Supplementary Material for You Only Need 90K Parameters to Adapt Light: a Light Weight Transformer for Image Enhancement and Exposure Correction

Ziteng Cui[1]
cui@mi.t.u-tokyo.ac.jp
Kunchang Li[2]
kc.li@siat.ac.cn
Lin Gu[3,1*]
lin.gu@riken.jp
Shenghan Su[4]
su2564468850@sjtu.edu.cn
Peng Gao[2]
gaopeng@pjlab.org.cn
Zhengkai Jiang[5]
zhengkjiang@tencent.com
Yu Qiao[2]
qiaoyu@pjlab.org.cn
Tatsuya Harada[1,3]
harada@mi.t.u-tokyo.ac.jp

[1] The University of Tokyo
[2] Shanghai AI Laboratory
[3] RIKEN AIP
[4] Shanghai Jiao Tong University
[5] Tencent Youtu Lab

A Analyse on Module Structure

For the global part g of the IAT module, here we simplify the ISP procedures [5, 9, 10] as the following equation:

$$G(\cdot) = Gamma(W_{ccm}(W_{wb}(\cdot))). \tag{1}$$

White balance (WB) function is an essential part in ISP pipeline. WB algorithm estimates the per channel gain on the image, to maintain the object's colour constancy under various different light colour. WB is usually represented as a 3×3 diagonal von Kris matrix W_{wb} in camera imaging pipeline [1, 3, 5, 12]. After that, camera color matrix (CCM) W_{ccm} converts the white-balanced data from camera internal color space cRGB to sRGB colour space [5, 10, 12]. At last gamma correction aims to match non-linearity of humans perception on dark regions. A standard gamma curve is usually represent as an exponential function with the exponential parameter γ, so we build our global branch $g_t(\cdot)$ following the equation:

Table A1: Comparison experiments of with (w) and without (w/o) raw-RGB supervision on exposure correction dataset [2].

Method	Expert A		Expert B		Expert C		Expert D		Expert E	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
w/o raw-RGB	19.90	0.817	21.65	0.867	**21.23**	**0.850**	19.86	0.844	19.34	0.840
w raw-RGB	**19.98**	**0.822**	**22.03**	**0.885**	21.16	0.843	**19.94**	**0.852**	**19.48**	**0.841**

$$g_t(\cdot) = (max(\sum_{c_j} W_{c_i,c_j}(\cdot), \varepsilon))^{\gamma}, c_i, c_j \in \{r,g,b\}, \tag{2}$$

where the W_{c_i,c_j} is a joint colour transform function consist of white balance matrix and CCM and γ is the gamma correction's exponential value, ε is a minimum number to keep non-negative. Final as we discussed in Sec.3.1, the input image I_i would separately pass by local branch f and global branch g to generate the prediction result $\hat{I}_t = g_t(f(I_i))$.

We also evaluate to train the model with corresponding raw-RGB data as additional supervision. Since it's hard to directly get raw-RGB data from the currect dataset, we then adopt the Invertible ISP [23] to generate corresponding raw-RGB data I_{raw} from the input image I_i, we use pre-train weights in [23] to generate I_{raw}. In the training stage, we additional add a loss function L_{raw} for raw-RGB supervision, the total loss function shown as follow:

$$\begin{aligned} L_{total} =& L_{rgb} + \lambda \cdot L_{raw} \\ =& L_1(g_t(f(I_i), I_t) + \lambda \cdot L_1(f(I_i), I_{raw}). \end{aligned} \tag{3}$$

L_{total} is the total loss function that consist of two parts: the first part L_{rgb} is L1 loss function between predict result $g_t(f(I_i))$ with ground truth image I_t, while the second part L_{raw} is the L1 loss function between middle representation $f(I_i)$ and raw-RGB image I_{raw} for raw-RGB part supervision, and λ is a balance parameter where we set it to 0.1 in our experiments. We make the comparison experiments on exposure correction dataset [2], the training and experiments' settings are follow the settings in Sec.4.2, only difference is the training strategy with or without raw-RGB supervision. The comparison results are shown in Table A1, we can find that with the additional supervision of raw-RGB data, most of evaluation metrics on exposure correction dataset [2] would be improved.

B Joint Training with High-level Framework

For high-level vision tasks under challenging lighting conditions,shown in Fig.B1, current high-level vision frameworks [7, 8, 13] usually well-trained on large scale normal-light datasets (*i.e.* MS COCO [14], ImageNet [11]), so directly take low-light/ strong-light data as input would cause the lightness in-consistency, on the other hand, using image enhancement methods (Sec.4.3 in main text) to pre-process images may cause target inconsistency (human vision *v.s.* machine vision) [10], since the goal of image restoration is image quality (*i.e.* PSNR, SSIM) and the goal of detection/ segmentation is machine-vision accuracy (*i.e.* mAP, mIOU).

An example is shown in Fig. B1, by attaching IAT to the downstream task module, our IAT could conduct object detection and semantic segmentation with the downstream frameworks. During training, we aim to minimise the downstream framework's loss function (*i.e.* object detection loss L_{obj} between detection prediction $\hat{t}$ and ground truth t) by jointly optimising the whole network's parameters (see Eq. 4). Compared to the subsequent high-level

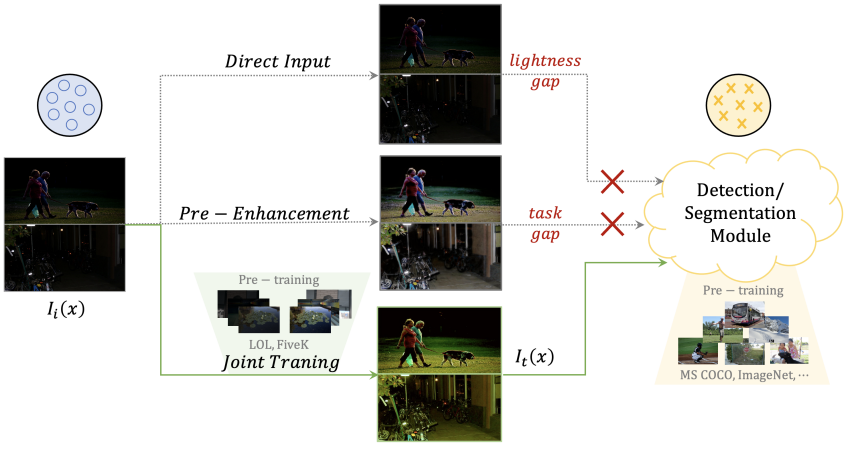

Figure B1: Joint training Enhancement Module with High-level Module.

Table B2: Comparison experiments on low-light detection dataset EXDark [15] and low-light semantic segmentation dataset ACDC [19].

	original	pre-enhancement		joint training		
		IAT (LOL)	IAT (MIT5K)	IAT (none)	IAT (MIT5K)	IAT (LOL)
EXDark (mAP↑)	76.4	77.2	76.9	77.1	77.6	**77.8**
ACDC (mIOU↑)	63.3	62.1	61.3	61.5	62.1	**63.8**

module, the time-complexity and model storage of our IAT main structure could be ignored (*i.e.* IAT main structure *v.s.* YOLO-V3 [18], 417KB *v.s.* 237MB).

$$\min_{i\in\mathbb{I},d\in\mathbb{D}} L_{obj}(\hat{t},t)$$
$$I_t(x)=\mathbb{I}(I_1(x)),\quad \hat{t}=\mathbb{D}(I_t(x)) \tag{4}$$

We make the comparison experiments on low-light detection dataset EXDark [15] and low-light semantic segmentation dataset ACDC [19]. For object detection task we adopt the YOLO-V3 [18] object detector and for segmentation task we adopt DeepLabV3+ [8] segmentation framework, the training and experiments' settings are follow the settings in Sec.4.3.

Experimental results are shown in Table. B2, "original" means to take the original low-light images for training and evaluation, "pre-enhancement" means to pre-enhancement the EXDark [15] and ACDC [19] datasets with IAT model trained on LOL-V1 dataset [22] ("IAT (LOL)") and MIT-Adobe FiveK dataset [6] ("IAT (MIT5K)"). The "joint training" means to joint train IAT with the following high-level framework, and IAT model is separately random initialize ("IAT (none)"), initialize with LOL pre-train weights ("IAT (LOL)") and initialize with MIT-Adobe FiveK weights ("IAT (MIT5K)"), from Table. B2 we could see that joint-training IAT with the high-level frameworks would further improve high-level visual performance, on both of object detection and semantic segmentation task.

C Ablation Studies

Table C3: Experiments on LOL-V2-real [22] dataset (SSIM, PSNR) and EXDark [15] dataset (mAP), shows each part's contribution of IAT.

Local Branch	Layer Norm	[20]'s Norm	Our Norm	Global (matrix)	Global (gamma)	PSNR↑	SSIM↑	mAP↑
√	√					18.80	0.762	75.8
√		√				19.61 (+0.81)	0.776 (+0.014)	75.8 (+0.0)
√			√			20.01 (+1.21)	0.786 (+0.024)	76.3 (+0.5)
√			√	√		21.95 (+3.15)	0.811 (+0.049)	76.5 (+0.7)
√			√		√	22.76 (+3.96)	0.805 (+0.043)	76.7 (+0.9)
√			√	√	√	**23.50** (+4.70)	**0.824** (+0.062)	**77.1** (+1.3)

Table C4: Blocks Number.

M \ A	2	3	4
2	22.10	22.85	22.34
3	22.24	**23.50**	22.67
4	22.42	23.00	23.48

Table C5: Channel Number.

#Channel:#Block	PSNR↑	SSIM↑	#Param.↓ (K)
Long and Thin (12:4)	22.60	0.807	**86.22**
Short and Thick (24:2)	22.70	0.815	101.03
Ours (16:3)	**23.50**	**0.824**	91.15

C.1 Contribution of each part.

To evaluate each part's contribution in our IAT model, we make an ablation study on the low-light enhancement task of LOL-V2-real [22] dataset, and the low-light object detection task of EXDark [15] dataset. We report the PSNR and SSIM results of the enhancement task and the mAP result of the detection task. We compare our normalization with LayerNorm [4] and ResMLP's normalization [20], and then evaluate different parts' contributions of the global branch (predict matrix and predict gamma value). The ablation results are shown in Table. C3.

C.2 Blocks & Channels Ablation.

To evaluate the scalability of our IAT model, we try the different block numbers and channel numbers in the local branch. We try different PEM numbers to generate M and A. The PSNR results on LOL-V2-real [22] dataset has been shown in Table.C4. It shows that keeping the same PEM number to generate M and A would be helpful to IAT's performance.

Keeping the same block number to generate M and A, we then evaluate with similar parameters to answer whether the local branch should be "short and thick" or "long and thin". The local branch's block number and channel number are respectively set to $2/24$ and $4/12$ for comparison. The results of PSNR, SSIM and model parameters are reported in Table. C5.

D Additional Qualitative Results.

In this section we show more qualitative results on low-level vision tasks: image enhancement (LOL (V1 & V2-real) [22], MIT-Adobe FiveK [6]) and exposure correction [2].

D.1 Image Enhancement Results

Fig. D1 shows the image enhancement results on LOL-V1 dataset [22] compare with RCT [13] and MBLLEN [16], Fig. D2 shows the image enhancement results on LOL-V2-real dataset [22] compare with MBLLEN [16] and KIND [24]. Fig. D3 shows the image enhancement results on MIT-Adobe FiveK dataset [6] compare with Deep-UPE [21] and Deep-LPF [17]. We could see that IAT can generate higher quality images which closer to reference target image I_t. Meanwhile IAT also take much fewer parameters and less inference time.

D.2 Exposure Correction Results

Fig. D4 shows the exposure correction results on [2] dataset, we show both under-exposure and over-exposure results of our IAT, and compare to five experts' results. IAT also generate high quality images, and have ability to handle under/over-exposure at same time.

References

[1] Mahmoud Afifi and Michael S. Brown. Deep white-balance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[2] Mahmoud Afifi, Konstantinos G. Derpanis, Bjorn Ommer, and Michael S. Brown. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[3] Mahmoud Afifi, Marcus A. Brubaker, and Michael S. Brown. Auto white-balance correction for mixed-illuminant scenes. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[4] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016.

[5] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[6] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, 2020.

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision*, 2018.

[9] Shiqi Chen, Huajun Feng, Keming Gao, Zhihai Xu, and Yueting Chen. Extreme-quality computational imaging via degradation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[10] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, 2009.

[12] Hakki Can Karaimer and Michael S. Brown. A software platform for manipulating the camera imaging pipeline. In *European Conference on Computer Vision*, 2016.

[13] Hanul Kim, Su-Min Choi, Chang-Su Kim, and Yeong Jun Koh. Representative color transform for image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014.

[15] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 2019.

[16] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *British Machine Vision Conference*, 2018.

[17] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[18] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[19] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[20] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021.

[21] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[22] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018.

[23] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *CVPR*, 2021.

[24] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, 2019.

Figure D1: Qualitative comparison results on LOL-V1 [22] dataset, compare with enhancement methods MBLLEN [16] and RCT [13].

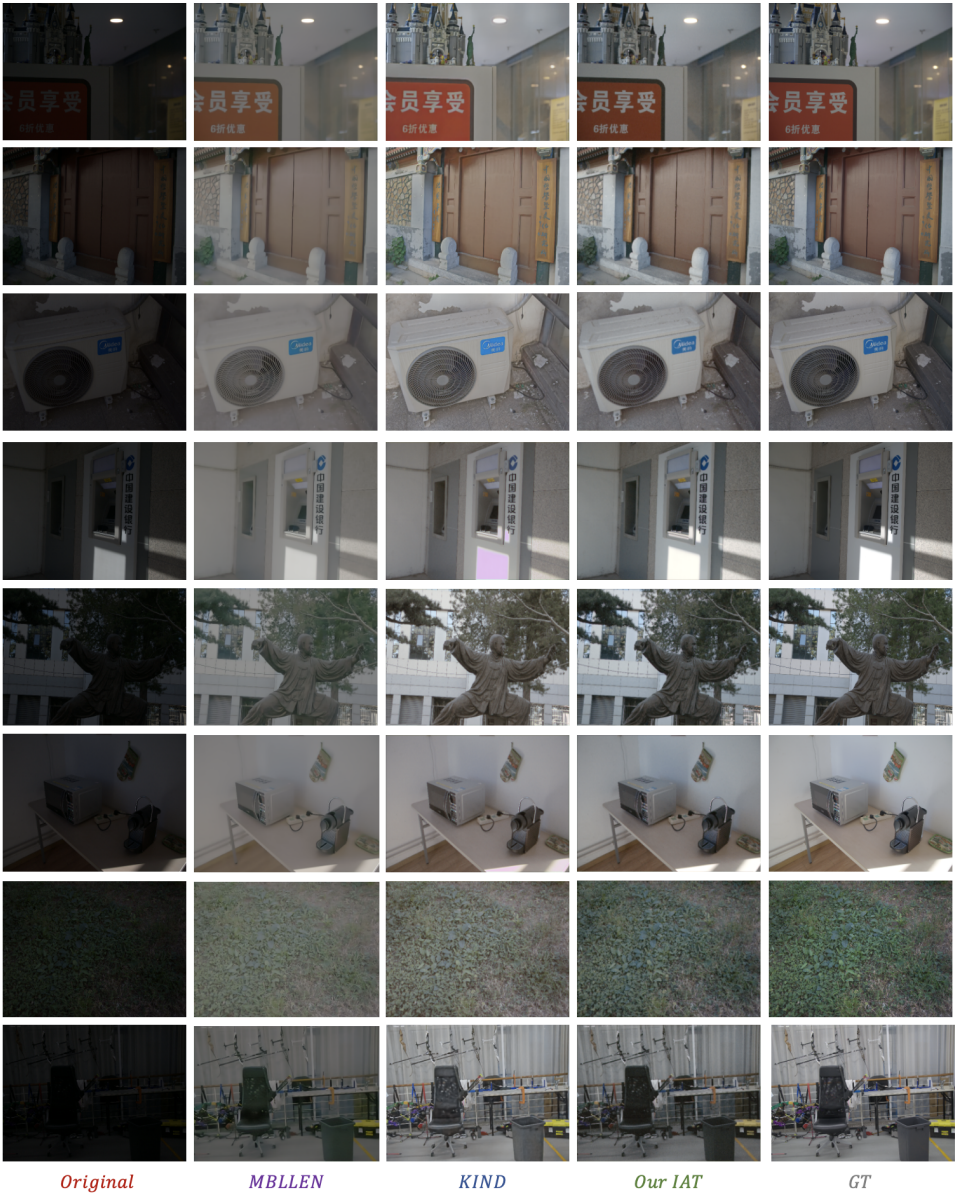

Figure D2: Qualitative comparison results on LOL-V2-real [22] dataset, compare with enhancement methods MBLLEN [16] and KIND [24].

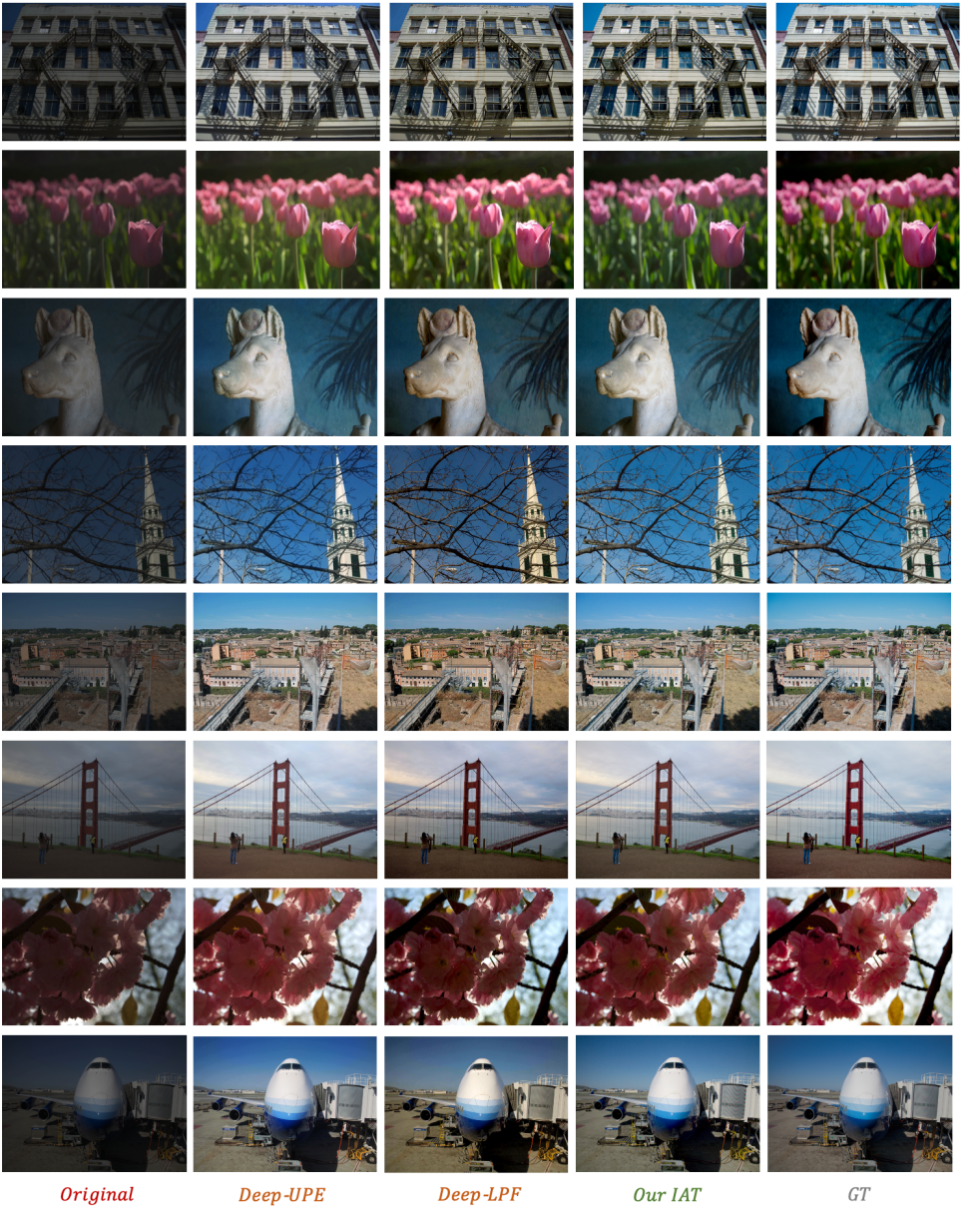

Figure D3: Qualitative comparison results on MIT-Adobe FiveK [6] dataset, compare with enhancement methods Deep-UPE [21] and Deep-LPF [17].

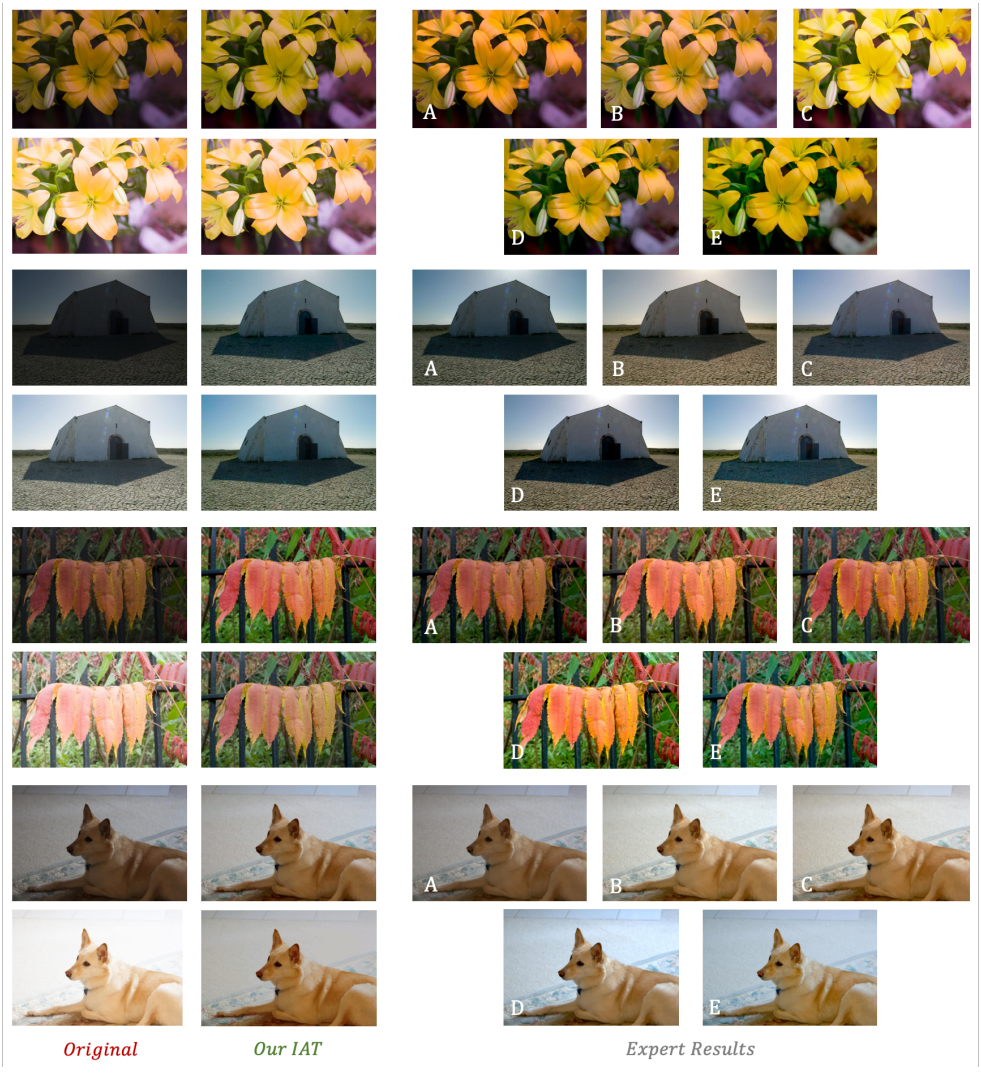

Figure D4: Qualitative comparison results of both under-exposure and over-exposure images on exposure correction dataset [0], left is input image, second row is output of our IAT, right are 5 experts' results.