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

Abstract

Challenging illumination conditions (low light, under-exposure and over-exposure) in the real world not only cast an unpleasant visual appearance but also taint the computer vision tasks. After camera captures the raw-RGB data, it renders standard sRGB images with image signal processor (ISP). By decomposing ISP pipeline into local and global image components, we propose a lightweight fast Illumination Adaptive Transformer (IAT) to restore the normal lit sRGB image from either low-light or under/over-exposure conditions. Specifically, IAT uses attention queries to represent and adjust the ISP-related parameters such as colour correction, gamma correction. With only ∼**90k** parameters and ∼**0.004s** processing speed, our IAT consistently achieves superior performance over State-of-The-Art (SOTA) on the benchmark low-light enhancement and exposure correction datasets. Competitive experimental performance also demonstrates that our IAT significantly enhances object detection and semantic segmentation tasks under various light conditions. Our code and pre-trained model is available at this url.

1 Introduction

Computer vision has witnessed great success on well-taken images and videos. However, the varying light conditions in the real world poses challenges on both human visual appearance

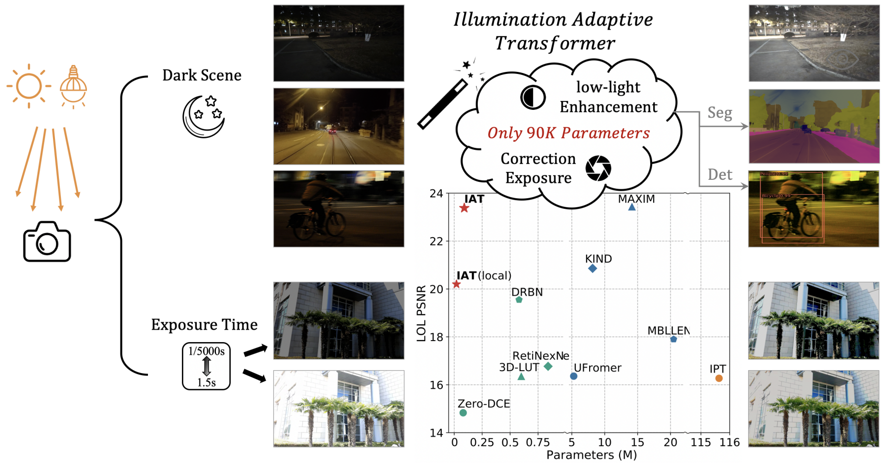


Figure 1: *Nature lay hid in Night*, apply our IAT, *and all was light*, middle figure shows our results compared with other SOTA methods on LOL-V1 dataset [60].

and downstream vision tasks (*e.g.*, semantic segmentation and object detection). Images taken under inadequate illumination (Fig.1 left top) suffer from limited photon counts and undesirable in-camera noise. On the other hand, outdoor scenes are often exposed to strong light such as direct sunlight (Fig.1 left down), making image saturated due to the limited range of sensors and non-linearity in the camera image pipeline. To make it worse, both the under and over exposure may exist together, *i.e.* spatial-variant illumination cast by shadow could make the contrast ratio to be 1000:1 or higher.

Multiple techniques such as low-light enhancement [19, 20, 31, 33, 39, 40, 42, 51, 53, 55, 69, 72], exposure correction [2, 43, 44, 64, 67] have been proposed to adapt to the difficult light condition. Low-light enhancement methods restore the details while suppressing the accompanying noises. Exposure correction methods focus on adjusting the under/over exposure image to reconstruct a clear image against short/long exposure time. While many efforts elaborate on improving human oriented visual perceptual, there also several methods enhance the high-level tasks by boosting their robustness against low light [15, 25, 36, 50] and over-exposure conditions [41]. As shown in Fig.1, we aim for a unified lightweight framework that improves both the visual appearance and consequent recognition tasks under challenging real-world light condition.

While sRGB images are most common to everyday life, many existing light adaptive methods operate on raw-RGB images which linearly proportion to the actual scene irradiance. To directly process the sRGB images, we specifically considers the image signal processor (ISP) in camera that renders sRGB from raw-RGB image. We propose a novel two-branches transformer based model to handle this issue, a pixel-wise local branch f coupled with global ISP branch g. In local branch f, we map the input image to latent feature space and replace transformer's attention block to depth-wise convolution for light-weight design. In global branch g, we use transformer's attention queries to control and adjust the global ISP-related parameters (*i.e.* colour transform matrix, gamma value). In addition, the learned queries could dynamic change under different light condition at same time (*i.e.* over-exposure and under exposure).

Extensive experiments are conducted on several real-world and synthetic datasets, *i.e.*, image enhancement dataset LOL [60] and photo retouching dataset MIT-Adobe FiveK [5], low-light detection dataset EXDark [38] and low-light segmentation dataset ACDC [49]. Results show that our IAT achieve state-of-the-art performance across a range of low-level and high-level tasks. More importantly, our IAT model is of only **0.09M** parameters, much smaller than current SOTA transformer-based models [8, 55, 59]. Besides, our average inference speed is **0.004s** per image, faster than the SOTA methods taking around 1s per image.

Our contribution could be summarised as follow:

- We have proposed a fast and light-weight framework, Illumination Adaptive Transformer (IAT), to adapt to challenging light conditions in the real world, which could both handle the low-light enhancement and exposure correction tasks.
- We have proposed a novel transformer-style structure to estimate ISP-related parameters to fuse the target sRGB image, wherein the learnable attention quires are utilised to attend the whole image, also we replace the layer norm to a new light normalisation, for better handling low-level vision tasks.
- Extensive experiments on several real-world datasets on 3 low-level tasks and 3 high-level tasks demonstrate the superior performance of IAT over SOTA methods. IAT is light weight and mobile-friendly with only **0.09M** model parameters and **0.004s** processing time per image. We will release the source code upon publication.

2 Related Works

2.1 Enhancement against Challenging Light Condition

Earlier low-light image enhancement solutions mainly rely on RetiNex theory [33] or histogram equalization [18, 53]. Since LLNet [39] utilised a deep-autoencoder structure, CNN based methods [17, 19, 31, 40, 42, 51, 58, 63, 64, 72] have been widely used in this task and gain SOTA results on the benchmark enhancement datasets [5, 60].

Similar to low-light enhancement, traditional exposure correction algorithms [43, 67] also use image histograms to adjust image intensities. The strategy then tends to correct exposure errors by adjusting the tone curve via a trained deep learning model [45, 66]. Very recently, Afifi *et al.* [2] propose a coarse-to-fine neural network to correct photo exposure, after that Nsampi *et al.* [44] introduce attention mechanism into this task.

Beyond low-level vision, low-light/ strong-light scenario also deteriorates the performance of high level vision [15, 25, 36, 41, 50, 73]. Several methods based on data synthesis [73], self-supervised learning [15] and domain adaptation [50] have been proposed to support high level vision tasks under challenging illumination conditions.

2.2 Vision Transformers

Transformer [57] was firstly proposed in NLP area to capture long-range dependencies by global attention. ViT [16] made the first attempt in vision task by splitting the image into tokens before sending into transformer model. Since then, Transformer based models have gained superior performances in many computer vision tasks, including image/video classification [34, 37], object detection [6, 71], semantic segmentation [62], vision-language model [46, 70] and so on.

In low-level vision area, transformer-based models has also made much progress on several sub-directions, such as image super-resolution [35], image restoration [8, 59, 68], image colorization [32] and bad weather restoration [56]. Very recently, MAXIM [55] use MLP-based model [52] in low-level vision area which also shows MLP's potential on low-level vision tasks. However, existing transformer & MLP models require much computational cost (*e.g.* 115.63M for IPT [8], 14.14M for MAXIM [55]), making it hard to implement on mobile and edge devices. Extreme lightweight of our method (0.09M) is particular important in low-level vision and computational photography.

3 Illumination Adaptive Transformer

3.1 Motivation

For a sRGB image I_i taken from light condition L_i, the input photons under light condition L_i would project through the lens on capacitor cluster, to pass by the in-camera process [61] and render with image signal processor (ISP) pipeline $G(\cdot)$ [4, 30]. Our goal is to match input sRGB I_i to the target sRGB image I_t (taken under light condition L_t). Existing deep-learning based methods tend to build an end-to-end mapping between I_i and I_t [2, 39, 40] or estimate some high-level representation to assist enhancement task (*i.e.* illumination map [58], colour transform function [31], 3D look-up table [69]). However, the actual lightness degradation happens in raw-RGB space, and the processes in camera ISP involves more elaborated non-linear operations such as white balance, colour space transform, gamma correction, *etc.* Therefore, much of research conducts image enhancement [7, 61] directly on raw-RGB data rather than sRGB images.

To this end, Brooks *et al.* [4] inverse each steps in ISP pipeline (*i.e.* gamma correction, tone mapping, camera colour transformation) to transform input sRGB image to "unprocessed" raw-RGB data. After that, Afifi and Brown [1] apply an encoder-decoder structure to edit the illumination of sRGB image from input light I_i to target light I_t as following:

$$I_t = G(F(I_i)), \tag{1}$$

where F is an unknown reconstruction function maps I_i to the corresponding raw-RGB data $D = F(I_i)$, and G is camera rendering function that transform D back to target sRGB image I_t. Here [1] use the network encoder f to represent F, before adding several individual decoders g_t upon encoder f. The function maps $f(I_i)$ to target I_t illumination conditions is represented below:

$$I_t = g_t(f(I_i)), \tag{2}$$

For the sake of lightweight network design, inspired by the DETR [6] which controls different object proposals via transformer queries, here we use different queries to control the ISP-related parameters in $g_t(\cdot)$. This re-configures parameters to make the image I_i adaptive to target light condition L_t. In training stage, the queries is dynamically updated in each iteration to match the target image I_t. Here we simplify the ISP procedures [4, 11, 15] into the equation 3 below. The simplification details could be found in the supplementary.

$$g_t(\cdot) = (max(\sum_{c_j} W_{c_i,c_j}(\cdot), \varepsilon))^{\gamma}, c_i, c_j \in \{r, g, b\}. \tag{3}$$

W_{c_i,c_j} is a 3×3 joint colour transformation matrix, considering the white balance and colour transform matrix. We adopt 9 queries to control W_{c_i,c_j}'s parameters. γ denotes the

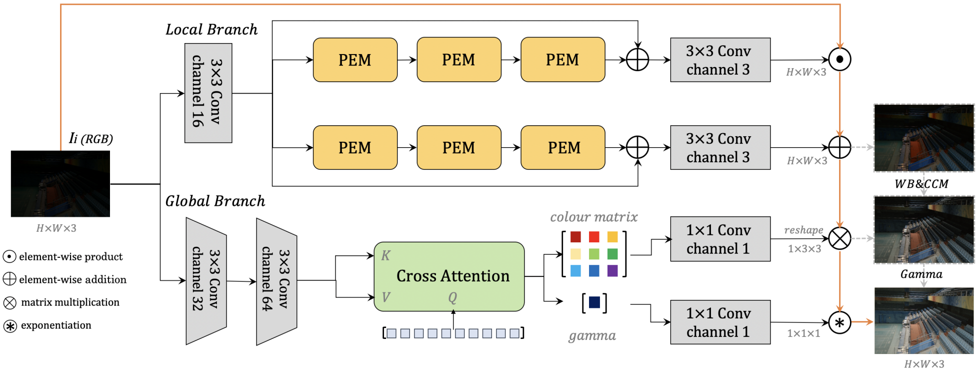


Figure 2: Structure of our Illumination Adaptive Transformer (IAT), the black line refers to the parameters generation while the yellow line refers to image processing.

gamma correction parameter which we use a single query to control. ε is a very small value to prevent numerical instability. Here we set $\varepsilon = 1e^{-8}$ in our experiments.

For process F, we apply a pixel-wise least squares model f. Our f consists of two individual branches to predict multiply map M and add map A. We then apply a least squares to process input sRGB image: $f(I_i) = I_i \odot M + A$. Here M and A has the same size with I_i to complete pixel-level multiplicative and additive adjustment. Finally, the equation of our IAT model follows:

$$I_t = (max(\sum_{c_j} W_{c_i,c_j}(I_i \odot M + A)), 0)^{\gamma}. \quad (4)$$

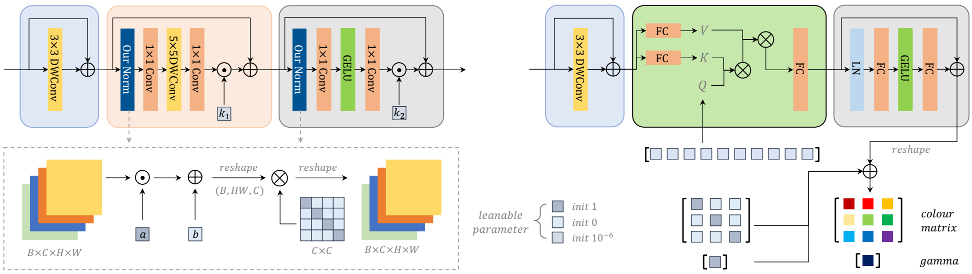
The non-linear operations are decomposed into a local pixel-wise components f and a global ISP components g. Thus, we design two individual transformer style branches: local adjustment branch and global ISP branch, to estimate the local pixel-wise components and global ISP components respectively.

3.2 Model Structure

Given an input sRGB image $I_i \in \mathbb{R}^{H \times W \times 3}$ under light condition L_i, where $H \times W$ denotes the size dimension and 3 denotes the channel dimension ($\{r, g, b\}$). As shown in Fig.2, we propose our Illumination Adaptive Transformer (IAT) to transfer the input RGB image I_i to a target RGB $I_t \in \mathbb{R}^{H \times W \times 3}$ under the proper uniform light L_t.

Local Branch. In the local branch, we focus on estimating the local components M, A to correct the effect of illumination. Instead of adopting a U-Net [48] style structure, which downsamples the images first before upsampling, we aim to maintain the input resolution through the local branch to preserve the informative details. Therefore, we propose a transformer-style architecture for the local branch. Compared to popular U-Net style structures [2, 40], our structure could deal with arbitrary resolution images without resizing them.

At first, we expand the channel dimension via a 3×3 convolution and pass them to two independent branches stacked by Pixel-wise Enhancement Module (PEM). For the lightweight design in the local branch, we replace self-attention with depth-wise convolution as suggested in the previous works [21, 34], depth-wise convolution could reduce parameters and



(a) Pixel−wise Enhancement Module (b) Global Prediction Module

Figure 3: Detailed structure of Pixel-wise Enhancement Module (PEM) and Global Prediction Module (GPM).

further save computation cost. As shown in Fig. 3 (a), our PEM first encodes the position information by 3×3 depth-wise convolution before enhancing local details with PWConv-DWConv-PWConv. Finally, we adopt two 1×1 convolutions to enhance token representation individually. Specially, we design new kind of normalisation names light normalisation, to replace transformer's Layer Normalisation [3]. As shown in Fig. 3 (a), light normalisation learns to scale a and bias b via two learnable parameters before fusing the channels via the learnable matrix. The matrix is initialised as an identity matrix. For better convergence, we adopt Layer Scale [54] which multiplies the features by a small number k_1/k_2.

We stack 3 PEMs in each branch and then connect the output features with the input features through element-wise addition. This skip connection [22] helps maintain the original image details. Finally, we decrease the channel dimension by a 3×3 convolutions and adopt ReLU/ Tanh function to generate the local components M/ A in Eq. 4.

Global ISP Branch. Global ISP branch accounts for part of the ISP pipeline [4, 23, 28, 30] (*i.e.* gamma correction, colour matrix transform, white balance) when transferring the target RGB image I_t. Specifically, the value of each pixel in the target image is determined by a global operation defined in Eq.3.

Inspired by Detection Transformer DETR [6], we design global component queries to decode and predict the W, γ to generate sRGB image I_t. This transformer structure allows capturing global interactions between context and individual pixels. As shown in Fig. 2, we first stack two convolutions as a lightweight encoder, which encodes the features in a high dimension with lower resolution, on the one hand lower resolution would save computational cost which contribute to the light-weight design, on the other hand higher feature representation would be helpful to extract image's global-level features. Then the generated features are passed to the Global Prediction Module (GPM), Fig. 3 (b) shows the detailed structure of GPM, different from original DETR [6] model, our global component queries Q are initialised as zeros without extra multi-head self-attention. Q is global component learnable embedding that attends keys K and values V generated from encoded features. The positional encoding for K and V is from a depth-wise convolution, which is friendly with different input resolutions. After feed forward network (FFN) [16] with two linear layers, we add two extra parameters with special initialisation to output colour matrix and gamma. This initialisation makes sure the colour matrix is identity matrix W and the gamma value g is one in the beginning, thus contributing to stable training.

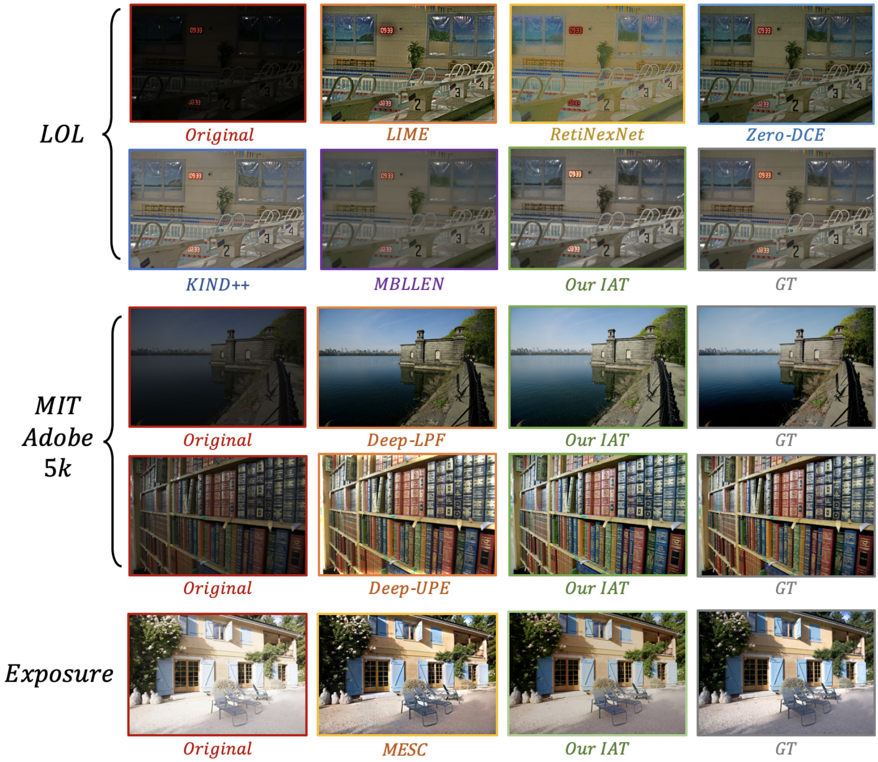


Figure 4: Results on enhancement dataset [5, 60]) and exposure correction dataset [2].

4 Experiments

We evaluate our proposed IAT model on benchmark datasets and experimental settings for both low-level and high-level vision tasks under different illumination conditions. Three low-level vision tasks include: (a) image enhancement (LOL (V1 & V2-real) [60]), (b) image enhancement (MIT-Adobe FiveK [5]), (c) exposure correction [2]. Three high-level visions tasks include: (d) low-light object detection (e) low-light semantic segmentation (f) various-light object detection. The number of PEM number in local branch are both set to 3, while the channel number in PEM is set to 16.

For all low-level vision experiments: $\{(a),(b),(c)\}$, the IAT model are trained on a single GeForce RTX 3090 GPU with batch size 8. We use Adam optimizer to train our IAT model while the initial learning rate and weight decay are separately set to $2e^{-4}$ and $1e^{-4}$. A cosine learning schedule has also been adopted to avoid over-fitting. For data augmentation, horizontal and vertical flips have been used to acquire better results.

4.1 Low-level Image Enhancement.

For (a) and (b) image enhancement task, we evaluate our IAT framework on benchmark datasets: LOL (V1 & V2-real) [60] and MIT-Adobe FiveK [5].

LOL [60] has two versions: LOL-V1 consists of 500 paired normal-light images and low-light images. 485 pairs are used for training and the other 15 pairs are for testing. LOL-

Table 1: Experimental results on LOL (V1 & V2) [60] datasets, best and second best results are marked in red and blue respectively, noted here [20] is non-deep learning method and [19] is self-supervised learning method.

Methods	LOL-V1		LOL-V2-real		Efficiency		
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	FLOPs(G)↓	#Params(M)↓	test time(s)↓
LIME* [20]	16.67	0.560	15.24	0.470	-	-	3.241 (M)
Zero-DCE* [19]	14.83	0.531	14.32	0.511	2.53	0.08	0.002 (P)
RetiNexNet [60]	16.77	0.562	18.37	0.723	587.47	0.84	0.841 (T)
MBLLEN [40]	17.90	0.702	18.00	0.715	19.95	20.47	1.981 (T)
DRBN [65]	19.55	0.746	20.13	0.820	37.79	0.58	1.210 (P)
3D-LUT [69]	16.35	0.585	17.59	0.721	7.67	0.6	0.006 (P)
KIND [72]	20.86	0.790	19.74	0.761	356.72	8.16	0e38 (T)
UFormer [59]	16.36	0.771	18.82	0.771	12.00	5.29	0.248 (P)
IPT [8]	16.27	0.504	19.80	0.813	2087.35	115.63	1.365 (P)
RCT [31]	22.67	0.788	-	-	-	-	-
MAXIM [55]	23.43	0.863	22.86	0.818	216.00	14.14	0.602 (P)
IAT (local)	20.20	0.782	20.30	0.789	1.31	0.02	0.002 (P)
IAT	23.38	0.809	23.50	0.824	1.44	0.09	0.004 (P)

Table 2: Experimental results on MIT-Adobe FiveK [5] dataset.

Metric	White-Box [26]	U-Net [48]	DPE [12]	DPED [27]	D-UPE [58]	D-LPF [42]	3D LUT [69]	IAT
PSNR↑	18.57	21.57	23.80	21.76	23.04	23.63	25.21	**25.32**
SSIM↑	0.701	0.843	0.880	0.871	0.893	0.875	**0.922**	0.920
#Params.↓	-	1.3M	3.3M	-	1.0M	0.8M	0.6M	**0.09M**

V2-real consists of 789 paired normal-light images and low-light pairs. 689 pairs are used for training and the other 100 pairs are for testing. The loss function between input image I_i and target image I_t for LOL dataset training is a mixed loss function [56] consisting of smooth L1 loss and VGG loss [29]. In LOL-V1 training, the images are cropped into 256×256 to train 200 epochs and then fine-tune on 600×400 resolution for 100 epochs. In LOL-V2-real training, the image resolution is maintained at 600×400 and trained for 200 epochs. Both LOL-V1 and LOL-V2-real testing the image resolution is maintained at 600×400. We compare our method with SOTA methods [8, 19, 20, 31, 40, 55, 59, 60, 65, 69, 72]. For image quality analysis, we evaluate the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). For the model efficiency analyze, we report three metrics: FLOPs, model parameters and test time, as shown in the last column of Table.1. We list different model's test time on their corresponding code platform (M means Matlab, T means TensorFlow, P means PyTorch). As shown in Table 1, **IAT (local)** only uses the local network to train the model and **IAT** refers to the whole framework. Our **IAT** gains SOTA result on both image quality and model efficiency, especially less than 100× FLOPs and parameters usage compare to the current SOTA methods MAXIM [55].

MIT-Adobe FiveK [5] dataset contains 5000 images, each was manually enhanced by five different experts (A/B/C/D/E). Following the previous settings [42, 58], we only use experts C's adjusted images as ground truth images. For MIT-Adobe FiveK [5] dataset training, we use a single L1 loss function to optimize IAT model. Our method is compared with SOTA enhancement methods [12, 26, 27, 42, 48, 58, 58, 69] on FiveK dataset. The image quality results (PSNR, SSIM) and model parameters are reported in Table. 2. Our **IAT** also gain satisfactory result in both quality and efficiency. Qualitative results of LOL [60] and FiveK [5] has been shown in Fig.4. More results could be found in supplementary material.

Table 3: Experimental results on exposure correction dataset [2]. Note here HE and LIME [20] are non-deep learning methods. PSNR, SSIM and PI results, reported by competing works, are from [2].

Method	Expert A		Expert B		Expert C		Expert D		Expert E		Avg		PI↓
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	
HE* [18]	16.14	0.685	16.28	0.671	16.52	0.696	16.63	0.668	17.30	0.688	16.58	0.682	2.405
LIME* [20]	11.15	0.590	11.83	0.610	11.52	0.607	12.64	0.628	13.61	0.653	12.15	0.618	2.432
DPED [27] (Sony)	17.42	0.675	18.64	0.701	18.02	0.683	17.55	0.660	17.78	0.663	17.88	0.676	2.806
DPE [12] (S-FiveK)	16.93	0.678	17.70	0.668	17.74	0.696	17.57	0.674	17.60	0.670	17.51	0.677	2.621
RetinexNet [60]	10.76	0.585	11.61	0.596	11.13	0.605	11.99	0.615	12.67	0.636	11.63	0.607	3.105
Deep-UPE [58]	13.16	0.610	13.90	0.642	13.69	0.632	14.80	0.649	15.68	0.667	14.25	0.640	2.405
Zero-DCE [19]	11.64	0.536	12.56	0.539	12.06	0.544	12.96	0.548	13.77	0.580	12.60	0.549	2.865
MSEC [2]	19.16	0.746	20.10	0.734	20.20	0.769	18.98	0.719	18.98	0.727	19.48	0.739	2.251
IAT (local)	16.61	0.750	17.52	0.822	16.95	0.780	17.02	0.773	16.43	0.789	16.91	0.783	2.401
IAT	**19.90**	**0.817**	**21.65**	**0.867**	**21.23**	**0.850**	**19.86**	**0.844**	**19.34**	**0.840**	**20.34**	**0.844**	**2.249**

Table 4: Experimental results on low-light detection dataset EXDark [38], low-light semantic segmentation dataset ACDC [49] and various light detection dataset TYOL [24].

Methods	(d) EXDark Detection [38]		(e) ACDC Segmentation [49]		(f) TYOL Detection [24]	
	mAP↑	time(s)↓	mIOU↑	time(s)↓	mAP↑	time(s)↓
base-line	76.4	0.033	**63.3**	0.249	88.4	0.023
MBLLEN [40]	76.3	0.086	63.0	0.332	95.3	0.105
DeepLPF [42]	76.3	0.138	61.9	0.807	94.5	0.223
Zero-DCE [19]	76.9	0.042	61.9	0.300	95.2	0.030
IAT	**77.2**	**0.040**	62.1	**0.280**	**95.8**	**0.027**

4.2 Exposure Correction.

For the (c) exposure correction task, we evaluate IAT on the benchmark dataset proposed by [2]. The dataset contains 24,330 sRGB images, divided into 17,675 training images, 750 validation images, and 5905 test images. Images in [2] are adjusted by MIT-Adobe FiveK [5] dataset with 5 different exposure values (EV), ranging from under-exposure to over-exposure condition. Same as [5], test set has 5 different experts' adjust results (A/B/C/D/E). Following the setting of [2], the training images are cropped to 512×512 patches and the test image is resized to have a maximum dimension of 512 pixels. We compare the test images with all five experts' results. Here we use L1 loss function for exposure correction training.

The evaluation result is shown in Table. 3. Our comparison methods include both traditional image processing methods (Histogram Equalization [18], LIME [20]) and deep learning methods (DPED [27], DPE [12], RetinexNet [60], Deep-UPE [58], Zero-DCE [19], MSEC [2]). Evaluation metrics are same as [2], including PSNR, SSIM and perceptual index (PI). Table. 3 shows that our **IAT** model has gained best result on all evaluation indexs. Compared to the second best result MSEC [2], IAT has much fewer parameters (0.09M v.s. 7M) and less evaluation time (0.004s per image v.s. 0.5s per image). Qualitative result has been shown in Fig.4 and more visual results are given in supplementary material.

4.3 High-level Vision

For high-level vision tasks: $\{(d),(e),(f)\}$, we use IAT to restore the image before feeding to the subsequent recognition algorithms based on mmdetection and mmsegmentation [9, 13]. For a fair comparison, we run all of the experiments in the same setting: same input size, same data augmentation methods (expand, random crop, multi-size, random flip...), same training epochs and same initial weights. We train the recognition algorithm on the datasets enhanced by IAT. We compare our methods with original datasets as well as datasets enhanced by other enhancement methods [19, 40, 42].

For object detection task in (d) EXDark dataset [38] and (f) TYOL dataset [24]. EXDark includes 7,363 real-world low-light images, ranging from twilight to extreme dark environment with 12 object categories. We take 80% images of each category for training and the other 20% for testing. TYOL includes 1680 images with 21 classes. We take 1365 images for training and other for evaluation. For both datasets, we perform object detection with YOLO-V3 [47], all the input images have been cropped and resized to 608×608 pixel size, we use SGD optimizer to train YOLO-V3 with batch size 8 for 25 epochs to EXDark and 45 epochs to TYO-L, the initial learning rate is $1e^{-3}$ and weight decay is $1e^{-4}$. The detection metric mAP and per-image evaluation time is shown in Table. 4. Our IAT model gains best results in both accuracy and speed compared to the baseline model and other enhancement methods [19, 40, 42].

For semantic segmentation in (e) ACDC dataset [49], we take 1006 night images in the ACDC dataset and then adopt DeepLab-V3+ [10] to train on the ACDC-night train set and test on ACDC-night val set. The DeepLab-V3+ [10] model is initialed by an Cityscape dataset [14] pre-train model, we tuned the pre-train model by SGD optimizer with batch size 8 for 20000 iters, initial learning rate is set to 0.05, momentum and weight decay are separately set to 0.9 and $5e^{-4}$. We show the segmentation metric mIOU and per-image evaluation time in Table. 4, we found that all the enhancements methods invalid in this setting, this may because the lightness condition in ACDC [49] is various and exceeds the generalisation ability of the enhancement model. For this problem, we propose to joint training our IAT model with following segmentation network (as well as detection network), which would solves this problem and improve the semantic segmentation/ object detection results in low-light conditions, detailed analyse please refer to Sec.B of supplementary material [1].

5 Conclusion

We propose a novel lightweight transformer framework IAT, by adapting ISP-related parameters to adapt to challenging light conditions. Despite its superior performance on several real-world datasets for both low-level and high-level tasks, IAT is extremely light with a fast speed. The lightweight and mobile-friendly IAT has the potential to become a standing tool for the computer vision community.

However, one mian drawback of the IAT module is that, the image signal processor (ISP) has been simplified due to the light-weight demand, we think that more detailed ISP-related parts could be concerned and interpolate to the IAT module. In further, we'd also like to implement IAT on 3D human relighting task, to solve more complex lighting problems under 3D condition.

6 Acknowledgement

Corner symbol '*' in the author name means the corresponding author. This work supported by JST Moonshot R&D Grant Number JPMJMS2011 and JST, ACT-X Grant Number JP-MJAX190D, Japan. This work also supported by National Natural Science Foundation of China (Grant No. 62206272) and Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

[1] For more experimental details and ablation analyse, please refer to the supplementary material.

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

[4] 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.

[5] 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.

[6] 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.

[7] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to see in the dark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[8] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[9] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[10] 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.

[11] 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.

[12] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[13] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] 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.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[17] Ying Fu, Yang Hong, Linwei Chen, and Shaodi You. Le-gan: Unsupervised low-light image enhancement network using attention module and identity invariant loss. *Knowledge-Based Systems*, 2022.

[18] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., USA, 2006. ISBN 013168728X.

[19] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong. Zero-reference deep curve estimation for low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[20] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 2017.

[21] Qi Han, Zejia Fan, Qi Dai, Lei Sun, Ming-Ming Cheng, Jiaying Liu, and Jingdong Wang. On the connection between local attention and dynamic depth-wise convolution. In *International Conference on Learning Representations*, 2022.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[23] Felix Heide and Steinberger et.al. Flexisp: A flexible camera image processing framework. *ACM Trans. Graph.*, 2014.

[24] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European Conference on Computer Vision*, 2018.

[25] Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *The British Machine Vision Conference*, November 2021.

[26] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Trans. Graph.*, 2018.

[27] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2017.

[28] H. Jiang, Q. Tian, J. Farrell, and B. A. Wandell. Learning the image processing pipeline. *IEEE Transactions on Image Processing*, 2017.

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

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

[31] 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.

[32] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *International Conference on Learning Representations*, 2021.

[33] Edwin H. Land. An alternative technique for the computation of the designator in the retinex theory of color vision. *Proceedings of the National Academy of Sciences of the United States of America*, 1986.

[34] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022.

[35] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE International Conference on Computer Vision Workshops*, 2021.

[36] Jiaying Liu, Dejia Xu, Wenhan Yang, Minhao Fan, and Haofeng Huang. Benchmarking low-light image enhancement and beyond. *International Journal of Computer Vision*, 2021.

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

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

[39] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 2017.

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

[41] Luca Minciullo, Fabian Manhardt, Kei Yoshikawa, Sven Meier, Federico Tombari, and Norimasa Kobori. Db-gan: Boosting object recognition under strong lighting conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[42] 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.

[43] Nayar and Branzoi. Adaptive dynamic range imaging: optical control of pixel exposures over space and time. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1168–1175 vol.2, 2003. doi: 10.1109/ICCV.2003.1238624.

[44] Ntumba Elie Nsampi, Zhongyun Hu, and Qing Wang. Learning exposure correction via consistency modeling. 2021.

[45] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

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

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[49] 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.

[50] Yukihiro Sasagawa and Hajime Nagahara. Yolo in the dark: Domain adaptation method for merging multiple models. In *Proceedings of European Conference on Computer Vision*, 2020.

[51] Aashish Sharma and Robby T. Tan. Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11972–11981, 2021. doi: 10.1109/CVPR46437.2021.01180.

[52] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24261–24272. Curran Associates, Inc.,

2021. URL https://proceedings.neurips.cc/paper/2021/file/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Paper.pdf.

[53] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, 1998.

[54] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[55] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. *CVPR*, 2022.

[56] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. *arXiv preprint arXiv:2111.14813*, 2021.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[58] 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.

[59] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

[61] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[62] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

[63] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17714–17724, June 2022.

[64] Kai-Fu Yang, Cheng Cheng, Shi-Xuan Zhao, Xian-Shi Zhang, and Yong-Jie Li. Learning to adapt to light, 2022. URL https://arxiv.org/abs/2202.08098.

[65] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[66] Runsheng Yu, Wenyu Liu, Yasen Zhang, Zhi Qu, Deli Zhao, and Bo Zhang. Deepexposure: Learning to expose photos with asynchronously reinforced adversarial learning. In *Advances in Neural Information Processing Systems*, 2018.

[67] Lu Yuan and Jian Sun. Automatic exposure correction of consumer photographs. In *European Conference on Computer Vision*, 2012.

[68] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021.

[69] 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*, pages 1–1, 2020. doi: 10.1109/TPAMI.2020.3026740.

[70] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[71] Renrui Zhang, Han Qiu, Tai Wang, Xuanzhuo Xu, Ziyu Guo, Yu Qiao, Peng Gao, and Hongsheng Li. Monodetr: Depth-aware transformer for monocular 3d object detection. *arXiv preprint arXiv:2203.13310*, 2022.

[72] 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.

[73] Yinqiang Zheng, Mingfang Zhang, and Feng Lu. Optical flow in the dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6749–6757, 2020.