The following publication J. Liang, H. Zeng, M. Cui, X. Xie and L. Zhang, "PPR10K: A Large-Scale Portrait Photo Retouching Dataset with Human-Region Mask and Group-Level Consistency," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 653-661 is available at https://doi.org/10.1109/CVPR46437.2021.00071.

PPR10K: A Large-Scale Portrait Photo Retouching Dataset with Human-Region Mask and Group-Level Consistency

Jie Liang[1,2*], **Hui Zeng**[1,2*], **Miaomiao Cui**[2], **Xuansong Xie**[2] and **Lei Zhang**[1,2†]
[1]The HongKong Polytechnic University, [2]DAMO Academy, Alibaba Group

Abstract

Different from general photo retouching tasks, portrait photo retouching (PPR), which aims to enhance the visual quality of a collection of flat-looking portrait photos, has its special and practical requirements such as human-region priority (HRP) and group-level consistency (GLC). HRP requires that more attention should be paid to human regions, while GLC requires that a group of portrait photos should be retouched to a consistent tone. Models trained on existing general photo retouching datasets, however, can hardly meet these requirements of PPR. To facilitate the research on this high-frequency task, we construct a large-scale PPR dataset, namely PPR10K, which is the first of its kind to our best knowledge. PPR10K contains $1,681$ *groups and* $11,161$ *high-quality raw portrait photos in total. High-resolution segmentation masks of human regions are provided. Each raw photo is retouched by three experts, while they elaborately adjust each group of photos to have consistent tones. We define a set of objective measures to evaluate the performance of PPR and propose strategies to learn PPR models with good HRP and GLC performance. The constructed PPR10K dataset provides a good benchmark for studying automatic PPR methods, and experiments demonstrate that the proposed learning strategies are effective to improve the retouching performance. Datasets and codes are available: https://github.com/csjliang/PPR10K.*

1. Introduction

Portrait photography has a vast range of applications in scenarios such as wedding, birthday, graduation, anniversaries, advertisements, personal recording or creations. To ensure high quality of the finalist photos, photographers tend to capture as many raw photos with high dynamic ranges as possible. However, a set of raw photos might be flat-looking and present inconsistent tone due to the variations of subject view, illumination condition, background contexts and camera settings, as shown in the top row of Figure 1. A fast retouching on a large set of raw photos is necessary before feeding back to customers for photo selection, followed by fine-grained editing.

While a set of common standards or styles of pre-retouching are widely accepted and followed, most portrait photos are retouched manually, which is very tedious and time-consuming on the large size and highly redundant raw photos. Automatic portrait photo retouching is thus highly desired as it can save a huge amount of tedious human labor and significantly improves the efficiency of the entire portrait photography pipeline, bringing better experience for both photographers and customers.

Different from general-purpose photo retouching tasks, portrait photo retouching (PPR) has two special and practical requirements: human-region priority (HRP) and group-level consistency (GLC). HRP means that human-related region in a portrait photo should have higher priority and be paid more attention. The first row of Figure 1(a) shows a set of typical examples, where the backgrounds are over-exposed while the human regions are under-exposed. For such cases, retouching should improve the exposure of human regions while preserving as many details as possible in the backgrounds. GLC requires a group of portrait photos, which are usually taken on the same subject at the same scene but have different subject views, lighting conditions and even camera settings, to be adjusted to a consistent tone, as shown in the bottom row of Figure 1.

To the best of our knowledge, existing general-purpose photo retouching or enhancement datasets [1, 2, 9, 13] and models [3, 5, 7, 10, 12, 17, 20, 26, 29, 30, 32] do not touch the above two requirements and thus can hardly satisfy the demands of automatic PPR. To facilitate research on this important and high-frequency task, in this paper, we construct the first large-scale PPR dataset which contains $11,161$ (in $1,681$ groups) high-quality raw portrait photos, namely PPR10K dataset hereafter. The raw photos are captured by various DSLR camera devices, covering a wide range of scenes, subjects, lighting conditions and camera settings. Each raw photo is independently adjusted by 3 expert retouchers with rich experience in professional photography

[*] Equal contribution.
[†] Corresponding author. This work is supported by the Hong Kong RGC RIF grant (R5001-18).

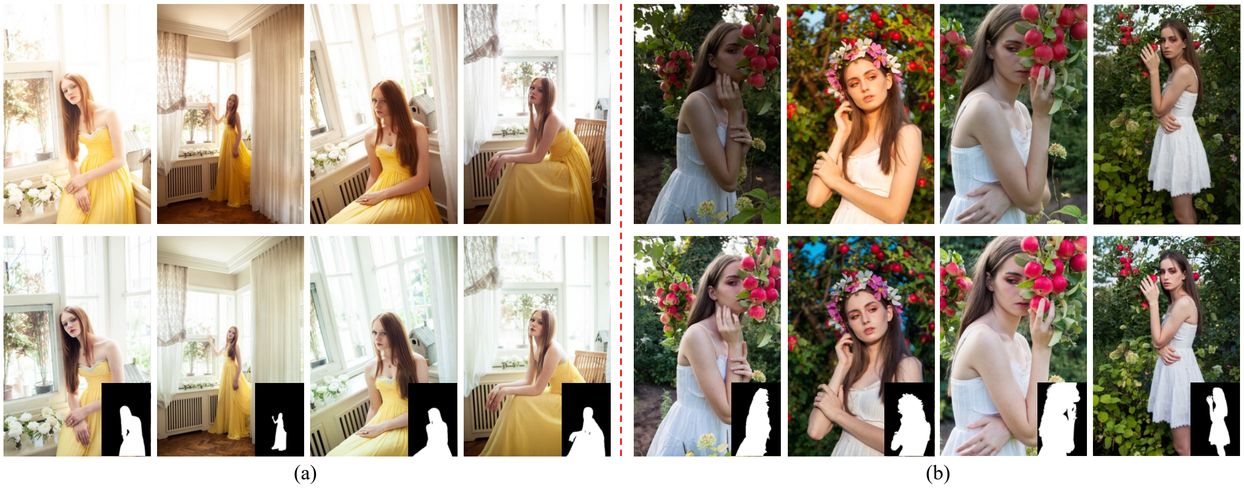


Figure 1: Examples of a group of photos from the PPR10K dataset (better viewed in color). *Top*: the raw photos; *bottom*: the retouched results from one expert and the human-region masks. The raw photos exhibit poor visual quality and large variance in subject views, background contexts, lighting conditions and camera settings. The retouched results demonstrate both good visual quality and group-level consistency.

studios, resulting in three versions of high-quality retouched targets. Besides the highly informative raw photos and their retouched results, we also provide a high-resolution human-region mask for each photo to make better use of HRP. Each group of photos is elaborately adjusted to ensure GLC. We believe this dataset will provide a valuable benchmark to facilitate the research on automatic PPR.

With the PPR10K dataset, we define a set of objective measures to evaluate the performance of automatic PPR in terms of both HRP and GLC. We also propose corresponding learning strategies to improve the retouching quality of trained PPR models. Specifically, we define human-region weighted measures based on the provided mask, which contribute to achieving better visual quality on subject areas. Explicitly defining the GLC on the image space is very challenging because of the large variations of content in a group of photos. We find that the GLC can be reliably evaluated based on the statistics in CIELAB color space. We also propose an efficient way to simulate the intra-group variations using individual images, which is proven effective to improve the GLC performance. Considering that previous general-purpose photo retouching models obtain poor performance on the PPR task, we re-implemented representative state-of-the-art photo retouching and enhancement methods, and report their performance on our dataset for a convenient and fair comparison.

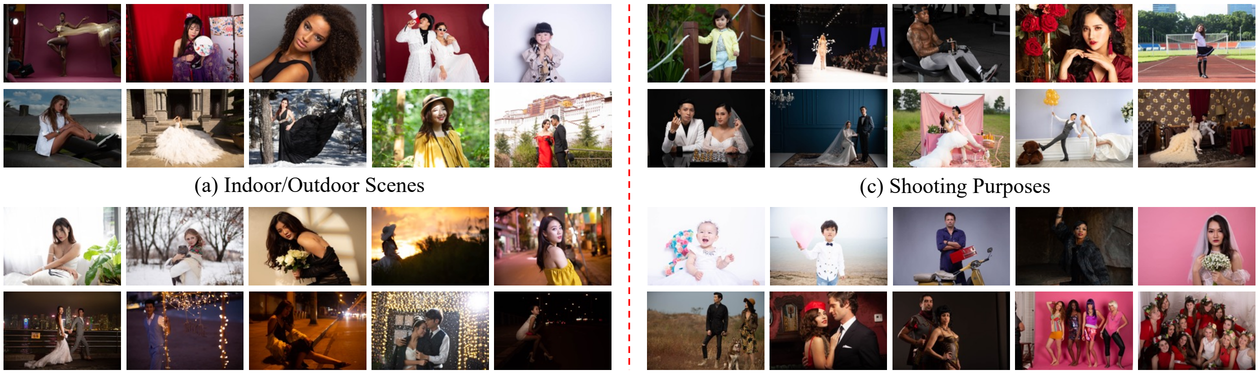
The contributions of this paper are two-fold. First, we construct the first large-scale and high-quality PPR dataset with human-region masks and group-level consistent targets, providing a valuable benchmark to facilitate research on this important task. Second, we propose a set of objective measures and learning strategies to evaluate and optimize the PPR models. Extensive experiments verified the effectiveness of the proposed dataset, measures and learning strategies both quantitatively and qualitatively.

2. Related Works

2.1. Photo Enhancement Datasets

High-quality datasets are the foundations of learning based photo enhancement or retouching research [1,2,9,13]. Bychkovsky *et al.* [1] constructed the pioneering FiveK dataset, which contains $5,000$ raw photos of general scenes together with five versions of retouched targets. This dataset has successfully facilitated the research of automatic photo retouching and enhancement [3, 7, 32]. Ignatov *et al.* [13] constructed the DPED dataset with an aim to learn a mapping from low-quality photos captured by mobile devices to the counterparts captured by high-end DSLR cameras. This dataset mainly consists of photos in general scenes such as landscapes and street views, and serves as a benchmark for general-purpose photo enhancement task. There are also datasets focusing on enhancing the dynamic range and contrast of photos [2, 9], where the ground-truths are elaborately generated via fusing multiple frames.

Despite the great efforts, the above datasets are constructed on general scenes, where portrait photos only take a minority and receive no special treatment. In addition, they only consider the visual quality of each individual photo rather than a group of photos that is commonly encountered in portrait photography. As a result, the models trained on them are unsuitable for the PPR task. In this paper, we elaborately construct a larger-scale PPR dataset, fulfilling the HRP and GLC requirements of portrait photography.



(a) Indoor/Outdoor Scenes

(c) Shooting Purposes

(b) Shooting Time/Lights

(d) Subjects

Figure 2: Visual examples to demonstrate the diversity of the proposed dataset, e.g., different indoor/outdoor scenes, light conditions, shooting purposes and human subjects.

2.2. Photo Retouching Methods

Photo retouching [2, 6, 8, 12, 16–19, 24, 31] aims to enhance the visual aesthetic quality of an image, which is conventionally achieved via professional tools, *e.g*., CameraRaw[*], or hand-craft operations like look-up tabels (LUTs) [15]. However, these manual tools rely heavily on the empirical knowledge and perceptual aesthetic judgment of the well-trained artists, therefore beyond the abilities of non-professional users. Some learning-based methods [1, 14, 21, 22, 29] based on hand-crafted features have been developed, yet can hardly satisfy the practical demands due to their limited representation capacity against the vast range of image contents and light conditions.

Various deep-learning-based schemes [3, 5, 7, 20, 25, 30] have recently been presented, benefiting from the FiveK dataset [1] and deep convolutional neural networks [11, 23]. Most of these deep models, however, are limited by the input resolutions or the processing time in practice. For real-time processing on high-resolution images, *e.g*., images with more than 24M pixels, Gharbi *et al*. [7] proposed the HDRNet putting most computation on downsampled images. He *et al*. [10] proposed to approximate a sequence of base operations such as brightness or contrast adjustments via a light-weight MLP, while Kosugi *et al*. [17] introduced a reinforcement learning framework to estimate the parameters of these operations. As a new state-of-the-art, Zeng *et al*. [32] proposed to learn an image-adaptive 3-dimensional LUT (3D LUT), which can retouch 4K images in a speed of more than 500fps with appealing tones. Nonetheless, the above-mentioned methods do not touch the HRP and GLC requirements, partially due to the lack of training data. In this paper, based on the constructed dataset, we propose two learning strategies to improve the performance of PPR, providing a benchmark for further research.

3. The PPR10K Dataset

As discussed before, existing photo retouching datasets and models cannot fulfill the requirements of PPR. To solve these problems, we construct a large-scale and high-quality PPR (PPR10K) dataset.

Challenges: To construct a valuable PPR dataset that fulfills the real-world requirements, we have to overcome several challenges. First, the photos should in raw format with high-quality. However, unlike the abundant and easily available compressed JPG images, raw photos are much more difficult to obtain on the internet. Second, the dataset should be large-scale and cover a wide range of real cases, in terms of shooting purpose, human subjects, background scenes, lighting conditions as well as the usage of camera devices, which further increases the cost of data collection. Third, high-quality retouched results (with both good visual quality and group-level consistency) and human-region masks should be provided to learn effective PPR models. These requirements make the labeling process expensive and cumbersome.

Data Collection and Selection: To obtain as many raw portrait photos as possible, we negotiated with many individual photographers and professional photography studios to purchase raw portrait photos in groups from them for research purposes only. We also purchased from several paid material websites that provide raw format portrait photos. During data collection, we have elaborately control the diversity of raw photos in terms of shooting purpose (e.g., wedding, birthday, graduation, anniversaries, advertisements, personal recording and creation), human subjects (including babies, children, younger, couples and worldwide people), background scenes (including indoor and outdoor, lighting conditions (from day to night, winter to summer), and usage of camera devices (covering a wide series of high-end DSLR cameras of Canon, Nikon and Sony).

[*] https://www.adobe.io/apis/creativecloud/camera-raw.html

The diversity of collected photos is shown in Figure 2.

We initially collected more than $25,000$ raw photos then conducted several rounds of selection. We first discarded photos without human subjects, with low quality such as severe motion blur or de-focus, or containing inappropriate information. We further carefully checked photos group by group, removing outliers (photos with very different content from the group) and duplicated ones (photos with almost the same content). After the screening, we finally obtained a total of $11,161$ portrait photos in $1,681$ groups, and each group contains $3 \sim 18$ photos with same the subjects captured in the same scene at consecutive time. Two typical groups of photos are shown in Figure 1.

Data Labeling: To obtain high-quality ground-truths, we hired 3 expert retouchers, all of whom have more than 5 years of experience working in the professional photographic industry, to retouch the raw photos independently, using the CameraRaw in PhotoShop. Each retoucher was required, based on their own domain knowledge, to retouch the raw photos to satisfy the output standards of professional portrait photography studios with two major requirements. First, each photo should be retouched to visually pleasing to the perception of common people, especially for the human-regions. Second, a group of photos should be adjusted to have a consistent tone. Retouchers were allowed to adjust any operations in CameraRaw without changing content or introducing geometric distortions. In addition, the retouching of each expert was also required to be self-consistent among similar scenes, which is important for learning a stable and robust retouching model. We also hired another expert to double-check the retouched results and conducted several rounds of feedback-and-repair to ensure high-quality of the ground-truths. The retouching style of the three experts is shown in **supplementary file**.

Considering the high priority of human regions and their complicated illuminations in portrait photos, we also provide human-region masks for learning better retouching models. To save the annotation cost, the masks were first generated using an internal-developed portrait segmentation algorithm which was trained on a set of human matting datasets and supports segmenting photos up to 100 megapixels. We then manually check and refine the failure cases on some difficult scenes such as underwater, extremely low light, glass-reflex and occluded cases.

Discussions: Despite the high-quality of our constructed dataset, it leaves several challenges in learning an effective portrait retouching model. First, both the raw photos and human-region masks have very high resolution ranging from 4K to 8K, which requires retouching models to be highly efficient. Second, the diversity of content and lighting conditions in various scenes requires the models to be flexible and content-adaptive. Third, the demand of group-level consistency requires the models to be robust and stable, which is critical for practical applications.

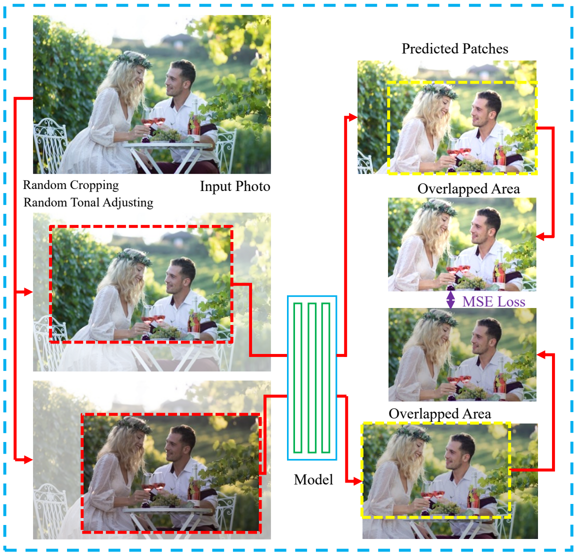


Figure 3: Illustration of the GLC learning strategy.

4. Measures and Learning Strategies

Based on the PPR10K dataset, we define a set of measures to quantitatively evaluate the performance of a PPR method. We also propose learning strategies to optimize the HRP and GLC requirements of the PPR task.

4.1. Basic Measures

Following the common practice in general-purpose photo enhancement task [1, 7, 32], we first define two basic measures, including the peak signal-to-noise ratio (PSNR) and the CIELAB color difference [28]. Given an input portrait photo $\boldsymbol{I}$, denote by $\hat{\boldsymbol{I}}$ and $\boldsymbol{Y}$ its predicted version of a PPR model and the target retouched by a human expert, respectively. We can easily obtain their conversions in Lab color space, which are denoted by $\boldsymbol{I}^{Lab}$, $\hat{\boldsymbol{I}}^{Lab}$ and $\boldsymbol{Y}^{Lab}$. Similar to the PSNR defined based on the L_2-distance in sRGB color space, the color difference is defined as the L_2-distance in CIELAB color space with $\triangle E_{ab} = \|\hat{\boldsymbol{I}}^{Lab} - \boldsymbol{Y}^{Lab}\|_2$. Compared to the sRGB color space, the CIELAB color space is more perceptually uniform and is widely used to tune the tones of photos [27].

4.2. Human-centered Measures

Considering the higher priority of human regions in portrait photos, we further define two human-centered measures, which can be naturally achieved by putting higher weights to human regions than background regions, leveraging the provided human-region masks in the PPR10K dataset. Given a photo $\boldsymbol{I}$ of $H \times W$ resolution, we can construct its weighting matrix $\boldsymbol{W_I} = [w_{i,j}] \in R^{H \times W}$,

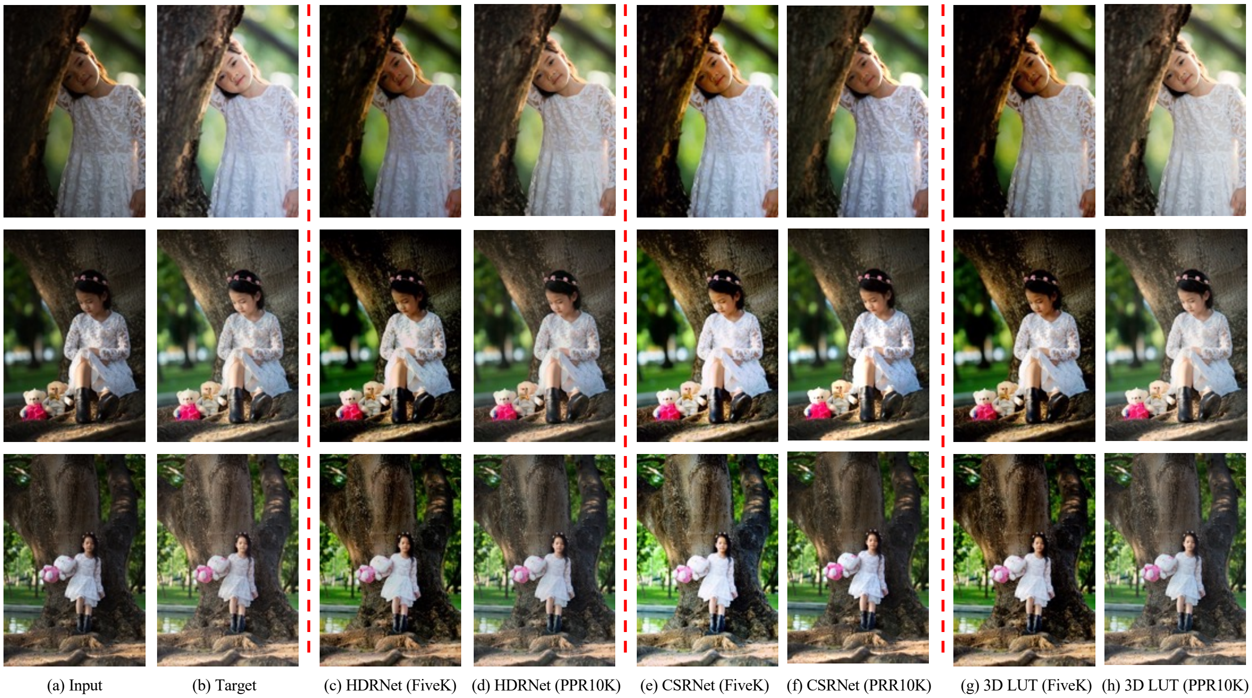


Figure 4: Visual comparisons between models trained on the FiveK (c, e, g) dataset and the proposed PPR10K (d, f, h) dataset. The targets in (b) are from PPR10K-a.

where $w_{ij} = 1$ for background regions human regions and $w_{ij} = \alpha$ $(\alpha \leq 1)$ for background regions. We empirically fix $\alpha = 0.5$. The human-centered PSNR ($PSNR^{HC}$) and color difference ($\triangle E_{ab}^{HC}$) can be consequently defined. To save space, we only provide the formula of $\triangle E_{ab}^{HC}$ as:

$$\triangle E_{ab}^{HC} = \|\boldsymbol{W_I} \odot \hat{\boldsymbol{I}}^{Lab} - \boldsymbol{W_I} \odot \boldsymbol{Y}^{Lab}\|_2. \qquad (1)$$

where $\odot$ denotes the element-wise matrix multiplication.

4.3. Group-level Consistency Measure

Different from the above measures based on individual photos, the group-level consistency (GLC) measures the variations in tone and color among a group of photos. This measure can hardly be defined in the image space since the image contents in a group of photos are not aligned (refer to Figure 1). A reasonable GLC measure should be sensitive to the change of global tone and color appearance, and simultaneously should be robust to the change of image content. It is worth mentioning that the content change is restricted in a group of photos that have the same subject and similar background. Inspired by the practice in white balance [6], we define the GLC measure based on the statistics of color components.

Specifically, given a group of predicted photos $[\hat{\boldsymbol{I}}_1, \hat{\boldsymbol{I}}_2, \cdots, \hat{\boldsymbol{I}}_m]$, we first calculate the mean color components of each photo $[\mu_{\hat{\boldsymbol{I}}_1}, \mu_{\hat{\boldsymbol{I}}_2}, \cdots, \mu_{\hat{\boldsymbol{I}}_m}]$ to represent their global tone and color appearance. The GLC measure is then defined as the variance of mean color components:

$$\mathcal{M}_{GLC} = \sum_c Var(\mu_{\hat{\boldsymbol{I}}_1^c}, \mu_{\hat{\boldsymbol{I}}_2^c}, \cdots, \mu_{\hat{\boldsymbol{I}}_m^c}), \qquad (2)$$

where c denotes a color channel which can be chosen from $\{R, G, B, L, a, b\}$ or the combinations of them. Through extensive quantitative studies, we empirically found that the GLC measure based on the combination of a and b channels is the most suitable and stable choice. The study details can be found in the **supplementary file**.

4.4. Learning Strategies

Optimizing the basic measures and human-centered measures is straightforward. We simply employ the human-region weighted MSE loss on sRGB color images to optimize a model with these measures:

$$\mathcal{L}_{HC} = \|\boldsymbol{W_I} \odot \hat{\boldsymbol{I}} - \boldsymbol{W_I} \odot \boldsymbol{Y}\|_2^2, \qquad (3)$$

where we set $\alpha = 1.0$ in $\boldsymbol{W_I}$ for the basic measures. For the human-centered measures, we set $w_{ij} = 1$ for backgrounds and $w_{ij} = 5$ for human-regions to accelerating training.

Explicitly optimizing the GLC measure is complicated since it introduces much additional cost including reading and processing a group of photos and color space conversion. To simplify and speed up the training process, we in-

Table 1: Quantitative comparisons among the baseline and ablation methods. The PPR10K-a/b/c indicate the GTs retouched by three experts. For each measure, the 'LR' column reports the results tested on 360p images and the 'HR' column reports the results tested on original resolutions. The ↑ and ↓ denote that larger or smaller is better, respectively.

#	Method	Dataset	$PSNR\uparrow$		$\triangle E_{ab}\downarrow$		$PSNR^{HC}\uparrow$		$\triangle E_{ab}^{HC}\downarrow$		$\mathcal{M}_{GLC}\downarrow$	
			LR	HR	LR	HR	LR	HR	LR	HR	LR	HR
1	HDRNet [7]	PPR10K-a	23.93	23.06	8.70	9.13	27.21	26.58	5.65	5.84	14.83	14.37
2	CSRNet [10]	PPR10K-a	22.72	22.01	9.75	10.20	25.90	25.19	6.33	6.73	12.73	12.66
3	3D LUT [32]	PPR10K-a	25.64	25.15	6.97	7.25	**28.89**	28.39	4.53	4.71	11.47	11.05
4	3D LUT+HRP	PPR10K-a	**25.99**	**25.55**	**6.76**	**7.02**	28.29	**28.83**	**4.38**	**4.55**	10.81	10.32
5	3D LUT+GLC	PPR10K-a	25.06	24.39	7.39	7.81	28.34	27.67	4.80	5.06	9.98	9.77
6	3D LUT+HRP+GLC	PPR10K-a	25.31	24.60	7.30	7.75	28.56	27.86	4.75	5.03	**9.95**	**9.68**
7	HDRNet [7]	PPR10K-b	23.96	23.51	8.84	9.13	27.21	26.55	5.74	5.92	13.21	13.04
8	CSRNet [10]	PPR10K-b	23.76	23.29	8.77	9.28	27.01	26.62	5.68	5.90	11.82	11.73
9	3D LUT [32]	PPR10K-b	24.70	24.30	7.71	7.97	27.99	27.59	4.99	5.16	9.90	9.52
10	3D LUT+HRP	PPR10K-b	**25.06**	**24.66**	**7.51**	**7.73**	**28.36**	**27.93**	**4.85**	**5.00**	9.87	9.60
11	3D LUT+GLC	PPR10K-b	24.16	23.39	8.15	8.70	27.48	26.71	5.25	5.61	9.17	8.92
12	3D LUT+HRP+GLC	PPR10K-b	24.52	23.81	7.93	8.42	27.82	27.12	5.12	5.44	**9.01**	**8.73**
13	HDRNet [7]	PPR10K-c	24.08	23.66	8.87	9.05	27.32	26.93	5.76	5.99	14.76	14.28
14	CSRNet [10]	PPR10K-c	23.17	22.85	9.45	9.87	26.47	26.09	6.12	6.54	14.64	14.22
15	3D LUT [32]	PPR10K-c	25.18	24.78	7.58	7.85	28.49	28.09	4.92	5.09	13.51	13.16
16	3D LUT+HRP	PPR10K-c	**25.46**	**25.05**	**7.43**	**7.69**	**28.80**	**28.38**	**4.82**	**4.98**	13.49	13.06
17	3D LUT+GLC	PPR10K-c	24.53	23.94	8.10	8.49	27.87	27.29	5.25	5.49	12.96	**12.75**
18	3D LUT+HRP+GLC	PPR10K-c	24.59	24.01	8.02	8.39	27.92	27.33	5.20	5.43	**12.76**	12.79

Table 2: Quantitative results of models trained on the FiveK dataset and evaluated on the PPR10K dataset, where a, b, and c denote the GTs retouched by three experts.

Method	Dataset	$PSNR$	$\triangle E_{ab}$	$PSNR^{HC}$	$\triangle E_{ab}^{HC}$	$\mathcal{M}_{GLC}$
HDRNet	a	18.20	17.22	21.44	11.27	20.76
CSRNet	a	19.86	14.07	23.06	9.15	13.97
3D LUT	a	19.92	13.75	23.79	8.90	13.85
HDRNet	b	18.74	16.31	22.00	10.63	20.76
CSRNet	b	19.65	14.47	22.83	9.40	13.97
3D LUT	b	19.74	14.08	23.42	9.22	13.85
HDRNet	c	19.71	14.81	22.96	9.65	20.76
CSRNet	c	19.81	14.57	23.06	9.46	13.97
3D LUT	c	20.03	13.90	23.01	8.85	13.85

troduce a strategy to simulate the group-level variation using a single image, the pipeline of which is shown in Figure 3. Specifically, given an input $\boldsymbol{I}$, we randomly crop two overlapped patches $\boldsymbol{I}_{C_1}$ and $\boldsymbol{I}_{C_2}$ to mimic the view change in a group of photos. We further randomly adjust the two crops regarding tonal attributes such as temperatures and exposures to synthesize the change of lighting condition and camera setting. We feed the two synthesized crops into a PPR model, obtaining two predictions $\hat{\boldsymbol{I}}_{C_1}$ and $\hat{\boldsymbol{I}}_{C_2}$ and their overlapped ranges $\hat{\boldsymbol{I}}^O_{C_1}$ and $\hat{\boldsymbol{I}}^O_{C_2}$. The GLC can be optimized using the following constraint:

$$\mathcal{L}_{GLC} = \|\hat{\boldsymbol{I}}^O_{C_1} - \hat{\boldsymbol{I}}^O_{C_2}\|_2^2. \tag{4}$$

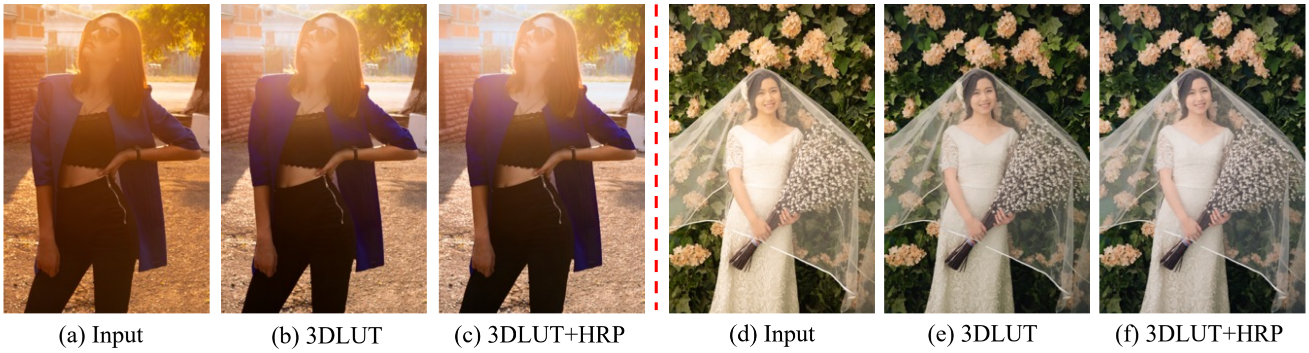
The total loss is calculated as $\mathcal{L} = \mathcal{L}_{HC} + \lambda\mathcal{L}_{GLC}$, where λ is a constant parameter to balance the two losses. We simply set $\lambda = 1$ in the experiments.

5. Experiments

5.1. Experiment settings

Datasets: We employed two datasets, including the constructed PPR10K dataset and the general-purpose FiveK [1] dataset, in our experiments. The PPR10K dataset is randomly divided into a training set with $1,356$ groups and $8,875$ photos, and a testing set with 325 groups and $2,286$ photos. The FiveK dataset is randomly divided into a training set with 4,500 images and a validation set with 500 images following the common practice. Input images are pre-processed in a 16-bit *tiff* format via CameraRaw to preserve as much information from the raw file as possible, while the target images are converted into 8-bit sRGB color space for convenient display on common devices. To speed up the training process, training images are resized to 360p (short side of the images) resolution. The testing images have two versions: the 360p resolution and the original resolution ranging from 4K to 8K.

Baseline Methods: Since in practice PPR needs to process very high-resolution photos, this hinders the real applications of most previous photo retouching/enhancement models because of their heavy computational and memory costs. We employ three competitive and efficient models, including the HDRNet [7], the CSRNet [10] and the 3D LUT [32], in our experiments (source codes released by



(a) Input (b) 3DLUT (c) 3DLUT+HRP (d) Input (e) 3DLUT (f) 3DLUT+HRP

Figure 5: Visual evaluation of the HRP learning strategy on example photos by using 3D LUT [32]. Leveraging the HRP loss leads to brighter faces and more natural temperature on both examples.

authors). To better model such a large-scale and diverse dataset, for the 3D LUT [32] mothod, we employ 5 LUTs and use the Resnet-18 [11] (initialized with the weights pretrained on ImageNet [4]) as the scene classifier.

Data Augmentation: Besides the commonly used data augmentation methods such as flipping and rotation, we also augment the training images by adjusting 6 visual attributes in CameraRaw, *i.e.*, temperature, tint, exposure, highlights, contrasts and saturation, to enrich the lighting and color distributions of the training set. The augmentation details can be found in the **supplementary material**.

5.2. Baseline Performance

We first evaluate the baseline performance of the three state-of-the-art photo retouching/enhancement methods on our PPR10K dataset. We retrained each model on each of the three expert sets independently and report their performance under five measures ($PSNR$, $\triangle E_{ab}$, $PSNR^{HC}$, $\triangle E_{ab}^{HC}$, $\mathcal{M}_{GLC}$) in Table 1 (rows 1-3, 7-9, 13-15). Each measure is evaluated on two resolutions (360p low resolution (LR) and original high resolution (HR)). Several observations can be made from the results.

First, all the three models can obtain reasonable results on $PSNR$ and $\triangle E_{ab}$, which indicates the high-quality and self-consistent annotations of the three experts. Among the three versions, the retouching style of expert-a is relatively easier to be learned as expected, since this expert prefers rendering a stronger and stable tonal style for all scenes, leading to a compact target space which is relatively easier to be modeled. In contrast, the other two experts prefer a mild rendition to preserve the naturalness of photos (visual examples are provided in the **supplementary file**). Among the three models, 3D LUT [32] achieves consistently better performance in most cases than HDRNet [7] and CSRNet [10]. Given its high performance and high efficiency, we choose 3D LUT as the baseline model to study the proposed learning strategies in Section 5.4.

5.3. Models Trained on FiveK and PPR10K

This section compares the PPR performance of the three methods by training them on the FiveK dataset and on our PPR10K dataset, respectively. We employed the commonly used expert C as the target on the FiveK dataset to train the three models. Input images were processed to have the same format as in our PPR10K dataset. We evaluated the trained models on the three testing sets of PPR10K and report the quantitative results in Table 2. Qualitative comparisons are shown in Figure 4. As expected, all models trained on the FiveK dataset obtain much worse performance on all measures compared to their counterparts trained on the PPR10K dataset (refer to Table 1), because of the domain gap between general-purpose photo enhancement and PPR. As shown in Figure 4, the results obtained by the FiveK models have two obvious problems. First, the tone and color appearance of each individual photo is unpleasing especially on the human regions. Specifically, the girl's face is dark in the shadow with unnatural color. Second, the retouched results in a group have large variations on both global tone and local contrast. For example, the third photo has obviously higher brightness and more natural color compared to the first one. In contrast, models trained on our PPR10K dataset achieve not only better individual visual quality but also higher group-level consistency.

5.4. Effectiveness of the Learning Strategies

This section evaluates the effectiveness of the proposed learning strategies using the 3D LUT model. On each of the PPR10K set, we trained three 3D LUT models using only HRP, only GLC, both HRP and GLC learning strategies and report the results in Table 1 (rows 4-6, 10-12, 16-18).

One can see that using the HRP loss brings better results on most individual measures. This is reasonable since all three experts paid special attention to the human regions during their retouching. Putting higher weights on human regions thus leads to better individual retouching quality.

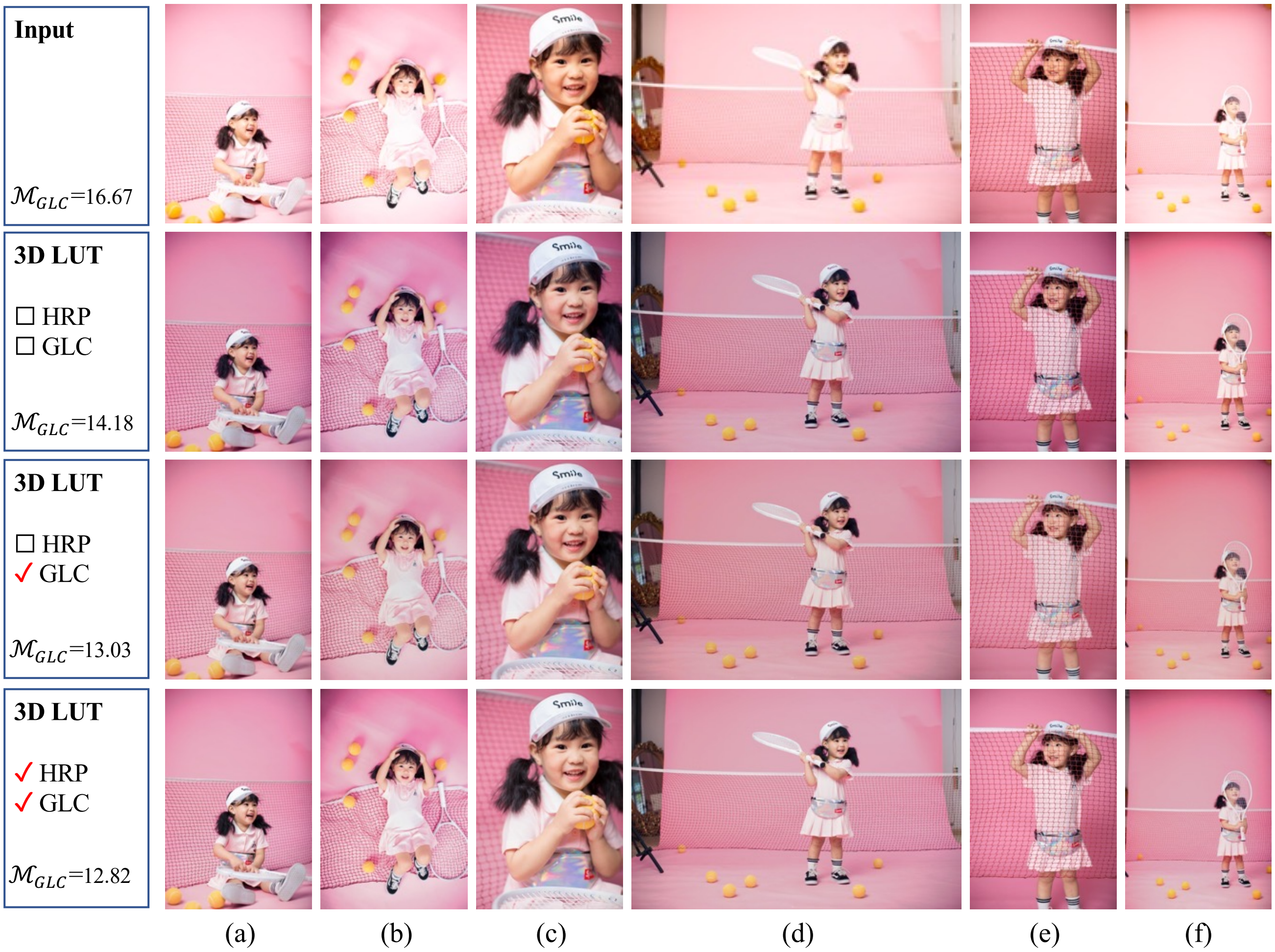

Input
$\mathcal{M}_{GLC}$=16.67
3D LUT
□ HRP
□ GLC
$\mathcal{M}_{GLC}$=14.18
3D LUT
□ HRP
✓ GLC
$\mathcal{M}_{GLC}$=13.03
3D LUT
✓ HRP
✓ GLC
$\mathcal{M}_{GLC}$=12.82
(a)
(b)
(c)
(d)
(e)
(f)

Figure 6: Visual evaluation of the GLC and GLC+HRP learning strategy by using 3D LUT. From *top* to *bottom*: the inputs, results of baseline 3D LUT, results of [3D LUT+GLC] and results of [3D LUT+GLC+HRP].

Two typical visual examples are shown in Figure 5. One can see that using the HRP loss leads to better visual quality (brighter faces and more natural temperature on both examples) on the human regions.

Using the GLC loss slightly deteriorates the four individual measures but improves the GLC measure. A qualitative example of learning with the GLC loss is shown in Figure 6. As shown in the figure, compared to the results obtained by baseline 3D LUT, the color of the background tends to be more consistent when GLC loss is employed. Specifically, the color of curtain in Figure 6 (b, d, f) varies in baseline 3D LUT, while being a consistent pink when the GLC loss is employed. Another observation is that combining the GLC and HRP losses further improves the GLC measure. This is possibly because jointly optimizing the HRP and GLC losses enables the model to learn complementary information and consequently achieves a good trade-off between individual visual quality and group-level consistency.

6. Conclusion

We constructed a large-scale PPR dataset, which was the first of its kind to the best of our knowledge. We collected high quality raw portrait photos with diverse contents from individual photographers and professional photography studios. After careful screening, $11,161$ portrait photos were selected, which fell into $1,681$ groups. High quality human region masks were provided in the dataset. We invited three expert retouchers to label the photos with priority to the human region and the tonal consistency within a group of photos. We defined a set of human-region centered and group-level consistency measures to faithfully evaluate the performance of a PPR model, and accordingly proposed learning strategies to train high quality PPR models. Extensive experiments were conducted to demonstrate the value of the constructed dataset, and the effectiveness of the proposed measures and learning strategies.

References

[1] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, 2011. 1, 2, 3, 4, 6

[2] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018. 1, 2, 3

[3] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs. In *CVPR*, 2018. 1, 2, 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7

[5] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *ACM Multimedia*, 2018. 1, 3

[6] Graham D Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. In *Color and Imaging Conference*, volume 2004, pages 37–41. Society for Imaging Science and Technology, 2004. 3, 5

[7] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics*, 36(4):118, 2017. 1, 2, 3, 4, 6, 7

[8] Arjan Gijsenij, Theo Gevers, and Joost Van De Weijer. Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, 20(9):2475–2489, 2011. 3

[9] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 35(6), 2016. 1, 2

[10] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *ECCV*, 2020. 1, 3, 6, 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 7

[12] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2):1–17, 2018. 1, 3

[13] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. DSLR-quality photos on mobile devices with deep convolutional networks. In *ICCV*, 2017. 1, 2

[14] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. Properties and performance of a center/surround retinex. *IEEE Transactions on Image Processing*, 6(3):451–462, 1997. 3

[15] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *ECCV*, 2016. 3

[16] Yeong-Taeg Kim. Contrast enhancement using brightness preserving bi-histogram equalization. *IEEE transactions on Consumer Electronics*, 43(1):1–8, 1997. 3

[17] Satoshi Kosugi and Toshihiko Yamasaki. Unpaired image enhancement featuring reinforcement-learning-controlled image editing software. In *AAAI*, 2020. 1, 3

[18] Rafał Mantiuk, Scott Daly, and Louis Kerofsky. Display adaptive tone mapping. In *ACM SIGGRAPH 2008 papers*, pages 1–10. 2008. 3

[19] Jayanta Mukherjee and Sanjit K Mitra. Enhancement of color images by scaling the dct coefficients. *IEEE Transactions on Image processing*, 17(10):1783–1794, 2008. 3

[20] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *CVPR*, 2018. 1, 3

[21] Zia-ur Rahman, Daniel J Jobson, and Glenn A Woodell. Multi-scale retinex for color image enhancement. In *ICIP*, 1996. 3

[22] Debashis Sen and Sankar K Pal. Automatic exact histogram specification for contrast enhancement and visual system based quantitative evaluation. *IEEE Transactions on Image Processing*, 20(5):1211–1220, 2011. 3

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[24] Yoshiharu Takahashi, Yasutoshi Fujii, Keiichi Kawazu, Seiichi Isoguchi, Kohichi Yamaguchi, and Katsuya Nagaishi. Photographic film reproducing apparatus using object brightness and exposure correction amount to develop photographed images, May 5 1998. US Patent 5,748,287. 3

[25] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, 2019. 3

[26] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, 22(9):3538–3548, 2013. 1

[27] Wikipedia contributors. CIELAB color space — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=CIELAB_color_space&oldid=982533770, 2020. 4

[28] Wikipedia contributors. Color difference — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Color_difference&oldid=979429323, 2020. 4

[29] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. A learning-to-rank approach for image color enhancement. In *CVPR*, 2014. 1, 3

[30] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM ToG*, 35(2):11, 2016. 1, 3

[31] Lu Yuan and Jian Sun. Automatic exposure correction of consumer photographs. In *ECCV*, 2012. 3

[32] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 3, 4, 6, 7