Photo Retouching
Image adaptive 3D LUT, CSRNet

Presenter: Minho Park

MIT-Adobe FiveK Dataset

- Automatic global adjustment using supervised machine learning

Figure 1. On this photo, the retouchers have produced diverse of outputs, from a sunset mood (b) to a day light look (f). There is no single good answer and the retoucher's interpretation plays a significant role in the final result. We argue that supervised machine learning is well suited to deal with the difficult task of automatic photo adjustment, and we provide a dataset of reference images that enables this approach. This figure may be better viewed in the electronic version.

Learning Photographic Global Tonal Adjustment with a Database of Input / Output Image Pairs, V. Bychkovsky, et al., CVPR 2011, https://ieeexplore.ieee.org/document/5995332

DeepUPE

- Predict illumination map

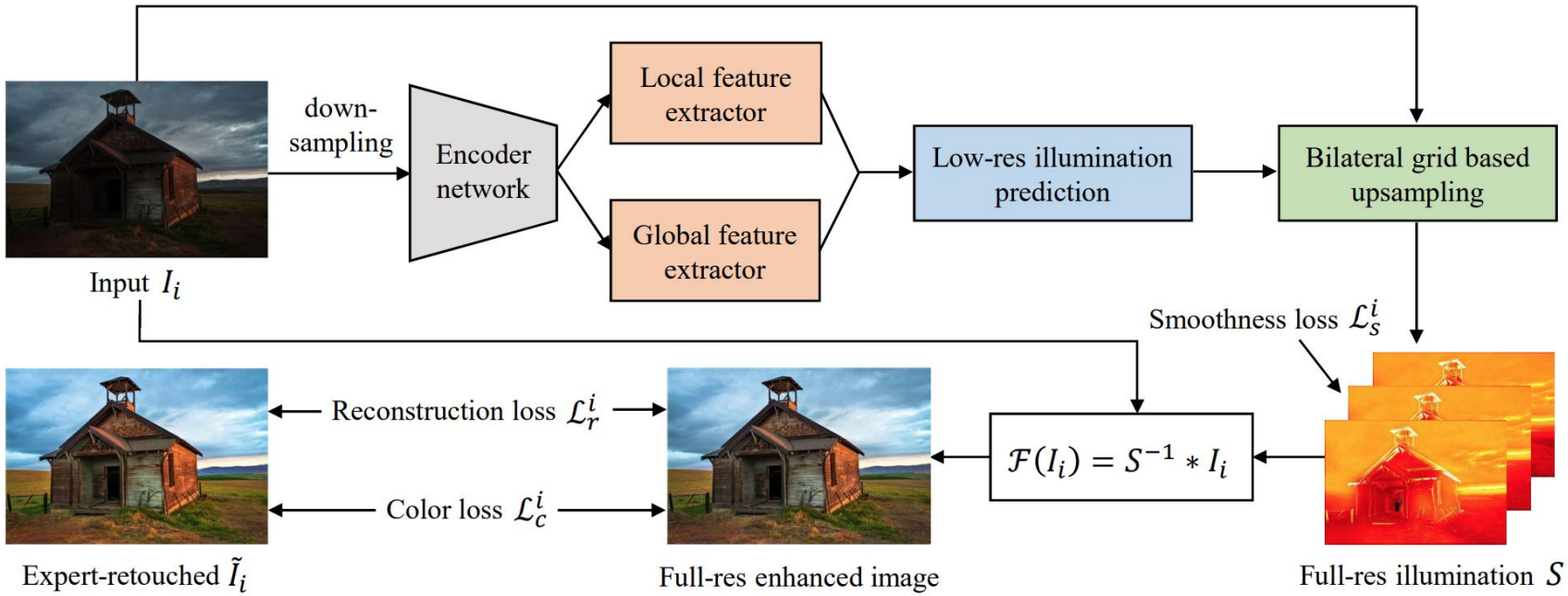


Figure 3: Overview of our network. First, we downsample and encode the input into a feature map, extract local and global features, and concatenate them to predict the low-res illumination via a convolution layer. Then we upsample the result to produce the full-res multi-channel illumination S (hot color map), and take it to recover the full-res enhanced image. We train the end-to-end network to learn S from image pairs $\{I_i, \tilde{I}_i\}$ with three loss components $\{L_r^i, L_s^i, L_c^i\}$.

Underexposed Photo Enhancement using Deep Illumination Estimation, Ruixing Wang et al., CVPR 2019, https://openaccess.thecvf.com/content_CVPR_2019/papers/Wang_Underexposed_Photo_Enhancement_Using_Deep_Illumination_Estimation_CVPR_2019_paper.pdf

Image Adaptive 3D LUT

- Overview

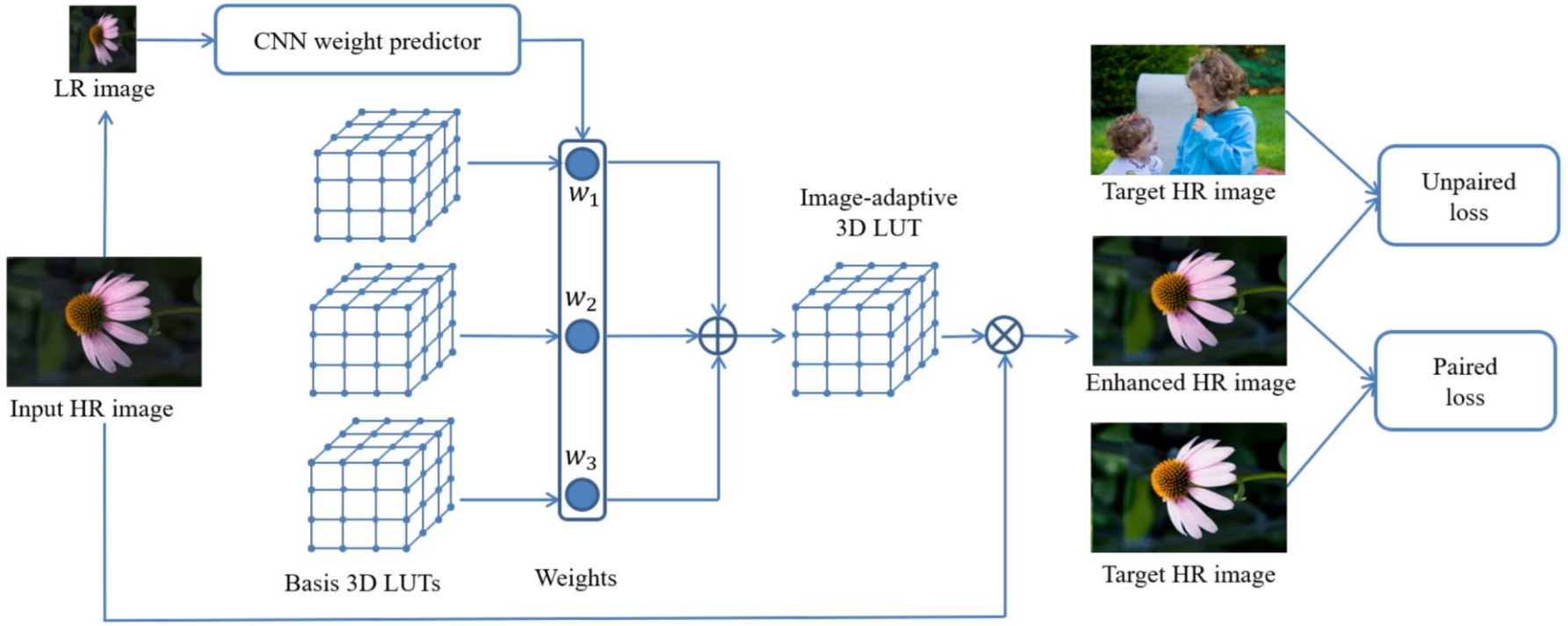


Fig. 1. Framework of the proposed image-adaptive photo enhancement method. Our method learns multiple basis 3D LUTs and a small CNN weight predictor. The CNN model works on a down-sampled version of the input image to predict content-dependent weights. The predicted weights are used to fuse the basis 3D LUTs into an image-adaptive LUT, which is then used to transform the source image. Our method can be trained using either paired or unpaired data in an end-to-end manner.

Learning Image-adaptive 3D Lookup Tables for High Performance Photo Enhancement in Real-time, Hui Zeng et al., https://arxiv.org/abs/2009.14468

3D Lookup Table

- Using $M \times M \times M$ (M=33) 3D Lookup Table, it contains $3M^3$ (108K) parameters.
- Trilinear Interpolation
  - 8 vertices affect the enhancement value

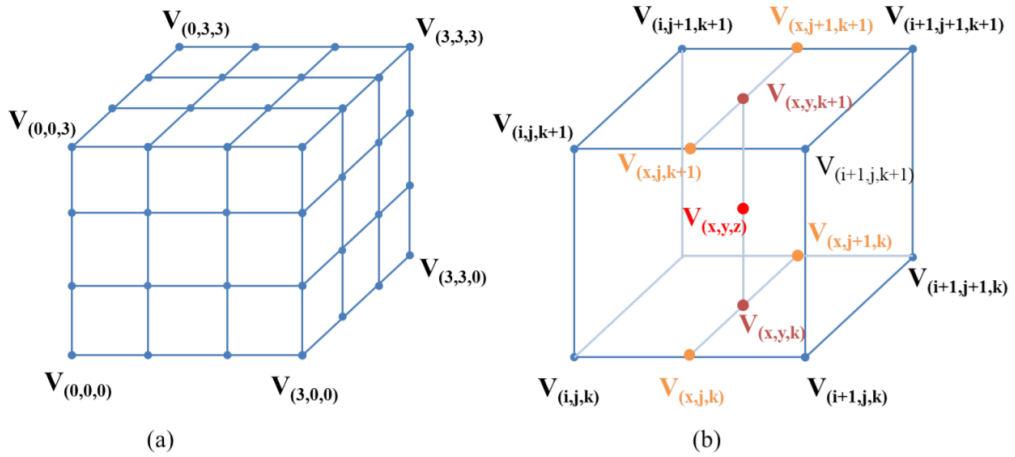


Fig. 2. Illustration of (a) a 3D LUT containing 4^3 elements and (b) the trilinear interpolation of one input.

CNN Weight Predictor

- $q = (\sum_{n=1}^{N} w_n \phi_n)(x)$

TABLE 1
Details of the CNN weight predictor. C, D, F and N represent convolutional block, dropout, fully-connected layer and number of weights, respectively.

Layer	C1	C2	C3	C4	C5	D	F
Input	256	128	64	32	16	8	8
Kernel	3	3	3	3	3	–	8
Channel	16	32	64	128	128	128	N

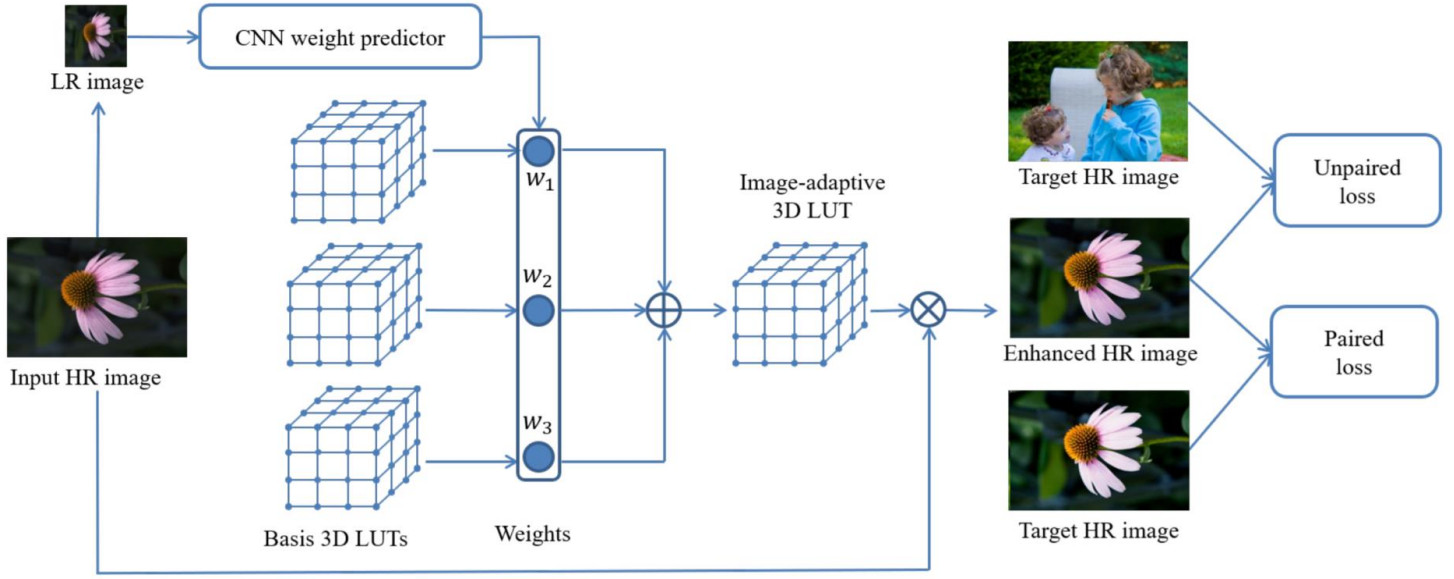


Fig. 1. Framework of the proposed image-adaptive photo enhancement method. Our method learns multiple basis 3D LUTs and a small CNN weight predictor. The CNN model works on a down-sampled version of the input image to predict content-dependent weights. The predicted weights are used to fuse the basis 3D LUTs into an image-adaptive LUT, which is then used to transform the source image. Our method can be trained using either paired or unpaired data in an end-to-end manner.

Learning Criteria

- Pairwise learning
  - MSE loss: $\mathcal{L}_{mse} = \frac{1}{T}\sum_{t=1}^{T}\|q_t - y_t\|^2$
- Unpaired learning
  - Using GAN loss: $\mathcal{L}_{gan} = \mathcal{L}_G + \mathcal{L}_D$
  - Generator: $\mathcal{L}_G = \underline{\mathbb{E}_x\big[-D\big(G(x)\big)\big]} + \underline{\lambda_1 \mathbb{E}[\|G(x) - x\|^2]}$ Preserves Content $\lambda_1 = 1000$
  - Discriminator: $\mathcal{L}_D = \underline{\mathbb{E}_x\big[D\big(G(x)\big)\big] - \mathbb{E}_y[D(y)]} + \underline{\lambda_2 \mathbb{E}_{\hat{y}}\left[\left(\left\|\nabla_{\hat{y}} D(\hat{y})\right\|_2 - 1\right)^2\right]}$

Adversarial Loss

Gradient Penalty $\lambda_2 = 10$

Regularization

- Smooth Regularization
  - Total Variation (TV) loss and L_2 regularization of the adaptive weight
  - $\mathcal{R}_{TV} = \sum_{c\in\{r,g,b\}} \sum_{i,j,k} \left(\left\| c^O_{(i+1,j,k)} - c^O_{(i,j,k)} \right\|^2 + \left\| c^O_{(i,j+1,k)} - c^O_{(i,j,k)} \right\|^2 + \left\| c^O_{(i,j,k+1)} - c^O_{(i,j,k)} \right\|^2 \right)$
  - $\mathcal{R}_s = \mathcal{R}_{TV} + \sum_n \|w_n\|^2$

Variance

- Monotonic Regularization
  - $\mathcal{R}_m = \sum_{c\in\{r,g,b\}} \sum_{i,j,k} \left(ReLU(c^O_{(i,j,k)} - c^O_{(i+1,j,k)}) + ReLU(c^O_{(i,j,k)} - c^O_{(i,j+1,k)}) + ReLU(c^O_{(i,j,k)} - c^O_{(i,j,k+1)}) \right)$

Monotonic

- Finally,
  - $\mathcal{L}_{total} = \mathcal{L}_{paired/unpaired} + \lambda_s \mathcal{R}_s + \lambda_m \mathcal{R}_m$
  - $\lambda_s = 10^{-4}, \lambda_m = 10$ via an ablation study

Summary

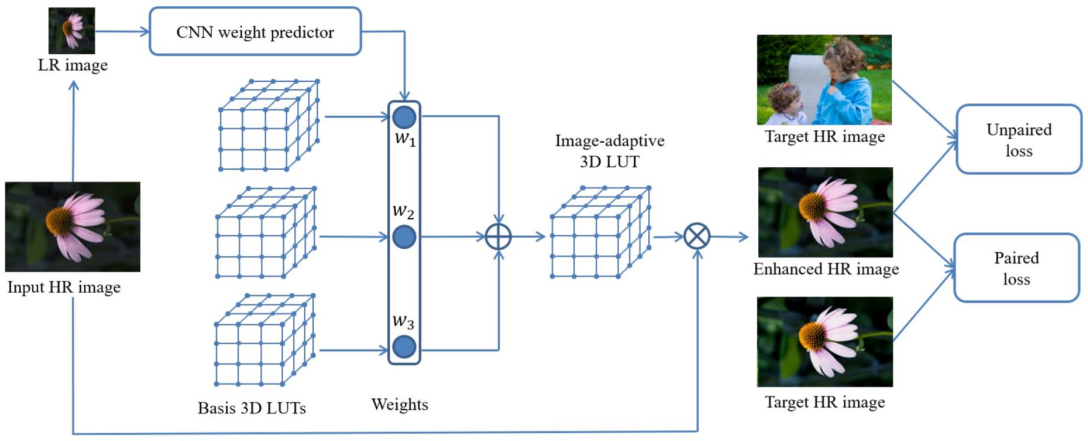


Fig. 1. Framework of the proposed image-adaptive photo enhancement method. Our method learns multiple basis 3D LUTs and a small CNN weight predictor. The CNN model works on a down-sampled version of the input image to predict content-dependent weights. The predicted weights are used to fuse the basis 3D LUTs into an image-adaptive LUT, which is then used to transform the source image. Our method can be trained using either paired or unpaired data in an end-to-end manner.

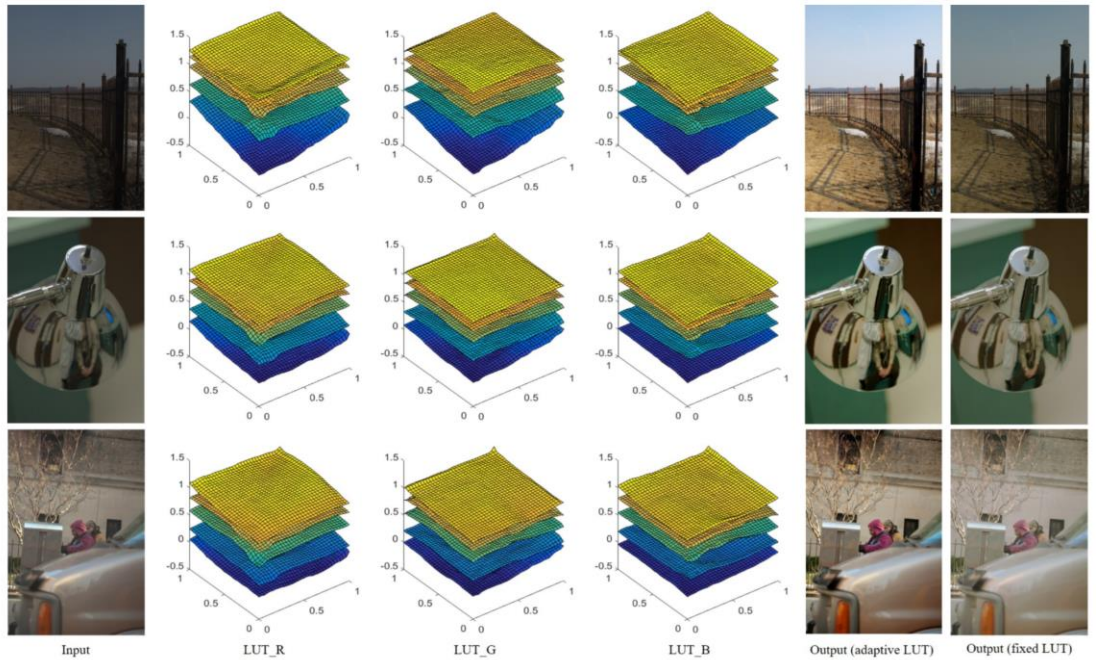


Fig. 6. Visualization of the image-adaptive LUTs (after combination) generated by our model for three different images and their corresponding enhancement results.

Quantitative Comparison

- Quantitative Comparison under the pair/unpaired setup

TABLE 3
Quantitative comparison of photo retouching results under the pairwise training setting on FiveK dataset. "N.A." means that the result is not available.

Method	FiveK (480p)			FiveK (original)		
	PSNR	SSIM	$\triangle E^*$	PSNR	SSIM	$\triangle E^*$
UPE [10]	21.88	0.853	10.80	21.65	0.859	11.09
Dis-Rec [8]	21.98	0.856	10.42	21.81	0.862	10.60
DPE [7]	23.75	0.908	9.34	N.A.	N.A.	N.A.
HDRNet [2]	24.32	0.912	8.49	24.03	0.919	8.68
Ours	25.21	0.922	7.61	25.10	0.930	7.72

TABLE 4
Quantitative comparison of photo retouching results under the unpaired training setting on the FiveK dataset. "N.A." means that the result is not available.

Method	FiveK (480p)			FiveK (original)		
	PSNR	SSIM	$\triangle E^*$	PSNR	SSIM	$\triangle E^*$
Camera Raw	21.61	0.854	11.83	21.55	0.861	11.98
Pix2Pix [49]	19.21	0.814	14.76	N.A.	N.A.	N.A.
CycGAN [50]	20.98	0.831	13.28	N.A.	N.A.	N.A.
White-Box [9]	21.32	0.864	12.65	21.17	0.875	12.81
DPE [7]	21.99	0.875	11.40	N.A.	N.A.	N.A.
UIE [11]	22.11	0.879	11.21	22.03	0.882	11.46
Ours	22.86	0.887	10.28	22.78	0.898	10.42

Qualitative Comparison

- MIT-Adobe FiveK dataset

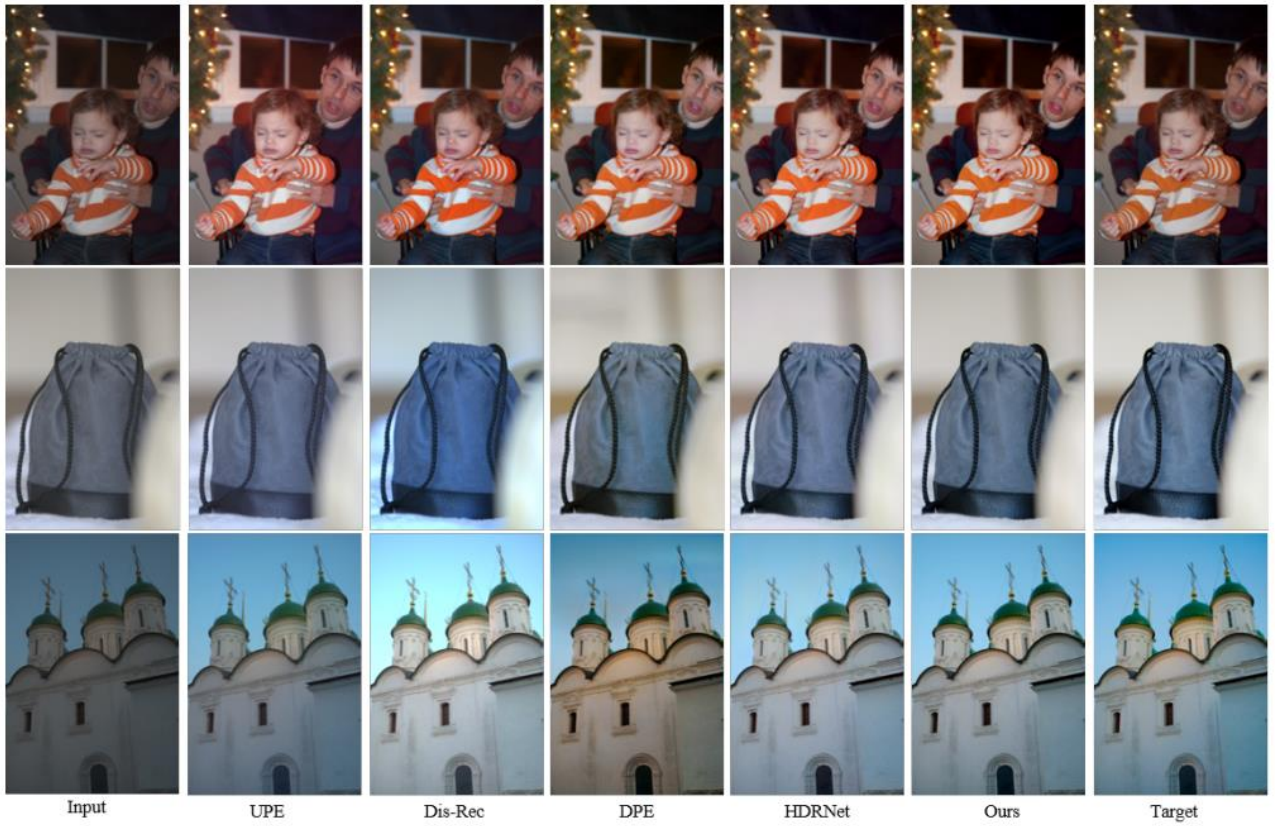


Fig. 7. Qualitative comparison of different paired learning methods for photo retouching on the FiveK dataset.

Qualitative Comparison

- More qualitative comparison with HDRNet

Fig. 8. Qualitative comparison between HDRNet and our method for photo retouching on the FiveK dataset. Note that some sky areas enhanced by HDRNet are biased to purple.

- User study

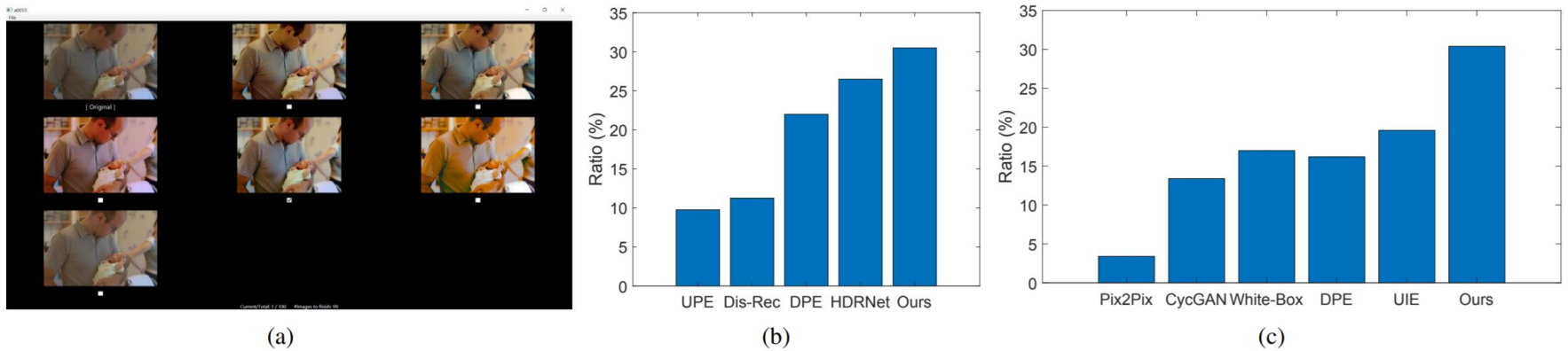


Fig. 11. (a) Interface of the user study tool; (b) voting statistics of paired learning methods; and (c) voting statistics of unpaired learning methods.

Ablation Study

- The number N of 3D LUTs

TABLE 2
Ablation studies on the number (N) of LUTs and the effect of CNN weight predictor.

N	1 (w/o CNN)	1	2	3	4	5
PSNR	20.37	23.15	24.86	25.21	25.26	25.29
SSIM	0.852	0.884	0.917	0.922	0.924	0.926
$\triangle E^*$	13.47	9.83	8.07	7.61	7.58	7.54

- Find λ_s and λ_m ($\mathcal{L}_{total} = \mathcal{L}_{paired/unpaired} + \lambda_s \mathcal{R}_s + \lambda_m \mathcal{R}_m$)

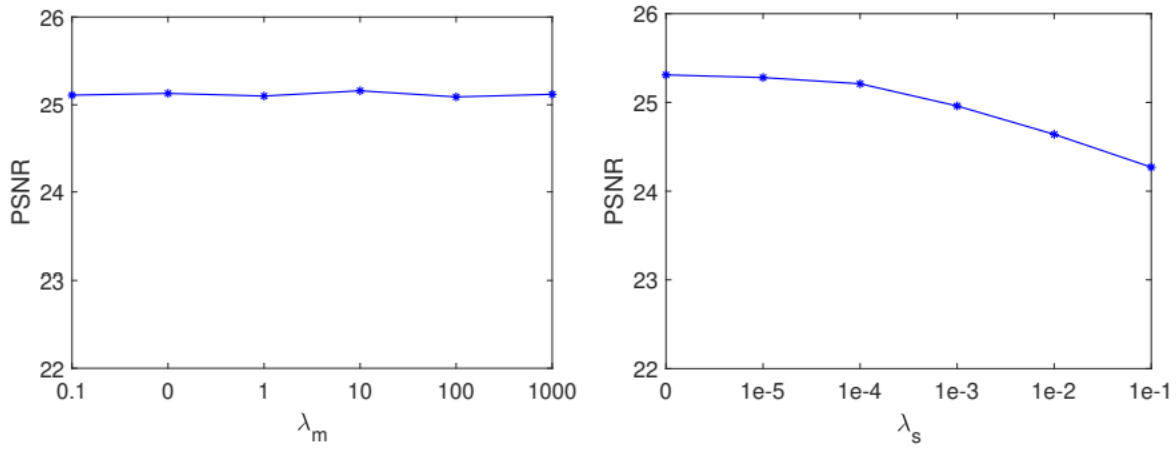


Fig. 3. Effects of different choices of parameters λ_s and λ_m.

Ablation Study

- Visualize Regularization

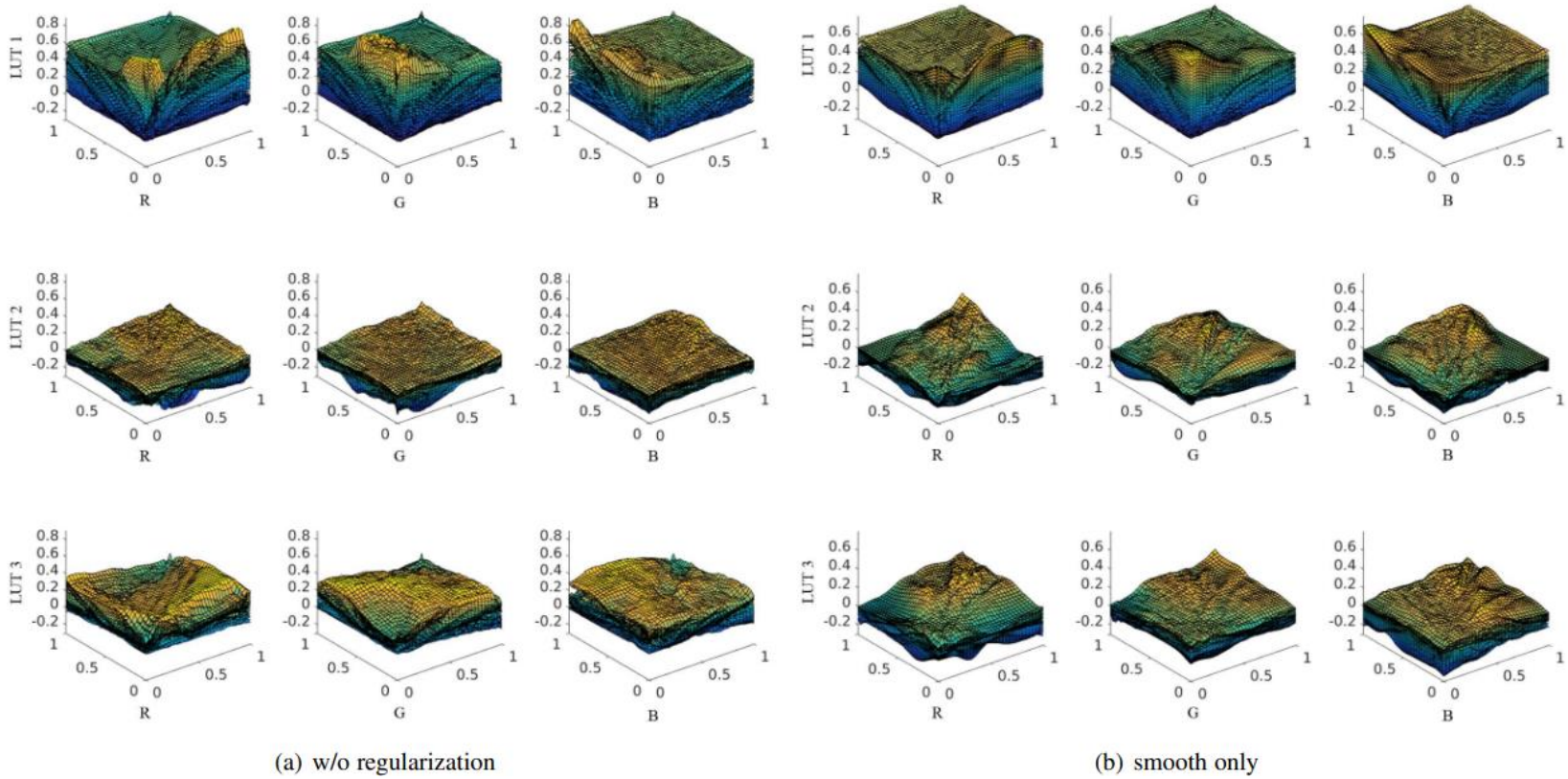
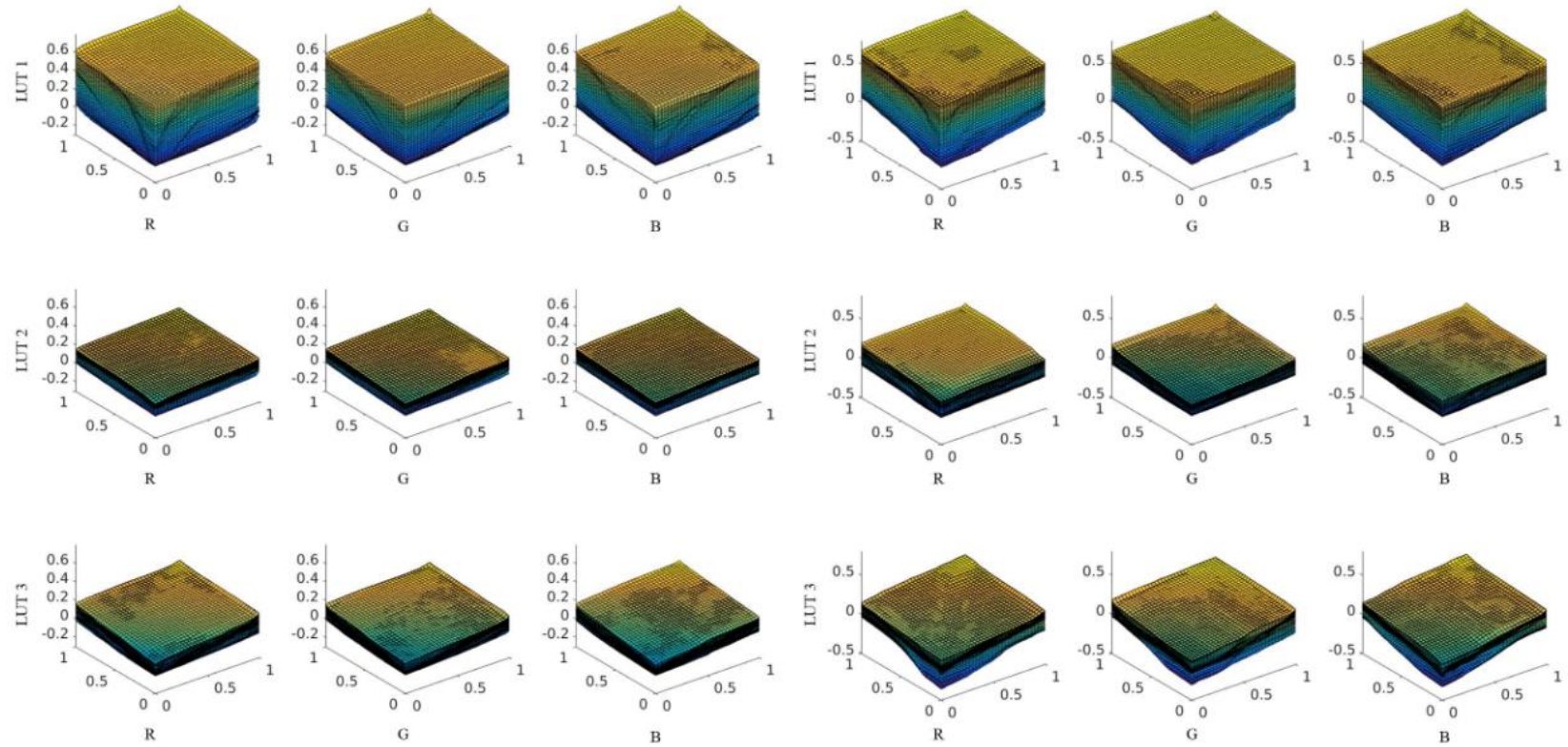


Fig. 4. Visualization of each of the R, G, B channels of the three basis 3D LUTs learned (a) without regularization, (b) with only smooth regularization, (c) with only monotonicity regularization and (d) with both smooth and monotonicity regularization.

Ablation Study

- Visualize Regularization



(c) monotonicity only (d) smooth + monotonicity

Fig. 4. Visualization of each of the R, G, B channels of the three basis 3D LUTs learned (a) without regularization, (b) with only smooth regularization, (c) with only monotonicity regularization and (d) with both smooth and monotonicity regularization.

Ablation Study

- Why regularization?

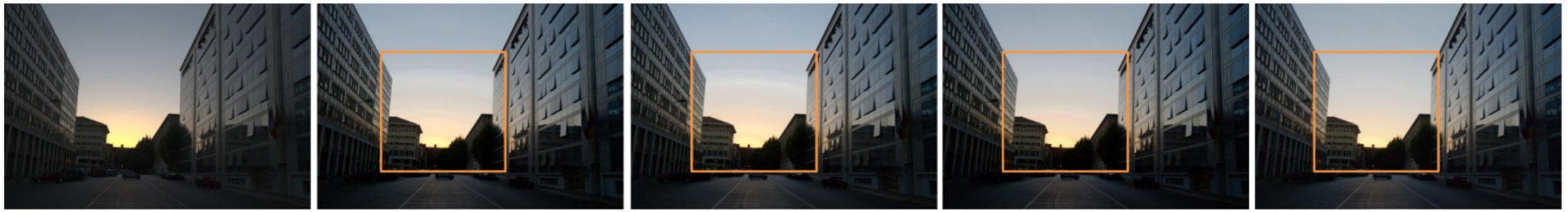


Fig. 5. Enhancement results of a typical input image (a) without regularization, (b) with only smooth or (c) only monotonicity regularization, and (d) with both smooth and monotonicity regularization. Best viewed in color.

Running Speed

- Evaluate the running speed in 1920×1080, 3840×2160, 6000×4000 sized image.

TABLE 7
Running time (in milliseconds) comparison between our model and state-of-the-art deep enhancement models at different resolutions. All models are tested using one Titan RTX GPU. "N.A." means that the result is not available.

Resolution	1920×1080	3840×2160	6000×4000
Pix2Pix [49]	1.2e2	N.A.	N.A.
CycGAN [50]	5.6e2	N.A.	N.A.
DPE [7]	8.6e1	N.A.	N.A.
White-Box [9]	5.0e3	9.1e3	2.0e4
Dis-Rec [8]	2.5e4	1.1e5	3.3e5
UIE [11]	1.0e4	2.0e4	3.3e4
HDRNet [2]	4.5e1	2.1e2	5.9e2
UPE [10]	4.5e1	2.1e2	5.9e2
Ours	**0.64**	**1.66**	**3.76**

Limitation and Discussion

- 3D LUT may produce less satisfactory results in some areas requiring local enhancement.
  - 3D LUT operation is the same for different local areas within the image.
- Limited solution is using local operator

Fig. 12. Enhancement results on an input scene with very high dynamic range. Our 3D LUT model effectively enhances the global color and tone, while the local contrast in the shadow area can be further improved by combining some local tone mapping operator with our model.

Limitation and Discussion

- As a compact and highly efficient operator, 3D LUT transforms each input “RGB” value independently

Fig. 13. Enhancement results generated by our 3D LUT model and Google’s image processing pipeline on a noisy input (with 16-bit dynamic range) from the HDR+ dataset. Although our model properly enhances the dark area, the noise is also amplified after the enhancement. Please zoom in to see the details.

CSRNet

- Very Lightweight Network

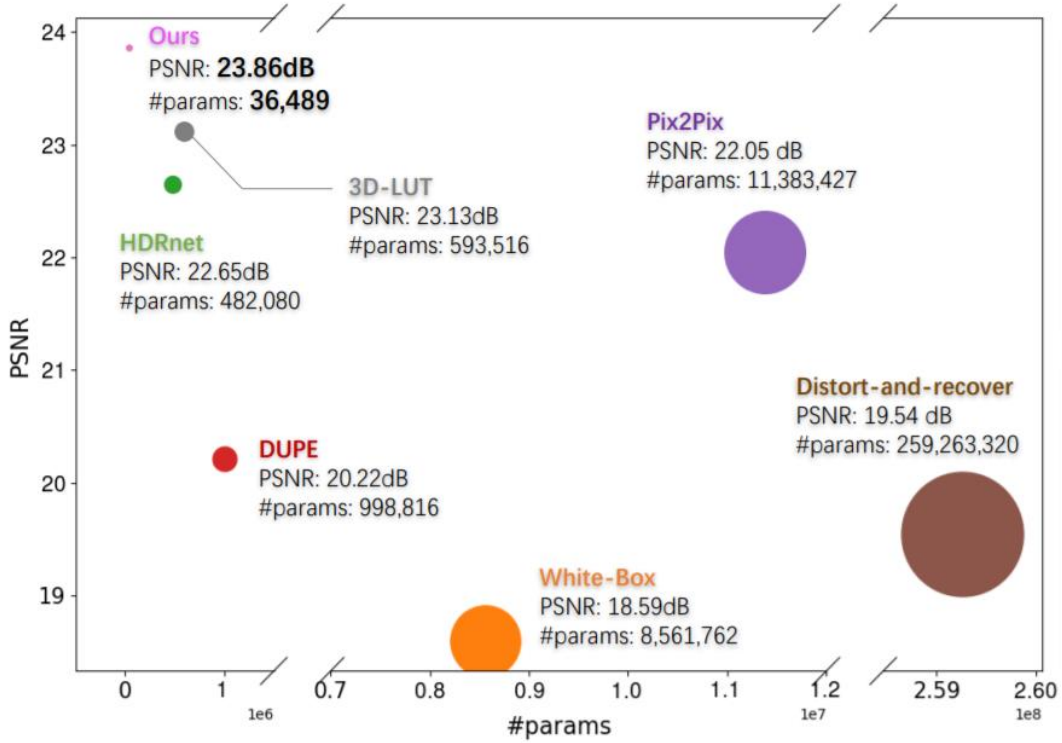


Fig. 1. Left: Compared with existing state-of-the-art methods

Very Lightweight Photo Retouching Network with Conditional Sequential Modulation, Yihao Liu et al., ECCV 2020, https://arxiv.org/abs/2104.06279

Retouching Operations and MLPs

- Commonly-used retouching operations can be regarded as classic MLPs used on input image.
- White-balancing
  - $I'_{RGB} = [I_R; I_G; I_B] \circ [\alpha_R; \alpha_G; \alpha_B]$
- Contrast Adjustment
  - $I'(x,y) = \alpha I(x,y) + (1-\alpha)\bar{I}$
  - $\bar{I} = \frac{1}{M \times N} \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} I(x,y)$
- Saturation Controlling
  - $I'(x,y) = \alpha I(x,y) + (1-\alpha)\bar{I}_{RGB}(x,y)$
  - $\bar{I}_{RGB} = \frac{1}{3}(I_R + I_G + I_B)$
- Tone Mapping
  - L: number of interval
  - $t_0, t_1, \cdots t_L$: height of tone curve

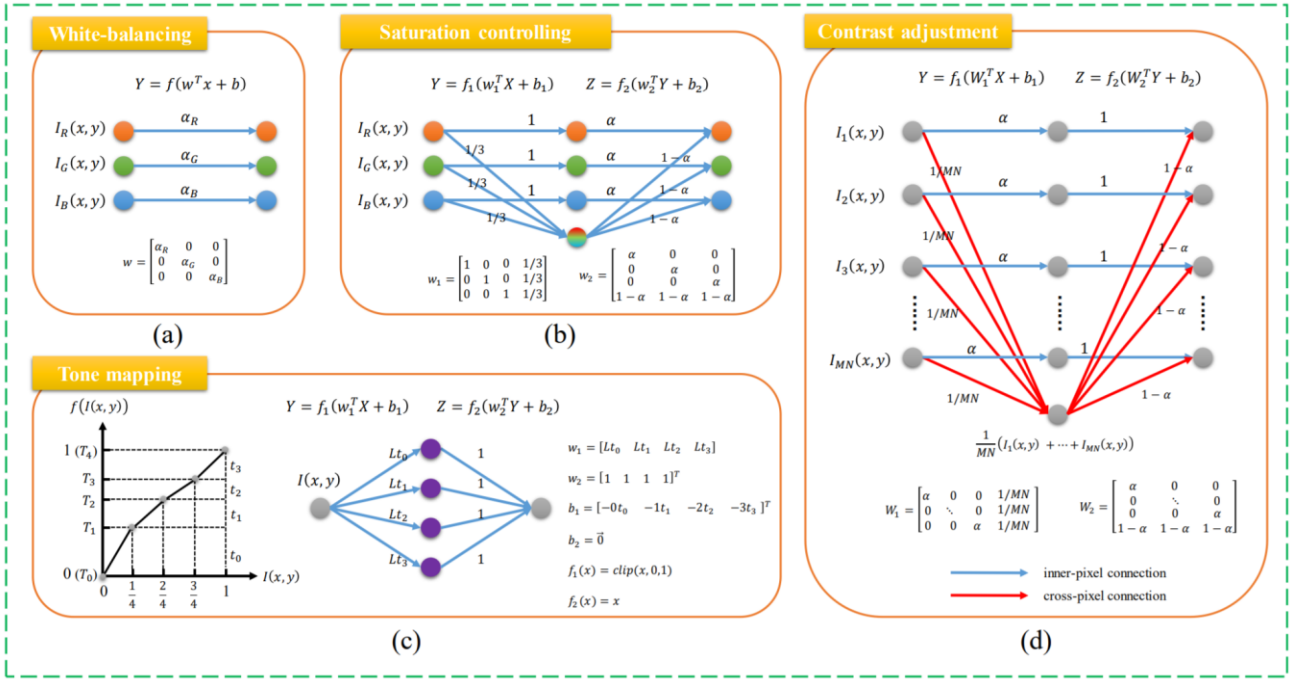
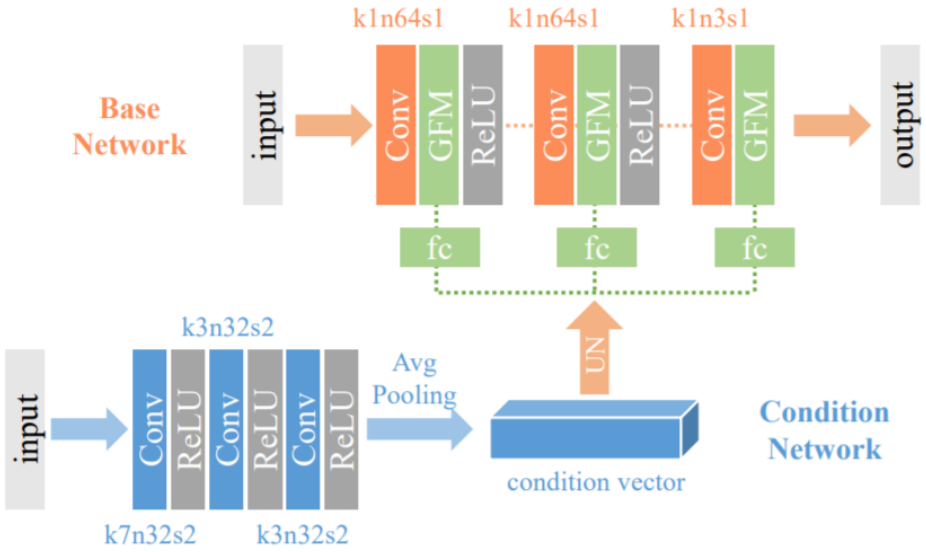


Fig. 2. Illustration of the equivalent MLPs for the corresponding retouching operations.

Model Architecture Overview

- Base network
  - Only composed 1×1 convolution layers
  - Using ReLU
- Condition network
  - Convolutional layers
  - Down-sampling
- GFM
  - $GFM(x) = \gamma \odot x + \beta$
- UN (Unit Normalization)
  - $UN(x_i) = \sqrt{N}\frac{x_i}{\|x\|_2}$



(b) CSRNet. (k: kernel size; n: number of feature maps; s: stride.)

Fig 3. (b) The proposed network consists of four key components – base network, condition network, GFM and UN.

Condition Network

- Role of condition network

TABLE 1
Demonstration experiment on simulating retouching operations. Our method can successfully handle commonly-used retouching opereations, which is consistent with the theoretical analysis. The results in "contrast" ajustment also show the significance of adopting the condition network.

Operations	Original (Input-GT)	Base Netwok	Condition Netwok	PSNR
brightness ($\alpha = 1.5$)	✓	×	×	14.7413
	×	✓	×	**69.7061**
brightness ($\alpha = 0.5$)	✓	×	×	12.8460
	×	✓	×	**69.0525**
tone-mapping* ($L = 4$)	✓	×	×	21.7580
	×	✓	×	**56.1175**
contrast ($\alpha = 1.5$)	✓	×	×	21.3584
	×	✓	×	28.6734
	×	✓	✓	**60.5206**

* The parameters for tone-mapping are set to $t_i = [3/8, 2/8, 1/8, 2/8]$.

TABLE 5
Results of ablation study for the condition network.

	Handcrafted prior	Dim	PSNR	#params
w/o condition network	None	0	20.47	4,611
	brightness	1	21.47	5,135
	average intensity	3	21.93	5,659
	histograms	768	22.90	206,089
w condition network	None (ours)	32	23.86	36,489
	brightness	1+32	23.41	36,751
	average intensity	3+32	23.67	37,275
	histograms	768+32	23.51	237,705

Quantitative Comparison

- On MIT-Adobe FiveK dataset (expert C).

TABLE 2
Quantitative comparison with state-of-the-art methods on MIT-Adobe FiveK dataset (expert C).

Method	Running Time	PSNR↑	SSIM↑	L2 error (L*a*b)↓	#params
White-Box [2]	1028.91ms	18.59	0.797	17.42	8,561,762
Distort-and-Recover [9]	4063.35ms	19.54	0.800	15.44	259,263,320
HDRNet [1]	6.03ms	22.65	0.880	11.83	482,080
DUPE [1]	8.47ms	20.22	0.829	16.63	998,816
MIRNet [35]	252.60ms	19.37	0.806	16.51	31,787,419
Pix2Pix [28]	181.98ms	21.41	0.749	13.26	11,383,427
3D-LUT [3]	1.60ms	23.12	0.874	11.26	593,516
CSRNet (ours)	1.92ms	**23.86**	**0.897**	**10.57**	**36,489**
DPE [6]	17.73ms	23.76	0.881	10.60	3,335,395
CSRNet (ours)	1.92ms	**24.37**	**0.902**	**9.52**	**36,489**

Multiple Styles and Strength Control

- The other networks, (DUPE, DPE, HDRNet) should train from scratch for fit to different photographers
- Flexibility of multiple style learning
- Only finetuning the condition network, which is much faster than training from scratch

TABLE 7
Performance for Multiple styles (A/B/D/E).

expert	PSNR (finetune)	PSNR (scrach)
A	22.49	22.35
B	25.63	25.60
D	23.01	23.12
E	23.93	23.89

References

- Learning Photographic Global Tonal Adjustment with a Database of Input / Output Image Pairs, V. Bychkovsky, et al., CVPR 2011, https://ieeexplore.ieee.org/document/5995332
- Underexposed Photo Enhancement using Deep Illumination Estimation, Ruixing Wang et al., CVPR 2019, https://openaccess.thecvf.com/content_CVPR_2019/papers/Wang_Underexposed_Photo_Enhancement_Using_Deep_Illumination_Estimation_CVPR_2019_paper.pdf
- Very Lightweight Photo Retouching Network with Conditional Sequential Modulation, Yihao Liu et al., ECCV 2020, https://arxiv.org/abs/2104.06279
- Learning Image-adaptive 3D Lookup Tables for High Performance Photo Enhancement in Real-time, Hui Zeng et al., https://arxiv.org/abs/2009.14468