to the candidate sets of cropping coordinates and the 5-th feature map is an $f \times f$ softmax probability map corresponding to f^2 possible predictions $\boldsymbol{\theta}_i^{crop} = [x_i, y_i, w_i, h_i], i \in \{1, 2, ..., f^2\}$, where $prob(\boldsymbol{\theta}^{crop} = \boldsymbol{\theta}_i^{crop} | \mathbf{I}) = p_i$.

We show in the experiment section that our cropping module is attentive to specific regions of the input that are relevant to the image content, hence producing reasonable crop-coordinate candidates resided on the corresponding neurons of the feature maps (see Fig. 4). Top-K average pooling [32] is adopted to aggregate the coordinate candidates with the highest probabilities (see Fig. 3*c*). Moreover, the cropping module could potentially produce only a small salient region due to instance-specific contrast in-

formation present in the feature response of a pre-trained network [5]. As the goal is to learn a good composition sub-image with better aesthetic quality rather than detecting isolated and salient objects, instance normalization [31] is adopted after the convolution layer. It is worth mentioning that directly extracting a sub-image from the input given the learned cropping coordinates $\boldsymbol{\theta}^{crop}$ is not differentiable. To ensure that the gradients can back-propagate to the cropping parameters, bilinear sampling from a sampling grid [15] is used to sub-sample the input image based on coordinates output $\boldsymbol{\theta}^{crop}$ of the cropping module (see Fig. 3*d*).

3.2.2 Color Enhancement Module

The color enhancement module learns a set of parameters $\boldsymbol{\theta}^{L}$ and $\boldsymbol{\theta}^{AB}$ targeting the adjustments for better lighting and color contrast. Specifically, brightness and lighting contrast can be adjusted based on the luminance channel of an image in the CIELab color space. We follow the idea behind gamma correction [24] to adjust the brightness and contrast of an image in pixel level (*i.e.*, the L channel of image $\mathbf{x}$ in the CIELab color space) by designing a piecewise transformation function $T_{\boldsymbol{\theta}^{L}}$ (see Fig. 5) defined on each pixel $m \in L$:

$$T_{\boldsymbol{\theta}^{L}}(m) = \begin{cases} 0 & \text{if } m \leq b \\ k_1 m^{\frac{1}{p}} & \text{if } b < m \leq a \\ m & \text{if } a < m \leq 1-a \\ k_2(m-k_3)^{\frac{1}{q}} + k_3 & \text{if } 1-a < m \leq 1-b \\ 1 & \text{if } m \geq 1-b \end{cases},$$

where $k_1 = a(a-b)^{-\frac{1}{p}}$, $k_2 = a(a-b)^{-\frac{1}{q}}$ and $k_3 = 1-a$ to ensure that $T_{\boldsymbol{\theta}^{L}}$ is continuous. We further constrain $p \geq 1$ in order to lighten the dark regions and $0 < q < 1$ to darken the over-exposed region. Similarly, we follow "The LAB Color Move"[5] and the curve adjustment instructed in [11] and design a similar process to enhance the image color. In particular, the adjustments T_A and T_B defined respectively on pixels $m \in A$ and $m \in B$ (*i.e.*, the a and b channels in image $\mathbf{x}$, see Figure 5) can be formulated as follows:

$$T_{\boldsymbol{\theta}^{A}}(m) = \begin{cases} 0 & \text{if } m \leq \alpha \\ \frac{1}{1-2\alpha}(m-\alpha) & \text{if } \alpha < m < 1-\alpha \\ 1 & \text{if } m \geq 1-\alpha \end{cases},$$

$$T_{\boldsymbol{\theta}^{B}}(m) = \begin{cases} 0 & \text{if } m \leq \beta \\ \frac{1}{1-2\beta}(m-\beta) & \text{if } \beta < m < 1-\beta \\ 1 & \text{if } m \geq 1-\beta \end{cases}.$$

[5] https://digital-photography-school.com/how-to-use-lab-color-in-photoshop-to-add-punch-to-your-images

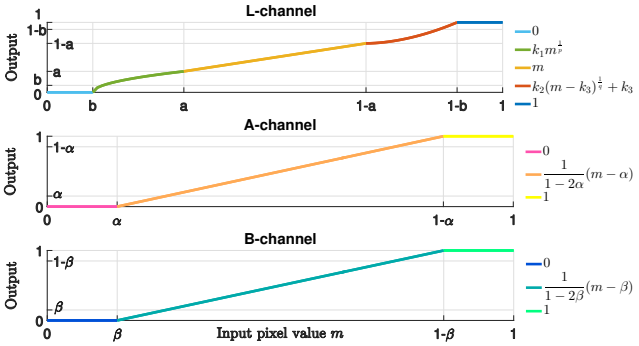

Figure 5. Transformation functions $T_{\boldsymbol{\theta}^{L}}$ (top), $T_{\boldsymbol{\theta}^{A}}$ (middle), $T_{\boldsymbol{\theta}^{B}}$ (bottom) for color enhancement. Pixel values are normalized in each channel.

Under this formulation, the parameter sets $\boldsymbol{\theta}^{L} = [a, b, p, q]$ and $\boldsymbol{\theta}^{AB} = [\boldsymbol{\theta}^{A}, \boldsymbol{\theta}^{B}] = [\alpha, \beta]$ can all be learned end-to-end by a convolution layer ($2048 \rightarrow 7$), where the 7-th feature map is also a $f \times f$ softmax probability map for Top-K averaging pooling similar to that of the cropping module (Fig. 3). More image enhancement modules with learnable parameter $\{\boldsymbol{\theta}\}$ can be easily extended in this manner provided that the transformations $T_{\{\boldsymbol{\theta}\}}$ are differentiable.

3.2.3 Generator Loss Function L_G

We formulate the loss function for the generator network as a weighted sum of an adversarial loss component, a perceptual loss component [17] and a regularization term. These terms are weighted to ensure that the loss terms are balanced in their scales. This formulation makes the training process more stable and has better performance, as compared to a baseline using only adversarial loss for the generator network. (see Sec. 6).

Adversarial Loss: Following Arjovsky *et al.* [1], the adversarial gradient to the generator network *G* is initially computed from the following loss function:

$$L_{gan} = \frac{1}{n}\sum_{i=1}^{n} D(\mathbf{I}_i^{\text{output}}) \tag{5}$$

Perceptual Loss: To prevent image cropping from a potential degradation of producing isolated high color-contrast patches from the user input, the design of the loss function should require that cropped image still possesses key content from the original image. This motivates us to adopt the feature reconstruction loss [17] to account for the semantic difference between the cropped image and the input as measured by feature similarity:

$$L_{per} = \frac{1}{n}\sum_{i=1}^{n} ||f_{vgg}(\mathbf{I}_i^{\text{output}}) - f_{vgg}(\mathbf{I}_i)||_2^2 \tag{6}$$

where f_{vgg} denotes the $fc7$ feature from the VGG-16 network [27] trained for ImageNet.

Regularization Loss: Moreover, the notion that an edited image should have better aesthetic quality (lower $f_W(\cdot)$ values) than the original further gives us an intuitive regularization for training the generator:

$$L_{reg} = \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{I}_i^{\text{output}}, \mathbf{I}_i) \tag{7}$$

where $\phi(\mathbf{I}_i^{\text{output}}, \mathbf{I}_i)$

$$= \begin{cases} 0 & \text{if } D(\mathbf{I}_i^{\text{output}}) < D(\mathbf{I}_i), \\ ||D(\mathbf{I}_i^{\text{output}}) - D(\mathbf{I}_i)||_2^2 & \text{otherwise.} \end{cases}$$

3.3. Discriminator Network (Net-D)

The proposed framework consists of a discriminator network that is able to assess image aesthetic quality. The discriminator network D is designed to share the ResNet-101 [10] base architecture of G during pre-training. As shown in Figure 2, the last layer for 1000-class classification in the original ResNet-101 is replaced by a 2-neuron fully-connected layer. We pre-train discriminator D for binary aesthetic classification with the cross-entropy loss as in [5]. After pre-training, the discriminator network D is appended with another 1-neuron fully-connected layer to perform output aggregation as an approximator to f_W in Eq. (2, 5, 6, 7). Deriving from Eq. 2, the loss function L_D in subsequent adversarial training can be written as:

$$L_D = E_{\mathbf{I}\sim p_{\text{good}}}[f_W(\mathbf{I}^{\text{good}})] - E_{\boldsymbol{\theta}\sim\mathbf{I}^{\text{bad}}}[f_W(\mathbf{I}^{\text{output}})], \tag{8}$$

where $\mathbf{I}^{\text{output}} = T_{\{\boldsymbol{\theta}\}}(\mathbf{I}^{\text{bad}})$, $\mathbf{I}^{\text{bad}} \sim p_{\text{bad}}$.

3.4. Multi-Stage Training and Constraint

Training such a framework for image enhancement is non-trivial as D may bias to overweight one of the image enhancement operations. To ensure D learns the aesthetic importance of each of the tasks, we propose a stage-wise training scheme. The base network of G is fixed after pre-training and is used as a shared feature extractor for all image enhancement modules. In the first stage, we only enable the image cropping module and fix the color enhancement transforms to be identity: $T_{\boldsymbol{\theta}^c}(m) = m$, where $c \in \{L, A, B\}$ (see Sec. 3.2.2). The A and B channels of the enhanced image are zero-masked before entering the discriminator D. The masking is intuitive as we want the discriminator D to focus more on image composition than color contrast. In the second stage, we freeze the image cropping module to learn the enhancement parameters $\boldsymbol{\theta}^L$ and $\boldsymbol{\theta}^{AB}$ for the color enhancement module. Zero-masking is dropped and D sees the full-color image in the second stage.

While pairing each of the low-quality images and the high-quality images, we ensure them to be close in their semantic class during training. This step is critical as we want to prevent the degenerated cases of enhancing an input based on wrong exemplars (*e.g.*, enhancing a colorful landscape photo based on a black-and-white portraiture close-shot), hence leading to a more stable adversarial gradient. This constraint is achieved by sampling from the k-nearest neighbor (k-NN) in the feature space of f_{vgg} of the given low-quality input when selecting the input data to train the discriminator.

4. Experiments

Our weakly-supervised EnhanceGAN is tasked to learn enhancement parameters based on the binary label on image aesthetic quality. Specifically, we train the EnhanceGAN with the benchmark datasets used in aesthetic quality assessment and perform quantitative evaluations on reserved testing data. A user study is also performed to confirm the validity of our quantitative evaluation.

4.1. Datasets for Training

CUHK-PhotoQuality Dataset (CUHK-PQ) [29] contains 4,072 high-quality images and 11,812 low-quality images. This dataset is used to pre-train the feature extractor (ResNet module) of our EnhanceGAN, with 10% of the images reserved for validation. We follow the training protocol as in [5] and pre-train the ResNet-module for binary image aesthetic assessment (see Sec.3.3), obtaining a balanced accuracy [5] of 94.3% on the validation set.

AVA Dataset [22]: The Aesthetic Visual Analysis (AVA) dataset is by far the largest benchmark for image aesthetic assessment. Each of the 255,530 images is labeled with aesthetic scores ranging from 1 to 10. We follow [22] and partition and images into high-quality set and low-quality set based on the average scoring. In this study we select a subset of low-quality images based on the semantic tags provided in the AVA data for analysis[6], covering a diverse set of images that require different enhancement operations for quality enhancement. In particular, the "Real" input to the discriminator in our EnhanceGAN is chosen from the top 30% of the high-quality images. The "Fake" inputs are the low-quality images that have an average score < 5. A total of $20,000$ low-quality images are used for training the EnhanceGAN, each paired with 5 high-quality images from the k-NN pool ($k = 5$ as in Sec. 3.4). This is the only form of supervision in our training framework, and **no groundtruth enhancement operation for the low-quality input is provided**. We reserve 100 random images from the standard testing partition of the AVA testing partition (denoted as Val^{100}) for evaluation.

[6]This corresponds to 9 semantic classes in the AVA dataset, including Landscape, Seascape, Cityscape, Rural, Sky, Water, Nature, Animals and Portraiture.

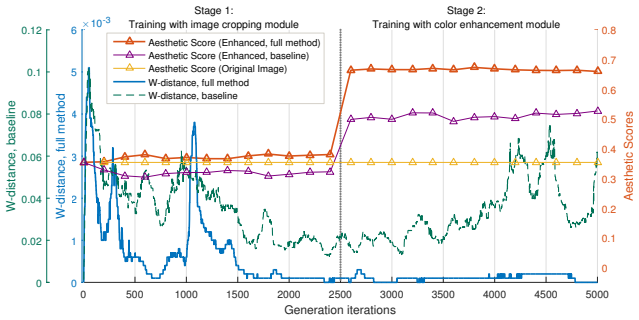

Figure 6. Training curve of our EnhanceGAN at different training iterations. The aesthetic score is produced by the softmax probability output of the VGG16 model in [5]. W-distance is plotted using the absolute value of L_D as in Eq. 8. We also plot the aesthetic scores of a baseline by training the generator without L_{per} and L_{reg}, which shows inferior performance than using the full loss L_G as in Sec. 3.2.3.

4.2. GAN Training Details

The generator network *G* in our EnhanceGAN is fully convolutional, allowing for arbitrary-sized input in CIELab color space. In our experiment we set the image input size to be 224×224, resulting in 2048 feature maps with spatial size of 7×7 (see Fig. 3). We found K = 3 in Top-K averaging pooling (see Fig. 3) to be robust in producing parameter candidates. Using a larger K value or global average pooling (*i.e.,* selecting from a large number of parameter candidates) could potentially suffer from noisy parameter predictions, while max pooling only concerns one single prediction and is prone to error. We use RMSprop [30] with a learning rate $lr_G = 5e^{-5}$ for the generator network and $lr_D = 5e^{-7}$ for the discriminator network after pretraining. The batch size is set to 64 for training the generator network and the batch size for the discriminator is 8 due to memory constraints. Each training stage (see Sec. 3.4) consists of 2500 generation iterations; in total, the EnhanceGAN is trained with 5000 generation iterations.

4.3. Quantitative Evaluation

Image Aesthetic Assessment: We monitor the training behavior by plotting the Wasserstein distance [1] against the average aesthetic score (range $[0, 1]$) on Val^{100} as predicted by the image aesthetic assessment model [5]. As shown in Fig. 6, the Wasserstein distance drops consistently across the training iterations as in [1]. More importantly, we observe that the final average score of the 100 images in Val^{100} has increased to 0.673 from 0.354 after training. This is consistent with the user study shown in Sec. 4.4.

Automatic Image Cropping: We evaluate the performance of our EnhanceGAN via two benchmark photo cropping datasets. (1) CUHK Image Cropping Dataset [37] contains 950 images that have 3 sets of cropping groundtruth by 3 different annotators. (2) Flickr Cropping Dataset [3] contains a standard partition of 1,394 training images and 345 testing images. We directly evaluate the image cropping performance on the 950 images of CUHK Image Cropping Dataset and on the 345 testing images in Flickr Cropping Dataset. We use the average overlap ratio and average boundary displacement error as in [3, 37] to evaluate the cropping performance:

$$\text{Overlap Ratio} = \frac{Area^{gt} \cap Area^{crop}}{Area^{gt} \cup Area^{crop}}, \quad (9)$$

$$\text{Displacement Error} = \sum_{k \in x,y,w,h} ||B_k^{gt} - B_k^{crop}||/4, \quad (10)$$

where $Area^{gt}$ is the area of the groundtruth crop window and $Area^{crop}$ is the area of the predicted crop; B_k is the normalized boundary coordinate k, $k \in \{x, y, w, h\}$. The results are summarized in Table 1 and Table 2. Note that our EnhanceGAN is NOT further finetuned with any groundtruth cropping labels from these datasets, yet it achieved competitive performance and even surpassed some methods with full supervision.

Table 1. Quantitative evaluation on CUHK Image Cropping Dataset [37]. The first number is average overlap ratio, higher is better. The second number (shown in parenthesis) is average boundary displacement error, lower is better. We show the ratios and errors evaluated against the groundtruth by *Photographer1* in [37] due to page limit. Complete results are shown in supplementary materials. Our EnhanceGAN is reasonably competitive as it is not trained with any cropping labels in this dataset.

Unsupervised methods	Ratio (Error)
EnhanceGAN w/o L_{per} & L_{reg}	0.4852 (0.1432)
Chen *et al.* [3]	0.6643 (0.092)
EnhanceGAN	**0.7145 (0.0765)**
Full supervision	
Park *et al.* [23]	0.6034 (0.1062)
Yan *et al.* [35]	0.7487 (0.0667)
Yan *et al.* [37]	0.7974 (0.0528)
Deng *et al.* [5]	0.8059 (0.0310)

Table 2. Quantitative evaluation on Flickr Cropping Dataset [3]. Apart from the fully-supervised baseline provided in [3], we select VGG and ResNet models pre-trained with ImageNet data and/or aesthetic data and finetune these baselines with the training partition in Flickr Cropping Dataset. Data augmentation is performed as in [5].

Unsupervised methods	Ratio (Error)
EnhanceGAN w/o L_{per} & L_{reg}	0.5103 (0.1282)
EnhanceGAN	**0.6333(0.0981)**
Full supervision	
Chen *et al.* [3]	0.6019 (0.1060)
VGG-19 (ImageNet)	0.6748 (0.0840)
VGG-16 (Deng *et al.* [5])	0.6786(0.0813)
ResNet-101 (ImageNet)	0.6822 (0.0815)

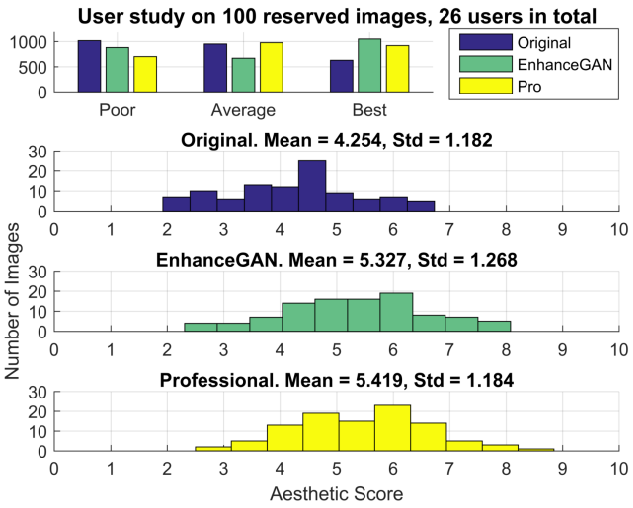

Figure 7. User study on Val^{100}. Our EnhanceGAN shows competitive performance as compared to human editing.

4.4. User Study

The subjective nature of image enhancement evaluation requires that our model should be validated through a human survey. For the purpose of the user study, we asked a professional editor to enhance each of the 100 images in Val^{100} in *Adobe Photoshop*, with restrictions to making adjustments on brightness, color and image compositions. Image editing options including the tools "Levels", "Curves", "Auto Tone", "Auto Contrast" and "Auto Color" in *Adobe Photoshop* were available to the professional editor. All enhanced images were stored using sRGB JPEG format with the highest quality (Quality = 12 in *Adobe Photoshop*). We wrote a ranking software and distributed to a total of 26 participants. All participants were shown a sequence of 100 image sets, where each image set contained the original image, the enhanced image by our EnhanceGAN, and the professionally edited image in random order. Participants were instructed to rank each set by clicking the best-quality image, the average-quality image, and the poor-quality image on the screen in order. No time constraints were placed.

Figure 7 shows the results of our user study. Each image in Val^{100} received 26 ratings, where we assign "best quality", "average quality" and "poor quality" to aesthetic scores of 10, 5 and 0, respectively. We observe that among the images ranked as the "best quality", the majority is from images enhanced by our EnhanceGAN and the ones edited by the professional. Our EnhanceGAN reaches a mean aesthetic score of 5.327, as compared to 5.419 for professional images and 4.254 for original input. Some of the images enhanced by EnhanceGAN receive even higher voting than the ones produced by the professional editor (see Fig. 8). The results demonstrate the effectiveness of our EnhanceGAN for automatic image enhancement and confirm our quantitative evaluation results as in Sec. 6.

Figure 8. Visual results from our user study: (a) original image; (b) enhanced by professional; (c) enhanced by EnhanceGAN. The box plot below shows the ranking for each image, and the amount of dots denotes the number of users who gives a particular rank as in {poor, average, best}. We show more examples and failure cases in the supplementary material.

5. Conclusion

In this work, we introduce EnhanceGAN for automatic image enhancement. Our framework only requires weak supervision in the form of binary label on aesthetic quality and learns enhancement parameters in an aesthetic-driven manner. Our EnhanceGAN is fully-differentiable, and it can be trained end-to-end in a stage-wise training setting, which allows extensibility to learn further enhancement capabilities by adding enhancement parameter modules in subsequent training stages. We quantitatively evaluate the performance of EnhanceGAN and show that the high-quality results produced by our enhancement model are on par with or even surpass professional editing by a user study.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv:1701.07875*, 2017. 2, 3, 4, 5, 7

[2] J. Chen, G. Bai, S. Liang, and Z. Li. Automatic image cropping: A computational complexity study. In *CVPR*, 2016. 1, 2, 3

[3] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, and B.-Y. Chen. Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study. *WACV*, 2017. 1, 3, 7, 10

[4] J. Choi and C. Kim. Object-aware image thumbnailing using image classification and enhanced detection of roi. *Multimedia Tools and Applications*, 75(23), 2016. 2

[5] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017. 2, 3, 5, 6, 7, 10

[6] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 38(2), 2016. 2

[7] H. S. Faridul, T. Pouli, C. Chamaret, J. Stauder, A. Trémeau, E. Reinhard, et al. A survey of color mapping and its applications. In *Eurographics (State of the Art Reports)*, 2014. 3

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 2, 3

[9] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *TPAMI*, 33(12), 2011. 2

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 4, 6

[11] J. Hosie-Bounar, K. Hart, and M. Geller. *New Perspectives on Adobe Photoshop CS5, Comprehensive*. Cengage Learning, 2011. 5

[12] J. Huang, H. Chen, B. Wang, and S. Lin. Automatic thumbnail generation based on visual representativeness and foreground recognizability. In *ICCV*, 2015. 2

[13] Y. Hwang, J.-Y. Lee, I. So Kweon, and S. Joo Kim. Color transfer using probabilistic moving least squares. In *CVPR*, 2014. 3

[14] M. B. Islam, W. Lai-Kuan, and W. Chee-Onn. A survey of aesthetics-driven image recomposition. *Multimedia Tools and Applications*, 2016. 2

[15] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 4, 5

[16] N. Jaiswal and Y. K. Meghrajani. Saliency based automatic image cropping using support vector machine classifier. In *International Conference on Innovations in Information, Embedded and Communication Systems*, 2015. 2

[17] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*. Springer, 2016. 1, 2, 3, 5

[18] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*. Springer, 2016. 2

[19] J.-Y. Lee, K. Sunkavalli, Z. Lin, X. Shen, and I. So Kweon. Automatic content-aware color and tone stylization. In *CVPR*, 2016. 1, 3, 10, 11

[20] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *ICCV*, 2015. 2

[21] L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *CVPR*, 2016. 2

[22] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 2, 6

[23] J. Park, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon. Modeling photo composition and its application to photo rearrangement. In *ICIP*, 2012. 7, 10

[24] D. G. Pelli and L. Zhang. Accurate control of contrast on microcomputer displays. *Vision research*, 1991. 5

[25] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016. 2, 3

[26] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5), 2001. 3

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 5

[28] W.-T. Sun, T.-H. Chao, Y.-H. Kuo, and W. H. Hsu. Photo filter recommendation by category-aware aesthetic learning. *arXiv:1608.05339*, 2016. 1, 3

[29] X. Tang, W. Luo, and X. Wang. Content-based photo quality assessment. *TMM*, 15(8), 2013. 2, 6

[30] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012. 7

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016. 3, 5

[32] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017. 4

[33] Z. Wang, D. Liu, S. Chang, Q. Ling, Y. Yang, and T. S. Huang. D3: Deep dual-domain based fast restoration of jpeg-compressed images. In *CVPR*, 2016. 2

[34] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, 2015. 4

[35] J. Yan, S. Lin, S. Bing Kang, and X. Tang. Learning the change for automatic image cropping. In *CVPR*, 2013. 1, 2, 7, 10

[36] J. Yan, S. Lin, S. Bing Kang, and X. Tang. A learning-to-rank approach for image color enhancement. In *CVPR*, 2014. 1, 3

[37] J. Yan, S. Lin, S. B. Kang, and X. Tang. Change-based image cropping with exclusion and compositional features. *IJCV*, 114(1), 2015. 1, 2, 7, 10

[38] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu. Automatic photo adjustment using deep neural networks. *ToG*, 35(2), 2016. 3

Abstract

The following are the supplementary materials corresponding to the paper "Aesthetic-Driven Image Enhancement by Adversarial Learning".

6. Detailed quantitative results on image cropping

We show more detailed results in Table 3 for the quantitative analysis of the performance of our EnhanceGAN on the CUHK Photo Cropping Dataset, corresponding to Table 1 at page 7 in the main paper.

Unsupervised methods	Photographer1	Photographer2	Photographer3
EnhanceGAN w/o L_{per} & L_{reg}	0.4932 (0.140)	0.4735 (0.148)	0.4889 (0.1422)
Chen *et al*. [3]	0.6643 (0.092)	0.6556 (0.095)	0.6439 (0.099)
EnhanceGAN	$\mathbf{0.7145(0.0765)}$	$\mathbf{0.7011(0.0805)}$	$\mathbf{0.7022(0.0795)}$
Full supervision			
Park *et al*. [23]	0.6034 (0.1062)	0.5823 (0.1128)	0.6085 (0.1102)
Yan *et al*. [35]	0.7487 (0.0667)	0.7288 (0.0720)	0.7322(0.0719)
Yan *et al*. [37]	0.7974 (0.0528)	0.7857 (0.0567)	0.7723 (0.0594)
Deng *et al*. [5]	0.8059 (0.0310)	0.7750 (0.0375)	0.7725 (0.0377)

Table 3. Quantitative Evaluation on CUHK Image Cropping Dataset [37]. The first number is average overlap ratio, higher is better. The second number (shown in parenthesis) is average boundary displacement error, lower is better. Our EnhanceGAN is reasonably competitive as we have not provided any cropping labels in this dataset.

7. More Visual Results

7.1. Color enhancement results compared with [19]

We present color enhancement results of our EnhanceGAN on the released 5-image evaluation set in [19]. Given an input image, the exemplar based approach in [19] produces five style-transferred images based on different retrieved exemplars. We select the Top-1 output produced by [19] for comparison, as shown in Fig. 9 (page 11). We observe that our EnhanceGAN produces more natural color enhancement results while the method in [19] features a more aggresive change in image styles.

7.2. Results on Val^{100} in our user study

We present more visual results of our EnhanceGAN on the reserved 100-image set Val^{100}. Similar to Figure 8 in our main paper at page 8, we show a box plot for user rankings below each of the images. See Fig. 10 for the animal category (page 12), Fig. 11 for scenery images (page 13), Fig. 12 for sky and seascapes (page 14) and Fig. 13 for nature. Potential failure cases could be seen from the over-exposure in some of our examples on portraiture images, see Fig. 14 (page 16).

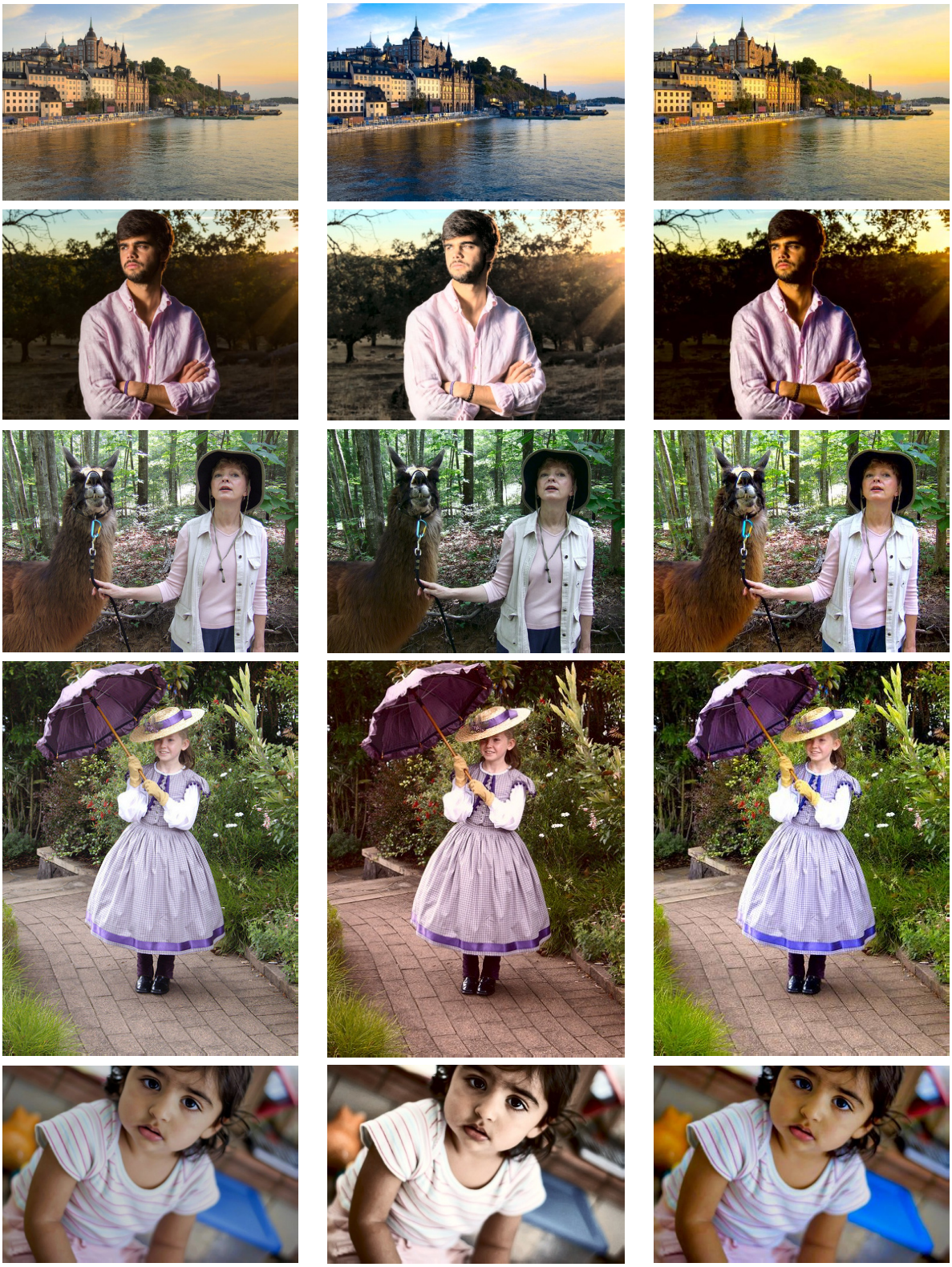

Figure 9. Visual results on the 5-image evaluation set in [19]. Column 1: input image; Column 2: output by the exemplar-based method in Lee et al. [19]; Column 3: output by our EnhanceGAN. Our EnhanceGAN produces more natural color enhancement results while the method in [19] features a more aggresive change in image styles. (Best viewed in color.)

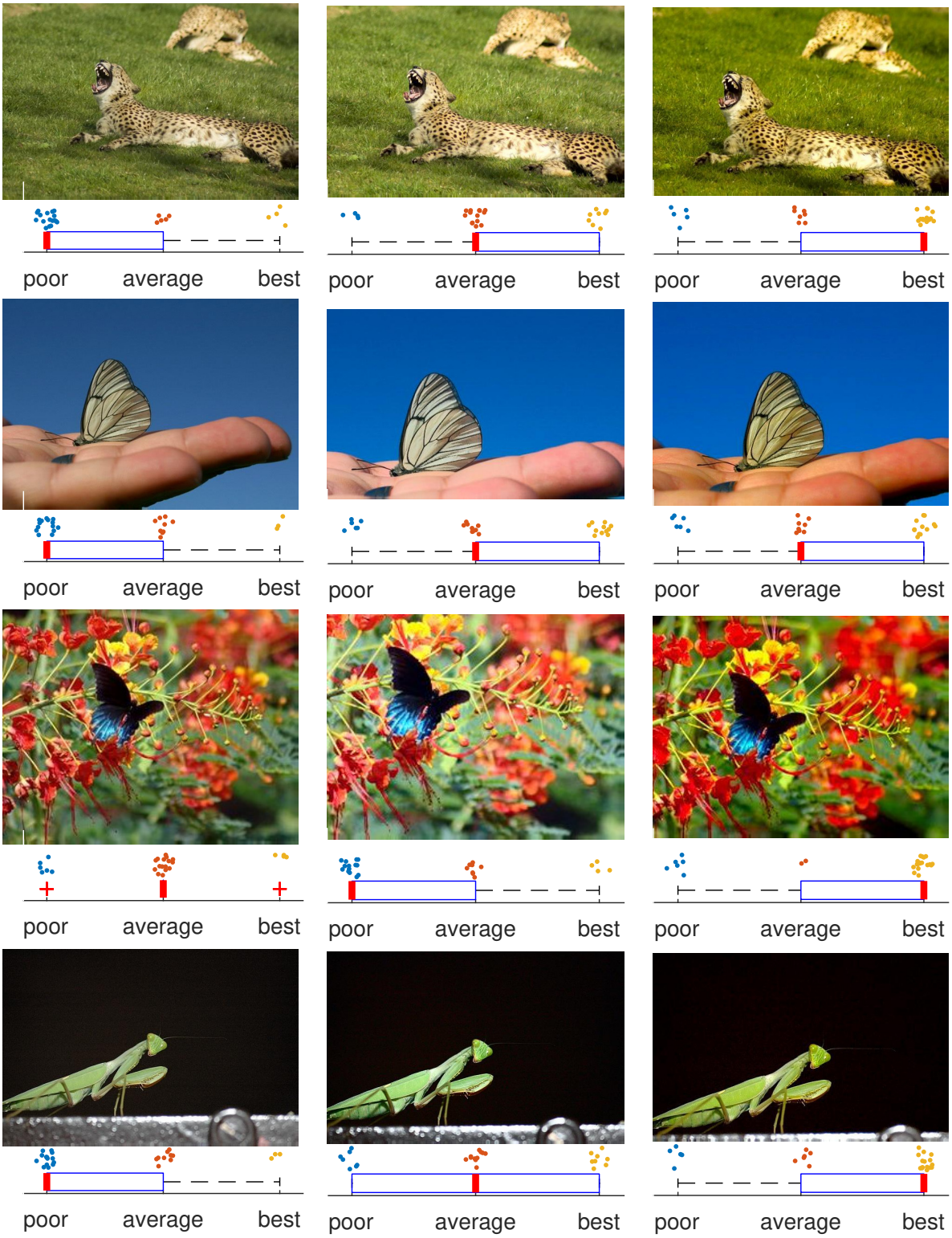

Figure 10. Visual results from our user study: (a) original image; (b) enhanced by professional; (c) enhanced by EnhanceGAN. The box plot below shows the ranking for each image, and the amount of dots denotes the number of users who gives a particular rank as in {poor, average, best}. (Best viewed in color.)

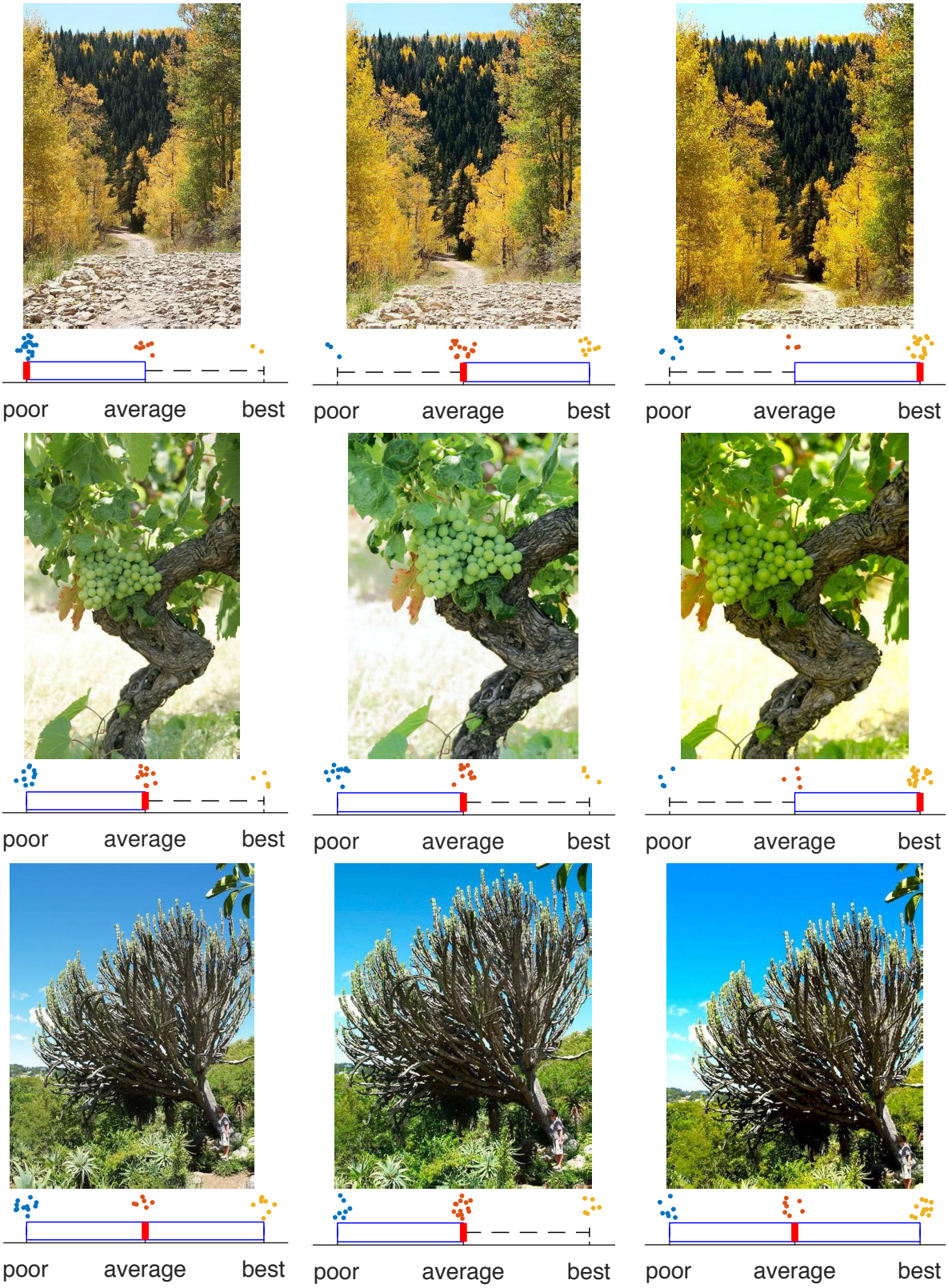

Figure 11. Visual results from our user study: (a) original image; (b) enhanced by professional; (c) enhanced by EnhanceGAN. The box plot below shows the ranking for each image, and the amount of dots denotes the number of users who gives a particular rank as in {poor, average, best}. (Best viewed in color.)

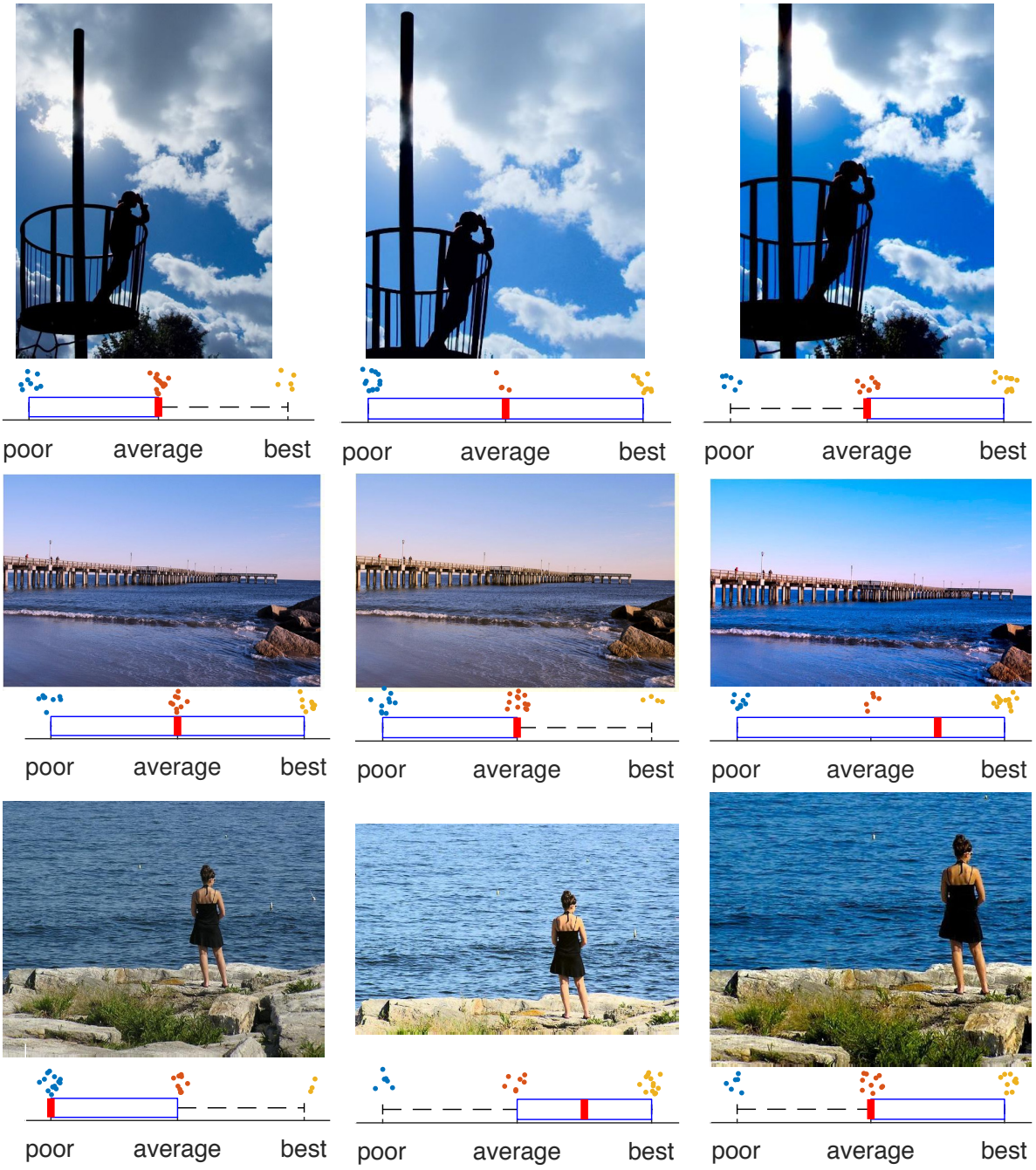

Figure 12. Visual results from our user study: (a) original image; (b) enhanced by professional; (c) enhanced by EnhanceGAN. The box plot below shows the ranking for each image, and the amount of dots denotes the number of users who gives a particular rank as in {poor, average, best}. (Best viewed in color.)

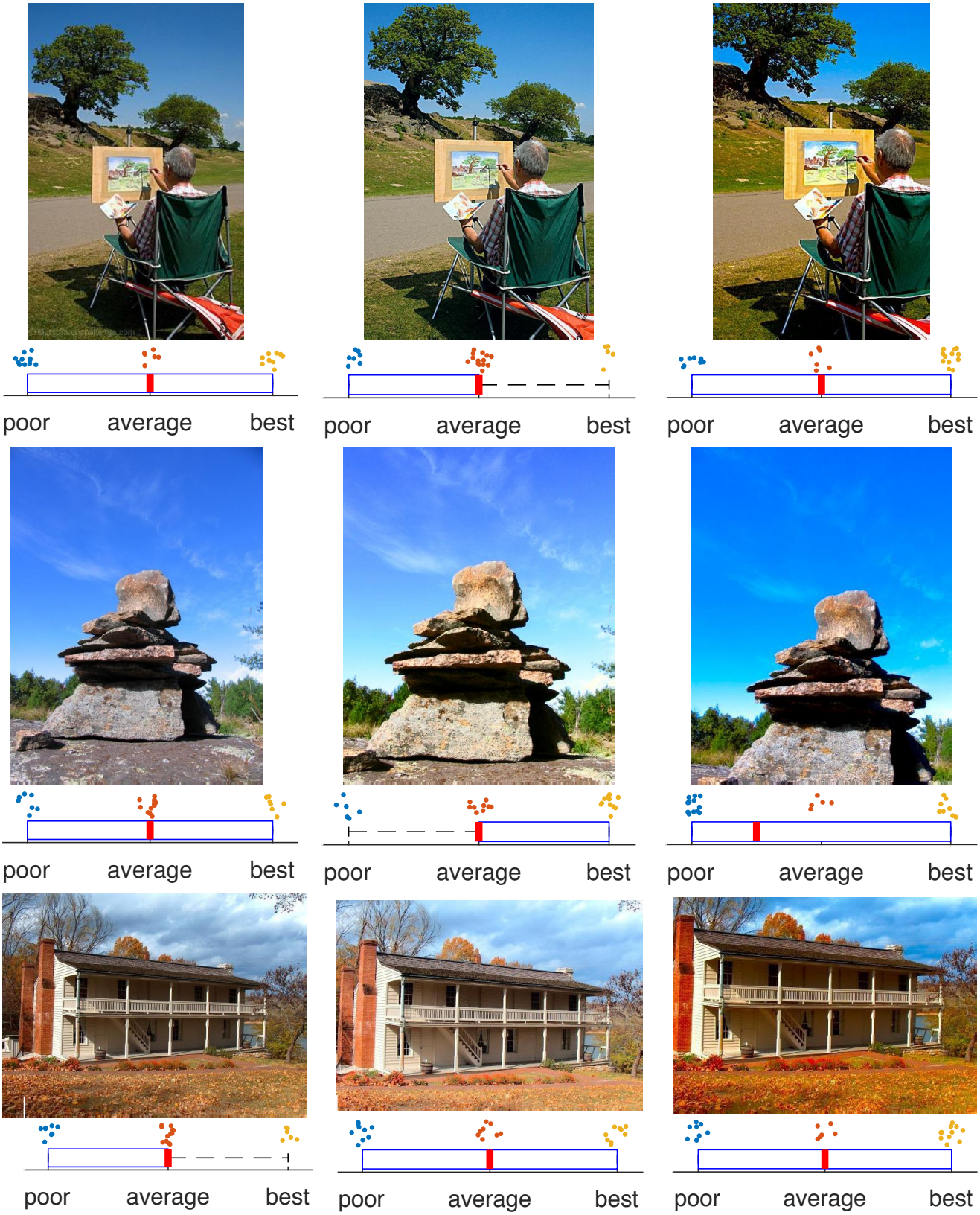

Figure 13. Visual results from our user study: (a) original image; (b) enhanced by professional; (c) enhanced by EnhanceGAN. The box plot below shows the ranking for each image, and the amount of dots denotes the number of users who gives a particular rank as in {poor, average, best}. (Best viewed in color.)

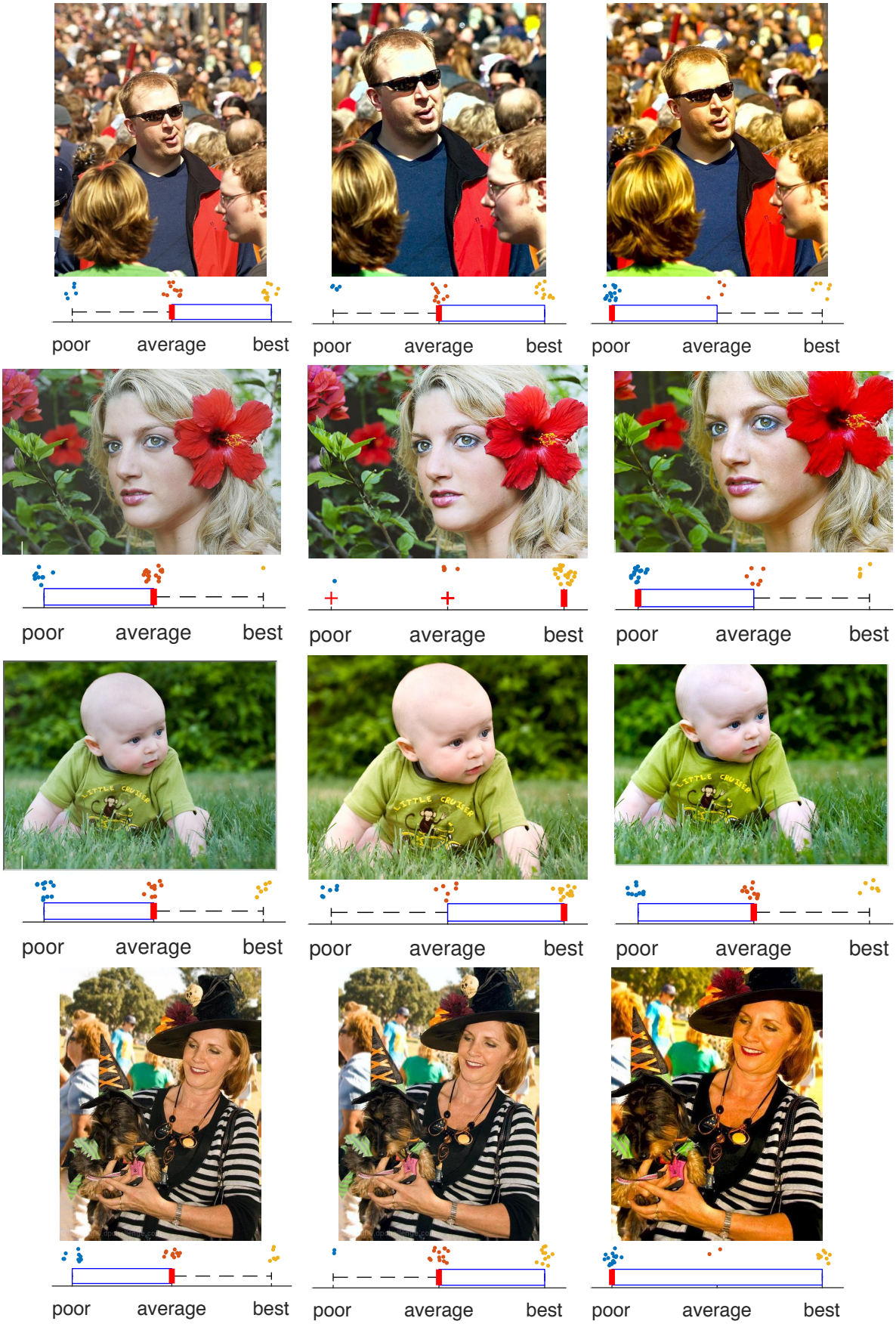

Figure 14. Visual results from our user study: (a) original image; (b) enhanced by professional; (c) enhanced by EnhanceGAN. The box plot below shows the ranking for each image, and the amount of dots denotes the number of users who gives a particular rank as in {poor, average, best}. We observe potential over-exposure in the face region in the exmaples in Row 1 and Row 4. (Best viewed in color.)