Supplementary of "RSFNet: A White-Box Image Retouching Approach using Region-Specific Color Filters"

Wenqi Ouyang[1], Yi Dong[2], Xiaoyang Kang[1], Peiran Ren[1], Xin Xu[1], Xuansong Xie[1]
[1]DAMO Academy, Alibaba Group, [2]Nanyang Technological University
wenqi.oywq@alibaba-inc.com, ydong004@ntu.edu.sg,
{peiran.rpr, kangxiaoyang.kxy, chris.xx, xingtong.xxs}@alibaba-inc.com

1. Filter Function

The retouched result is represented as equation:

$$Y = X + \sum_m \sum_n (F_{m,n}(\theta_{m,n}, X) - X) \odot M_m, \tag{1}$$

For different filters, $F_{m,n}(\theta_{m,n}, X)$ has different expressions:

$$\begin{aligned}
F_{contrast}(\theta, X) - X &= \theta(X - mean(X)) \\
F_{saturation}(\theta, X) - X &= \theta(X - L(X)) \\
F_{hue}(\theta, X_c) - X_c &= \begin{cases} \alpha_h \theta X_c, X_c \in \{X_R, X_G\} \\ -\frac{1}{2}\alpha_h \theta X_c, X_c = X_B \end{cases} \\
F_{temperature}(\theta, X_c) - X_c &= \begin{cases} \alpha_{t,1}\theta X_c, X_c = X_R, \theta \geq 0 \\ \alpha_{t,2}\theta X_c, X_c = X_R, \theta < 0 \\ 0.0, X_c = X_G, \theta \geq 0 \\ \alpha_{t,3}\theta X_c, X_c = X_G, \theta < 0 \\ \alpha_{t,4}\theta X_c, X_c = X_B, \theta \geq 0 \\ \alpha_{t,5}\theta X_c, X_c = X_B, \theta < 0 \end{cases} \\
F_{shadows}(\theta, X) - X &= \theta(1 - X) \\
F_{midtones}(\theta, X) - X &= \theta(0.25 - (X - 0.5)^2) \\
F_{highlights}(\theta, X) - X &= \theta X \\
F_{shift}(\theta, X_c) - X_c &= \begin{cases} \alpha_{s,1}, X_c = X_R \\ \alpha_{s,2}, X_c = X_G \\ \alpha_{s,3}, X_c = X_B \end{cases}
\end{aligned} \tag{2}$$

Where $mean(X)$ denotes the mean value of the entire image, while $L(X)$ represents the L channel of the image in the CIE LAB color space. Additionally, $X_c \in X_R, X_G, X_B$ denotes the RGB color channels of the image. The adjustment factor, denoted by α, is a scalar that satisfies $\alpha > 0$. In our experiments, values of α are determined according to traditional color grading tools.

2. Variation of RSFNet with Controlled Region Shape

Ground Truths Mask Generation. We adopt the palette-based method proposed in [1] to generate the main colors $C = C_1, ..., C_n$ of an image, along with distance maps from pixels to N color centers. We obtain region masks by applying a

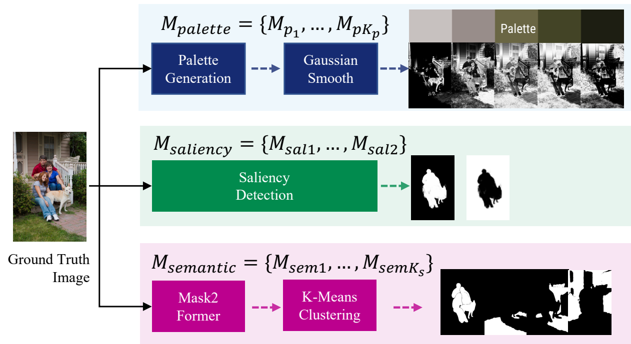

Figure 1: Ground thuths masks generation process.

Method	360p expert a		360p expert b		360p expert c	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DeepLPF [8]	24.97	0.939	24.33	0.930	24.65	0.926
CSRNet [3]	24.38	0.938	24.41	0.940	24.53	0.931
3D-LUT+AdaInt [11]	27.31	0.954	26.62	0.945	26.67	0.929
Harmonizer [4]	25.02	0.916	23.84	0.895	25.22	0.920
RSFNet-map	27.25	0.956	26.61	0.954	26.76	0.945
RSFNet-saliency	25.98	0.946	25.87	0.948	25.88	0.937

Table 1: Quantitative comparisons for retouching tasks on PPR10K dataset [5]. All the models are trained on data without augmentations.White-box methods are colored as violet.

Gaussian smoothing function to these distance maps, resulting in $M_{palette} = M_{p1}, \ldots M_{pK_p}$. We also predict saliency masks using pre-trained networks from [6], yielding $M_{saliency} = M_{sal1}, M_{sal2}$.

Since previous works on panoptic segmentation split objects into more than one hundred classes, which is redundant for our task, we aim to identify the most significant pixel groupings. To accomplish this, we follow the practice in [7, 10] and train a self-attentioned network with pairwise retouching data. First, we predict semantic masks using the networks presented in [2]. We then apply a clustering algorithm (e.g., K-means [9]) to the output features of the self-attentioned network with masked images as input. Masks assigned with the same cluster index are merged, resulting in $M_{semantics} = M_{sem1}, \ldots M_{semK_s}$. For each of the three sets of masks, we train a separate model with the corresponding output channel numbers. The entire process is illustrated in Figure 1.

Differentiable Adaptive Smooth Kernel. To ensure smooth transition across mask edges, we have incorporated a differentiable adaptive smooth kernel module into our main network. We fix the Gaussian smooth kernel to a suitable size σ_{max}, such as 51×51 for the original input with a resolution of 256×256. The standard variance of the kernel is a learnable parameter, which adapts to the inputs.

3. Additional Results

Quantitative Results. We valuate our methods using a random split setup for PPR10K [5]. The results are demonstrated in Table 1. For more implementation details, please refer to our codebase at `https://github.com/Vicky0522/RSFNet`.

Qualitative Results. We present more results of RSFNet-saliency, RSFNet-palette and RSFNet-map in Figure 2 and 3, including generated masks and corresponding filter arguments.

Figure 2: Editable white-box retouching. Arguments and masks generated by RSFNet-saliency trained with saliency masks are shown in the second column. Retouched result is shown in the third column. Three versions of adjustments conducted on the retouched results are shown in the three columns on the right. Ground truths is shown in the right-most column. Numbers in green boxes indicate relative variation.

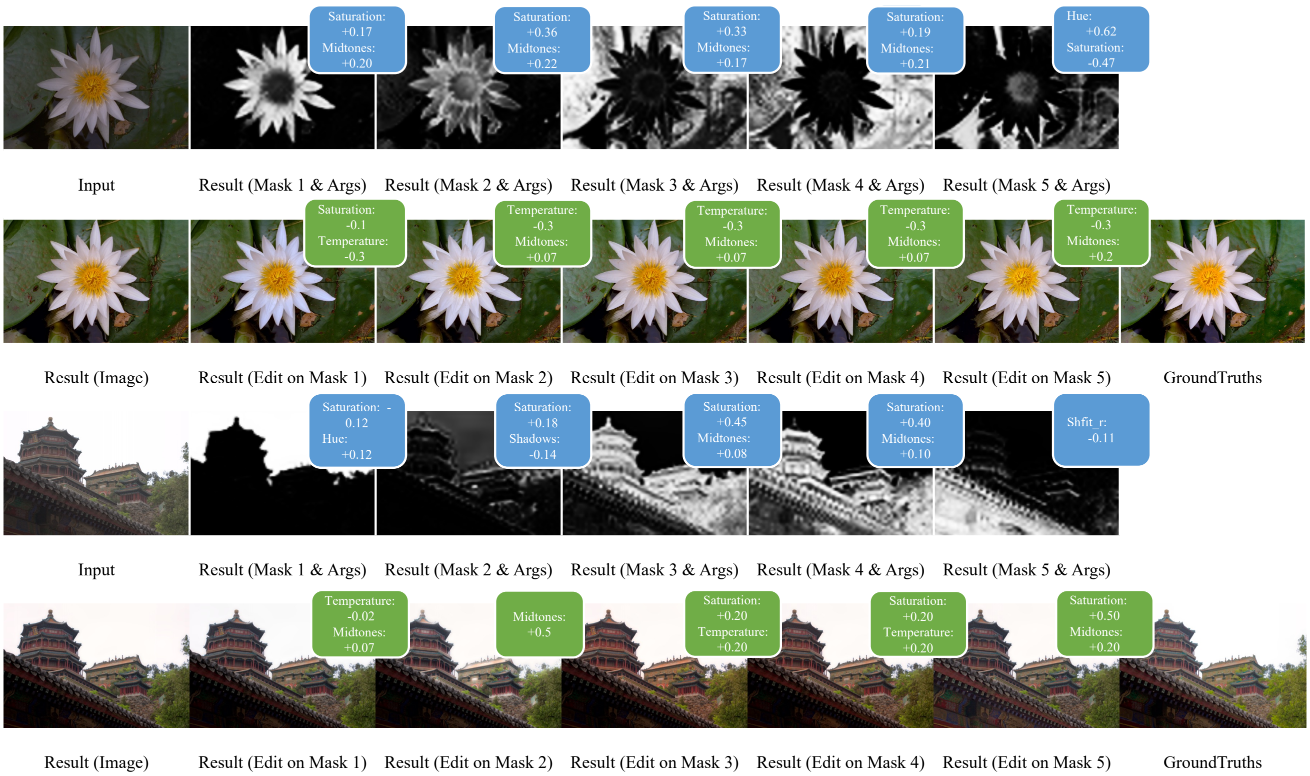

Figure 3: Editable white-box retouching. Arguments and masks generated by RSFNet-palette trained with palette-based masks are shown in the first row. Only two of the most significant arguments are presented. Retouched result is shown in the first column of the second row. Five versions of adjustments conducted on the retouched results and corresponding masks are shown in the rest columns of the second row. Ground truths is shown in the right-most column. Numbers in green boxes indicate relative variation.

References

[1] Huiwen Chang, Ohad Fried, Yiming Liu, Stephen DiVerdi, and Adam Finkelstein. Palette-based photo recoloring. *SIGGRAPH*, 34(4), July 2015. 1

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2021. 2

[3] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *ECCV*, pages 679–695, 2020. 2

[4] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson W.H. Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *ECCV*, pages 690–706, 2022. 2

[5] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *CVPR*, pages 653–661, 2021. 2

[6] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, pages 3917–3926, 2019. 2

[7] Caron Mathilde, Touvron Hugo, Misra Ishan, Jégou Hervé, Mairal Julien, Bojanowski Piotr, and Joulin Armand. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 3917–3926, 2021. 2

[8] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory G. Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *CVPR*, pages 12823–12832, 2020. 2

[9] Lloyd Stuart. Least squares quantization in pcm. *IEEE TIT*, 28(2):129–137, 1982. 2

[10] Van Gansbeke Wouter, Vandenhende Simon, and Van Gool Luc. Discovering object masks with transformers for unsupervised semantic segmentation. *arXiv preprint arXiv:2206.06363*, 2022. 2

[11] Canqian Yang, Meiguang Jin, Xu Jia, Yi Xu, and Ying Chen. Adaint: Learning adaptive intervals for 3d lookup tables on real-time image enhancement. In *CVPR*, pages 17501–17510, 2022. 2