Real-Time Exposure Correction via Collaborative Transformations and Adaptive Sampling

Ziwen Li[1], Feng Zhang[1], Meng Cao[2], Jinpu Zhang[1], Yuanjie Shao[3], Yuehuan Wang[1]*, Nong Sang[1]
[1]National Key Laboratory of Multispectral Information Intelligent Processing Technology,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Mohamed bin Zayed University of Artificial Intelligence
[3]School of Electronic Information and Communication, Huazhong University of Science and Technology
{D201980722,fengzhangaia,shaoyuanjie,yuehwang,nsang}@hust.edu.cn, {mengcaopku,zjphust}@gmail.com

Abstract

Most of the previous exposure correction methods learn dense pixel-wise transformations to achieve promising results, but consume huge computational resources. Recently, Learnable 3D lookup tables (3D LUTs) have demonstrated impressive performance and efficiency for image enhancement. However, these methods can only perform global transformations and fail to finely manipulate local regions. Moreover, they uniformly downsample the input image, which loses the rich color information and limits the learning of color transformation capabilities. In this paper, we present a collaborative transformation framework (CoTF) for real-time exposure correction, which integrates global transformation with pixel-wise transformations in an efficient manner. Specifically, the global transformation adjusts the overall appearance using image-adaptive 3D LUTs to provide decent global contrast and sharp details, while the pixel transformation compensates for local context. Then, a relation-aware modulation module is designed to combine these two components effectively. In addition, we propose an adaptive sampling strategy to preserve more color information by predicting the sampling intervals, thus providing higher quality input data for the learning of 3D LUTs. Extensive experiments demonstrate that our method can process high-resolution images in real-time on GPUs while achieving comparable performance against current state-of-the-art methods. The code is available at https://github.com/HUST-IAL/CoTF.

1. Introduction

Exposure correction [1] is a fundamental problem in the field of computational photography and computer vision, and has been extensively studied over the last few decades. Its purpose is to automatically correct over- or underexposed images taken under undesirable lighting conditions. Exposure correction plays an important role in many applications such as autonomous driving [14] and video understanding [3–8].

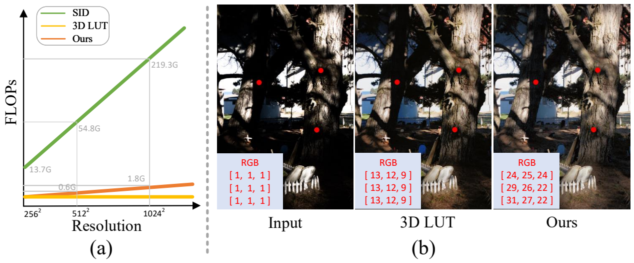


Figure 1. Comparison of different transformation methods. (a) shows the computational effort of the different methods. We can see that the computational effort of the pixel transformation method SID increases significantly with resolution, while our method remains efficient at high resolution. (b) shows the pixel mapping relations for different transformations. 3D LUT performs a fixed global transformation based on pixel values, resulting in some unsatisfactory local contrast. While our method considers the pixel context and yields favorable results.

Recently, with the rapid development of deep learning, many learning-based exposure correction methods [1, 15, 16, 18, 36] have been proposed and achieved promising performance. However, most of them elaborate complex network structures and learn dense pixel-wise transformations with the computational burden proportional to the resolution of the input image. This leads to the huge computations and endure the curse of dimensionality, when confronted with high-resolution images. For example, in Figure 1(a), when processing 256 × 256 images, SID [9] requires only 13.7G FLOPs of floating point operations. When the image

[*] Corresponding author

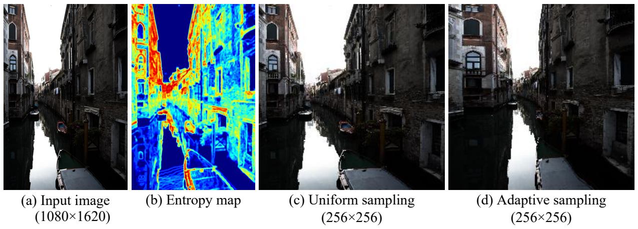

(a) Input image (1080×1620)
(b) Entropy map
(c) Uniform sampling (256×256)
(d) Adaptive sampling (256×256)

Figure 2. Illustration of non-uniform distribution of color information. We use local entropy to reflect the richness of colors. From the entropy map, we can see that the color information is unevenly distributed. In addition, we show that our adaptive sampling retains more color information than uniform sampling.

scale raise to 1024×1024, the computation demand rises to 219.3G FLOPs accordingly, which is unacceptable for real-world and real-time applications on mobile devices. Thus, our goal is to design a network that achieves comparable enhancement quality but with fast speeds and low operations.

An intuitive way to reduce the cost of pixel-wise transformations is to apply a downsampling step before enhancement, but this may lose high-frequency detail and lead to blurring effects. Another efficient alternative is to resort to the global transformation, such as 3D Lookup Table (3D LUT). The 3D LUT simulates arbitrary nonlinear functions by predicting the control points of a curve, providing strong color translation capabilities. It is also an efficient data structure that replaces complex calculations with fast lookup operations. Recently, some excellent works [33, 42, 49] have utilized neural networks to predict image-adaptive 3D LUTs, which achieved pleasing image quality and efficient computation. However, as pointed out by [49], the 3D LUT performs global transformations and ignores the local context of each pixel, and thus cannot finely manipulate the pixel transformations in local regions, leading to globally sub-optimal enhancement results. As shown in Figure 1(b), the mapping ability of 3D LUT is limited by the fixed transformation of pixel values, which produces some unsatisfactory contrast in local areas.

Moreover, existing learnable 3D LUT methods use bilinear downsampling to reduce the image resolution to save computation resources. This way samples pixels uniformly according to a predetermined fixed position without considering the adaptation to the image content. Indeed, it is easy to observe that the distribution of color information in an image is *spatially inhomogeneous*. As shown in Figure 2(b), we use local entropy to visualize the color distribution in an image. The sky region is flat with less color information, while the building region is colorful. Uniform sampling in colorful regions may lose useful color information, while color redundancy occurs in flat regions. The color conversion capability of a learnable 3D LUT depends heavily on the color information in the input image, so uniform downsampling can limit the learning of the 3D LUT and lead to subsequent performance degradation.

To alleviate the above problems, in this work, we propose a collaborative transformations framework (CoTF) for real-time exposure correction that integrates global transformation and pixel-wise transformation in an efficient way. The idea is inspired by professional retouchers, who usually make global adjustments and then fine-tune local areas. In particular, the global transformation predicts image-adaptive 3D LUTs to adjust the appearance holistically, obtaining decent global contrast and sharp edge details. The pixel-wise transformation uses an encoder-decoder to extract the fine-grained local context, which can be performed at low resolution to ensure efficiency. In order to effectively combine these two components, we design a relation-aware modulation (RAM) module to compensate local contrast information for global transformation results via cross-resolution interaction. In addition, we propose an adaptive sampling strategy to retain more color information during downsampling. It predicts an image-adaptive sampling grid, which enables dense sampling in colorful regions and sparse sampling in flat regions, as shown in Figure 2(d). In this way, we provide higher-quality input data for the learning of 3D LUTs to enhance color transformation capability. Benefiting from the above design, our approach is able to achieve a good balance between performance and efficiency, which can process high-resolution images in real time on GPUs. Extensive experiments on several exposure correction datasets demonstrate that our method outperforms the state-of-the-art methods both qualitatively and quantitatively.

Overall, our main contributions are as follows:

- We present a collaborative transformations framework (CoTF) that integrates the advantages of global transformation and pixel-wise transformation in an efficient manner. In addition, we design a relation-aware modulation module to modulate global transformation results with local context via cross-resolution interaction.
- We propose an adaptive sampling strategy to retain more color information and provide more higher-quality input data for the learning of 3D LUTs.
- Extensive experiments demonstrate that the proposed method outperforms the existing state-of-the-art methods on performance and efficiency.

2. Related Work

2.1. Exposure Correction

Exposure correction methods can be broadly categorized into traditional and learning-based methods. Traditional methods use histogram equalization [29, 56], curve mapping [47] and Retinex models [10, 13, 23, 32] to adjust contrast and brightness. However, these methods rely on

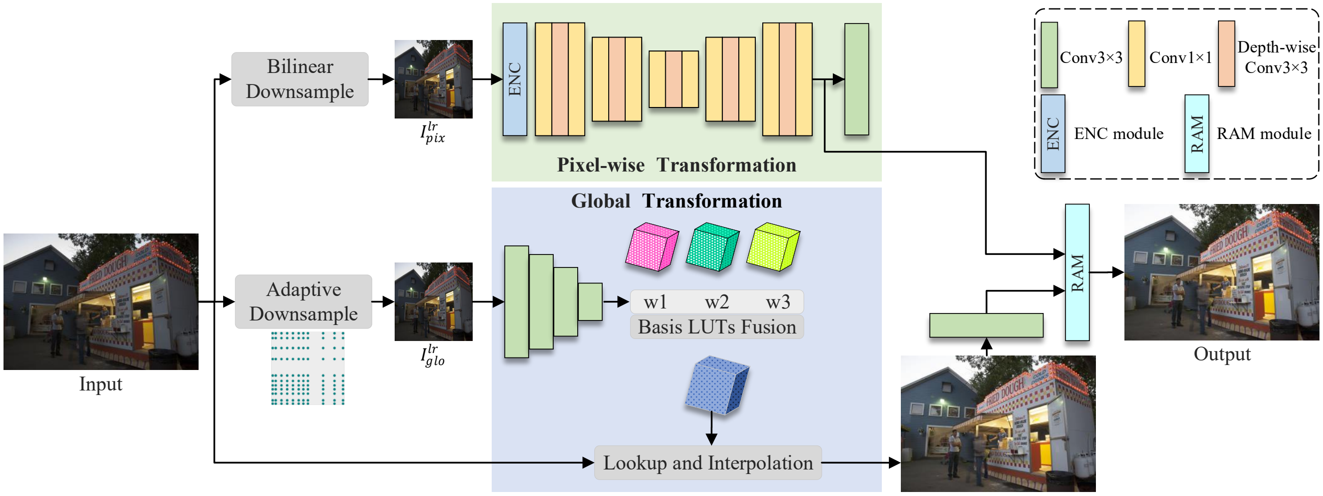


Figure 3. Illustration of proposed collaborative transformations framework (CoTF). It consists of three main components, 1) the global transformation based on the learnable 3D LUT, 2) the pixel-wise transformation in low resolution space and 3) the relation-aware modulation module to perform cross-resolution interactions. For the global transformation, we design an adaptive sampling strategy to provide higher quality input data.

hand-crafted priors and may not be robust enough to tackle complex scenarios.

Learning-based methods are rapidly evolving due to the powerful learning capabilities of deep neural networks. Some methods combine neural networks with physical models, including Retienx models [26, 31, 38, 39, 46, 52, 53, 55] and curve mapping [12, 22]. For example, RetinexNet [38] introduces subnets to decompose illumination and reflection components and enhance them separately. ZeroDCE [12] proposes a higher-order pixel-wise curve to enhance underexposed images. Another class of methods [9, 19, 34, 35, 40, 41, 44, 45, 54] learns the pixel mapping relationship between degraded and clear images. For example, DRBN [44] proposes to decompose images into different bands and recombine them under perceptual guidance. However, these methods mainly focus on enhancing underexposed images, ignoring various exposure scenes in practical applications.

Recently, some works have built a single model to correct both overexposed and underexposed images. Afifi et al. [1] presented a large-scale dataset and designed a multi-scale Laplace pyramid network. To reduce the representation gap across exposures, CMEC [28], ENC [15], and ECLNet [17] map features to exposure-invariant space. Huang et al. [16] propose a Fourier-based network with complementary interactions in the spatial and frequency domains. Wang et al. [30] proposed local color distributions to deal with non-uniform illumination. Wang et al. [36] proposed decoupling contrast enhancement and detail restoration in convolutional operations. Huang et al. [18] proposed to learn the sample relations and perform joint optimization in a mini-batch. CuDi [21] proposes curve distillation to extract knowledge from large curve-based teacher networks. CLIP-LIT [24] proposes a prompt learning framework including prompt initialization, enhancement network training and prompt refinement. Unlike these methods that use only pixel-wise transformations, in this work, we effectively unify the global transformation and pixel-wise transformations in a framework that is flexible and scalable for high-resolution images.

2.2. Lookup Tables

3D LUTs enable efficient color mapping and are widely used in camera imaging pipelines and photo editing software. Recently, learnable LUT methods [25, 43, 49, 50] have sprung up for image enhancement. Zeng et al. [49] were the first to propose image-adaptive 3D LUTs, which consists of several basic 3D LUTs and adaptive weights. Wang et al. [33] proposed spatial-aware 3D LUT considering spatial information. AdaINT [42] learns adaptive intervals to achieve more flexible sample point allocation for 3D LUT. However, all these methods use bilinear downsampling to reduce resolution, which is a kind of uniform sampling that may lose rich color information. In contrast, our proposed adaptive sampling strategy is able to retain more color information at a given size. In addition, unlike these methods that only utilize the LUTs, we efficiently integrate LUT-based global transformation and pixel-wise transformation to correct exposure collaboratively.

3. Method

The pipeline for CoTF is shown in Figure 3. We first downsample the high-resolution image I^{hr} to obtain I^{lr}_{glo} and I^{lr}_{pix}

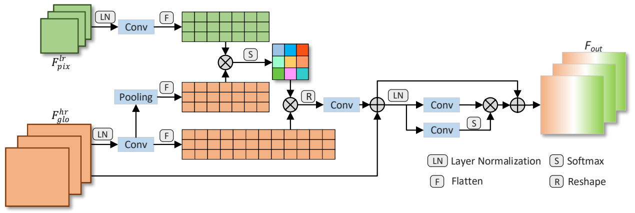


Figure 4. Illustration of the Relation-Aware Modulation (RAM) module, which modulates global transformation results with local contexts via cross-resolution interactions.

to perform global and pixel-wise transformation in the low-resolution space to reduce the computational complexity and memory burden. The global transformation predicted at low resolution can be flexibly scaled to high-resolution images, while the low-resolution pixel transformation focuses on low-frequency local context. After the global and pixel-wise transformations, we design a cross-resolution RAM module that compensates the global transformation results with fine-grained local context.

3.1. Global Transformation

As a typical global transformations tool, 3D LUT can flexibly express nonlinear mappings and adjust attributes, such as lighting, hue and saturation. A 3D LUT can be represented as a 3D array of size N^3 that discretizes each dimension of the RGB color space into N bins, where index-value pairs are used as input-output pairs. When transforming, 3D LUT use the color (r, g, b) of the input pixel as an index to look up the nearest neighbor point, and then compute the transformed color using trilinear interpolation.

Since different exposures (e.g., under- and overexposure) require different 3D LUTs, we utilize multiple 3D LUTs $\{T_m\}_{m=1}^{M}$ to handle various lighting conditions. We follow the practice of [49] to adaptively fuse these 3D LUTs. It contains two sub-mappings, one for predicting the basis 3D LUTs, and the other for learning content-dependent weights $\{w_m\}_{m=1}^{M}$. These basis 3D LUTs are linearly combined with adaptive weights to obtain the image-adaptive 3D LUT, which flexibly covers the transformation space from different exposures to normal exposures. We use the learned 3D LUT to transform high-resolution inputs I^{hr} to yield globally enhanced results $\hat{I}_{glo}^{hr}$ with sharp details.

3.2. Pixel-wise Transformation

The color conversion of 3D LUTs works only on pixel values, which may lead to undesirable results in local areas. In contrast, pixel-wise transformation adjusts pixels with reference to the local context. Since 3D LUT preserves high-frequency details well, pixel-wise transformation focuses only on low-frequency content and can be performed at low resolutions to reduce computational burden.

We employ an encoder-decoder consisting of pointwise and depthwise convolutions to accomplish the pixel transformation. Besides, we introduce a simplified ENC module [15] to reduce the discrepancy between different exposure features F_{ex}, which is expressed as:

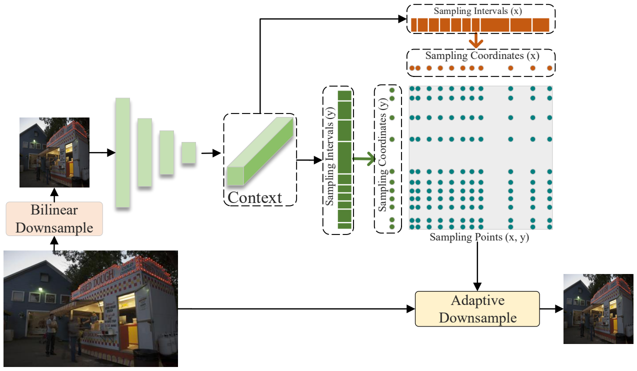


Figure 5. Illustration of adaptive sampling strategy. We first use a small CNN to extract the context of a low-resolution image. Then we learn the sampling interval and convert it into sampling coordinates to adaptively downsample the original image.

$$\hat{F}_{ex} = [IN(F_{ex}), F_{ex}], \tag{1}$$

$$\widetilde{F}_{ex} = Sigmoid(FC(GAP(\hat{F}_{ex}))) \cdot \hat{F}_{ex}, \tag{2}$$

where $[\cdot]$, IN, and GAP denote concatenation, instance normalization, and global average pooling, respectively. After pixel-wise transformation, we can obtain a low resolution result $\hat{I}_{pix}^{lr}$ with good local contrast.

3.3. Relation-Aware Modulation

After global and pixel-wise transformation, we obtain a high-resolution result $\hat{I}_{glo}^{hr}$ and a low-resolution result $\hat{I}_{pix}^{lr}$, which are complementary inherently. However, these two transformations share inconsistent resolutions and characteristics. Therefore, directly upsampling $\hat{I}_{pix}^{lr}$ and then simply blending it with $\hat{I}_{glo}^{hr}$ can lead to blurring effects and sub-optimal performance.

To address this issue, we design a lightweight Relation-Aware Modulation (RAM) module that modulates global transformation results with local contexts via cross-resolution interactions, as depicted in Figure 4. To avoid the loss of local context information and repeated extraction of features, we use the last feature map F_{pix}^{lr} of the encoder-decoder instead of the image $\hat{I}_{pix}^{lr}$. We use a convolution with kernel size 3×3 to extract the features F_{glo}^{hr} of the $\hat{I}_{glo}^{hr}$ and expand the channel dimensions to be consistent with F_{pix}^{lr}. Subsequently, we pool F_{glo}^{hr} and compute the cross-attention with F_{pix}^{lr} to obtain the relation map A, which reflects their information relationship. This operation is defined as:

$$A = Softmax(Pooling(F_{glo}^{hr}) \times F_{pix}^{lr}). \tag{3}$$

Inspired by [48], we compute it along the channel dimension to reduce complexity. Then we use A to modulate the features F_{glo}^{hr} to dynamically aggregate local contexts, *i.e*.

$$F_{out} = FFN(F_{glo}^{hr} \times A + F_{glo}^{hr}), \tag{4}$$

where we learn the residuals to stabilize the training and use feed-forward networks (FFN) to obtain a better feature representation.

3.4. Adaptive Sampling

The previous learnable 3D LUT methods use uniform sampling to reduce the image resolution, which limits the learning of 3D LUTs. To address this issue, in this work, we propose an adaptive sampling strategy to preserve more color information during downsampling.

A naive way to perform adaptive sampling is to learn the sampling coordinates directly, but this is hard to optimize because it is non-differentiable. We add two constraints: 1) sampling covers the entire spatial range $(0, 1)$, and 2) maintaining monotonically incrementality of the coordinates. In this way, we can learn the sampling intervals instead of learning the coordinates directly. Note that here we choose the horizontal and vertical directions (denoted by X and Y) as two separate sampling directions.

As shown in Figure 5, we employ a lightweight CNN to learn sampling intervals at low resolution, which preserves the original color distribution. Assuming a given sample size of $K_x \times K_y$, *i.e*., there are K_x and K_y sampling points along the X and Y directions, respectively, which means that we need to learn $K_{\{x,y\}} - 1$ sampling intervals $P_{\{x,y\}}$ in each direction. Next, we use `Softmax` to normalize the interval, ensuring that the samples cover the entire image without exceeding the range. Subsequently, we convert the $K_{\{x,y\}} - 1$ normalized sampling intervals $\hat{P}_{\{x,y\}}$ into K sampling points $Q_{\{x,y\}}$ via an accumulation operation. Since the value of each interval is positive, the accumulation operation ensures monotonic increment of the coordinates.

Finally, the sampling grid G is obtained by computing the Cartesian product of the X- and Y-direction coordinates, which is denoted as $G = Q_x \otimes Q_y = \{(Q_{x,i}, Q_{y,j}) | i \in \{1, 2, ..., K_x\}, j \in \{1, 2, ..., K_y\}\}$. We downsample the original image by applying the sampling grid, which adapts to the image content. Compared with uniform sampling, adaptive sampling is a superior strategy to densely sample colorful regions and sparsely sample flat regions. In this way, more color information can be retained during downsampling, which provides higher quality data and thus improves the color translation capability of the 3D LUTs.

It is worth noting that we only applied the adaptive sampling strategy to the global transformation. This is because the 3D LUT is a spatially independent model, *i.e*., the color transform is only related to the color values and not to the position. In contrast, the pixel-wise transformation is a spatially correlated model that requires positional consistency, so we still use bilinear downsampling for it.

4. Experiments

4.1. Experimental settings

Datasets. We evaluate proposed method on three datasets, including two exposure correction datasets, (*i.e*., MSEC [1] and SICE [2]), and a non-uniform illumination dataset (*i.e*., LCDP [30]). The MSEC [1] dataset renders images using relative EVs of -1.5 to +1.5 and contains a total of 17675 training images, 750 validation images, and 5905 test images. Following the settings of [15] for SICE, we treat the second and second-last exposure levels as underexposed and overexposed images, and the middle exposure levels as ground truth. It contains 1000 training images, 24 validation images and 60 test images. The LCDP dataset exhibits non-uniform illumination due to both overexposure and underexposure occurring in single images. It contains 1415 training images, 100 validation images, and 218 test images.

Implementation Details. We use the small CNN in [49] as a backbone for global transformation and adaptive sampling, which contains only 5 convolutional layers. We set 3 basis 3D LUTs, with the dimension of each LUT set to 3×17^3. We initialize the first 3D LUT as a identity mapping and the others as zero mappings. The sampling grid is initialized to a uniform state. Consistent with [15], we use L1 loss, L_1, perceptual loss, L_{per} and SSIM loss, L_{ssim} to train the network, which is expressed as $L_{total} = L_1 + \beta_1 L_{per} + \beta_2 L_{ssim}$, where the coefficients β_1 and β_1 are empirically set to 0.1 and 0.5, respectively.

During training, we use the ADAM [20] optimizer to minimize L_{total} in an end-to-end manner. The mini batch size is set to 2. We set the initial learning rate to $4e^{-4}$ and update it using the cosine annealing strategy. For adaptive sampling, the learning rate is decayed by 0.1 to stabilize the training. We downsample the image to 256 × 256 to feed the network. For MSEC, SICE, and LCDP datasets, the training process consists of 50, 200 and 200 epochs, respectively. Our models are implemented using Pytorch and run on NVIDIA TITAN V GPUs.

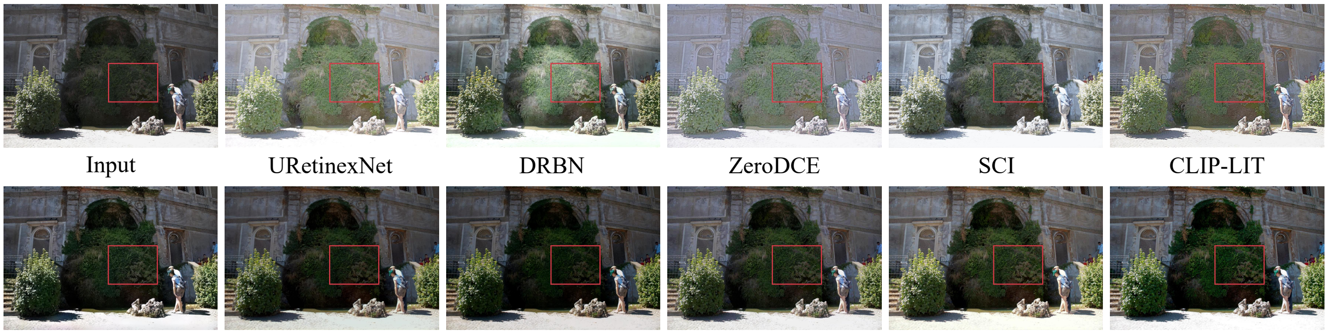
4.2. Comparison with State-of-the-Art Methods

We use PSNR, SSIM [37] and LPIPS [51] metrics for performance evaluation, as well as parameters, FLOPs and inference times for efficiency evaluation.

Quantitative Comparisons. Table 1 reports the quantitative results on the MSCE and SCIE datasets. We can see that our method has the best overall performance. On the MSEC dataset, our method has the best performance with 23.44dB PSNR, 0.8728 SSIM and 0.1232 LPIPS. On the SICE

Methods	MSEC							SICE						
	Under		Over		Average			Under		Over		Average		
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	LPIPS↓
HE [29]	16.52	0.6918	16.53	0.6991	16.53	0.6959	0.2920	14.69	0.5651	12.87	0.4991	13.78	0.5376	0.3738
CLAHE [56]	16.77	0.6211	14.45	0.5842	15.38	0.5990	0.4744	12.69	0.5037	10.21	0.4847	11.45	0.4942	0.4688
LIME [13]	13.98	0.6630	9.88	0.5700	11.52	0.6070	0.2758	16.48	0.5832	6.67	0.4041	11.58	0.4937	0.3712
WVM [10]	18.67	0.7280	12.75	0.645	15.12	0.6780	0.2284	15.16	0.5915	8.03	0.4485	11.60	0.5200	0.3432
RetinexNet [38]	12.13	0.6209	10.47	0.5953	11.14	0.6048	0.3209	12.94	0.5171	12.87	0.5252	12.90	0.5212	0.4312
URetinexNet [39]	13.85	0.7371	9.81	0.6733	11.42	0.6988	0.2858	17.39	0.6448	7.40	0.4543	12.40	0.5496	0.3549
DRBN [44]	19.74	0.8290	19.37	0.8321	19.52	0.8309	0.2795	17.96	0.6767	17.33	0.6828	17.65	0.6798	0.3891
SID [9]	19.37	0.8103	18.83	0.8055	19.04	0.8074	0.1862	19.51	0.6635	16.79	0.6444	18.15	0.6540	0.2417
MSEC [1]	20.52	0.8129	19.79	0.8156	20.08	0.8145	0.1721	19.62	0.6512	17.59	0.6560	18.58	0.6536	0.2814
ZeroDCE [12]	14.55	0.5887	10.40	0.5142	12.06	0.5441	0.2923	16.92	0.6330	7.11	0.4292	12.02	0.5311	0.3532
Zero-DCE++ [22]	13.82	0.5887	9.74	0.5142	11.37	0.5583	0.3121	11.93	0.4755	6.88	0.4088	9.41	0.4422	0.3623
RUAS [26]	13.43	0.6807	6.39	0.4655	9.20	0.5515	0.4819	16.63	0.5589	4.54	0.3196	10.59	0.4393	0.5122
SCI [27]	9.97	0.6681	5.83	0.5190	7.49	0.5786	0.3116	17.86	0.6401	4.45	0.3629	12.49	0.5051	0.4239
PairLIE [11]	11.78	0.6596	8.37	0.5887	9.73	0.6171	0.3605	16.67	0.5995	6.26	0.3846	11.47	0.4921	0.4138
ENC-SID [15]	22.59	0.8423	22.36	0.8519	22.45	0.8481	0.1827	21.30	0.6645	19.63	0.6941	20.47	0.6793	0.2797
ENC-DRBN [15]	22.72	0.8544	22.11	0.8521	22.35	0.8530	0.1724	21.89	**0.7071**	19.09	0.7229	20.49	0.7150	0.2318
CLIP-LIT [24]	17.79	0.7611	12.02	0.6894	14.32	0.7181	0.2506	15.13	0.5847	7.52	0.4383	11.33	0.5115	0.3560
FECNet [16]	22.96	0.8598	23.22	0.8748	23.12	0.8688	0.1419	22.01	0.6737	19.91	0.6961	20.96	0.6849	0.2656
LCDPNet [30]	22.35	**0.8650**	22.17	0.8476	22.30	0.8552	0.1451	17.45	0.5622	17.04	0.6463	17.25	0.6043	0.2592
FECNet+ERL [18]	23.10	0.8639	23.18	0.8759	23.15	0.8711	/	22.35	0.6671	20.10	0.6891	21.22	0.6781	/
CoTF(Ours)	**23.36**	0.8630	**23.49**	**0.8793**	**23.44**	**0.8728**	**0.1232**	**22.90**	0.7029	**20.13**	**0.7274**	**21.51**	**0.7151**	**0.1924**

Table 1. Quantitative comparisons on the MSEC and the SICE datasets. Some are absent ("/") due to the unavailable source code. The best results are highlighted in bold.


Input URetinexNet DRBN ZeroDCE SCI CLIP-LIT

MSEC ENC-DRBN FECNet LCDPNet Ours GT

Figure 6. Visual comparison with state-of-the-art methods on the MSEC dataset.

dataset, our method has the highest PSNR and the second highest SSIM score. Table 2 shows the quantitative results on the LCDP dataset. As can be seen, our method improves 0.65dB PSNR and 0.0161 SSIM compared to the second best LCDPNet method. Overall, our method can achieve comparable performance with state of the art methods.

Efficiency Evaluation. We report the efficiency comparisons of the different methods in Table 2. Our method significantly reduces the computational cost and meets the requirements of real-time processing. For example, compared to the pixel-wise transformation method FECNet, our method requires only 2% FLOPs and 8% runtime. Our method has 93% fewer FLOPs and is 80% faster compared to LCDPNet, which is partially run at low resolution. This is because our method unifies pixel-wise and global transformations in an efficient way that is insensitive to the number of pixels. These results demonstrate the efficiency and practicality of our method.

Qualitative Comparisons. We provide qualitative comparisons in Figure 6, Figure 7 and Figure 8. As can be seen, other methods always suffer from over- or under-enhancement, color deviation and blurring effects. And our method succeeds in restoring proper global brightness and local contrast, consistent colors, and sharp details. These results prove that our method produces more pleasing visual effects. More visual results can be found in the supplementary material.

4.3. Ablation Studies

We perform ablation studies on the LCDP dataset to verify the effectiveness of each component of the our method.

Effectiveness of each component. We set up different vari-

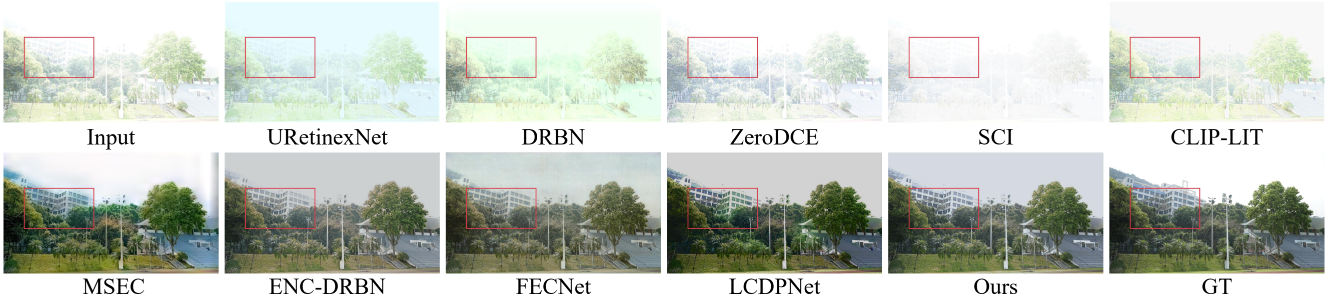


Figure 7. Visual comparison with state-of-the-art methods on the SICE dataset.

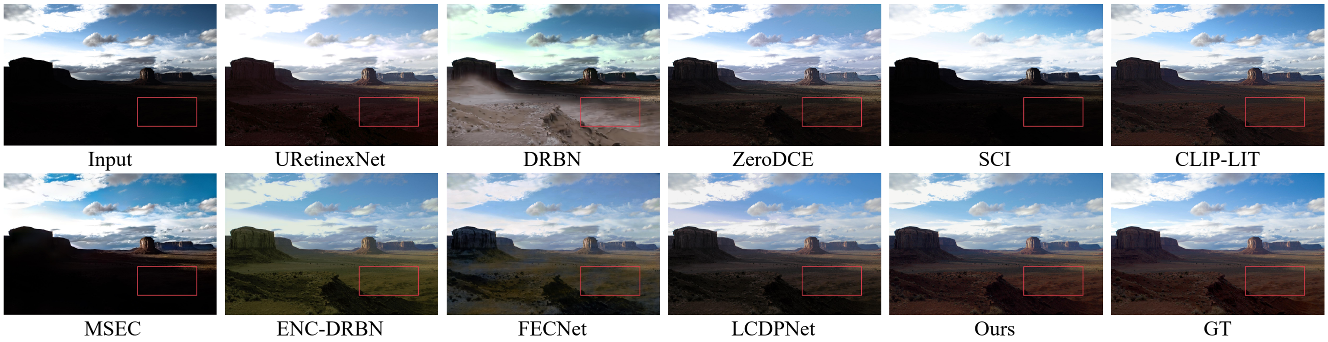


Figure 8. Visual comparison with state-of-the-art methods on the LCDP dataset.

Methods	PSNR↑	SSIM↑	LPIPS↓	Param(M)↓	FLOPs(G)↓	Time(s)↓
HE [29]	15.98	0.6840	0.3871	-	-	-
CLAHE [56]	16.33	0.6420	0.5054	-	-	-
LIME [13]	17.34	0.6860	0.2759	-	-	-
WVM [10]	18.16	0.7390	0.2123	-	-	-
RetinexNet [38]	16.20	0.6304	0.2940	0.84	566.08	0.1529
URetinexNet [39]	17.67	0.7369	0.2504	1.32	913.36	0.1877
DRBN [44]	15.47	0.6979	0.3149	0.58	170.55	0.1226
SID [9]	21.89	0.8082	0.1781	7.40	219.29	0.0387
MSEC [1]	17.07	0.6428	0.3151	7.04	154.28	0.0468
ZeroDCE [12]	18.96	0.7743	0.2055	0.079	83.27	0.0229
Zero-DCE++ [22]	18.42	0.7669	0.2204	0.01	**0.21**	0.0024
RUAS [26]	13.93	0.6340	0.3458	0.003	3.88	0.0281
SCI [27]	15.96	0.6646	0.2913	**0.0003**	0.55	**0.0021**
PairLIE [11]	16.51	0.6667	0.2945	0.34	358.37	0.0716
ENC-SID [15]	22.66	0.8195	0.1631	7.45	278.76	0.0647
ENC-DRBN [15]	23.08	0.8302	0.1536	0.58	227.73	0.1869
CLIP-LIT [24]	19.24	0.7477	0.2262	0.28	292.56	0.0877
FECNet [16]	22.34	0.8038	0.2334	0.15	94.61	0.1261
LCDPNet [30]	23.24	0.8420	0.1368	0.96	27.12	0.0472
FECNet+ERL [18]	/	/	/	0.15	94.61	0.1261
CoTF(Ours)	**23.89**	**0.8581**	**0.1035**	0.31	1.81	0.0095

Table 2. Quantitative comparison on LCDP datasets. Some are absent ("/") due to the unavailable source code. We also report efficiency comparisons where FLOPs and runtimes are measured with 1024 × 1024 images. Runtimes are averaged over 10 images on the NVIDIA TITAN V GPU. The best results are highlighted in bold.

Setting	Pixel Trans	Global Trans	Feature Mod	PSNR	SSIM
1	✓			20.08	0.5983
2	✓(HR)			22.34	0.8073
3		w/o AdaSamp		23.07	0.8298
4		✓		23.35	0.8343
5	✓	✓	CAT	23.43	0.8372
6	✓	✓	CA	23.62	0.8407
7	✓	✓	RAM*	23.79	0.8550
8	✓	✓	✓	**23.89**	**0.8581**

Table 3. Ablation study on the key components of the CoTF. HR denotes high resolution. AdaSamp denotes adaptive sampling. CAT, CA, and RAM* denote the use of concatenation, channel attention, or RAM module after upsampling.

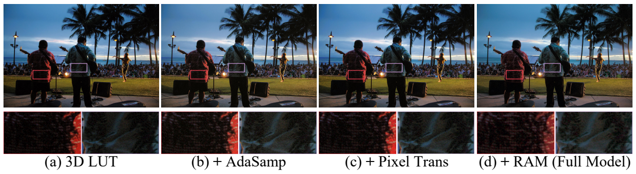

(a) 3D LUT (b) + AdaSamp (c) + Pixel Trans (d) + RAM (Full Model)

Figure 9. Visual results of ablation study on the key components of the CoTF.

ants to validate the effectiveness of the proposed framework. The results are listed in Table 3. Setting 1 has poor performance using only low-resolution pixel-wise transformations. Setting 2 is performed at high resolution, but the shallow network is still not impressive enough. Settings 3 and 4 show the effectiveness of 3D LUTs and proposed adaptive sampling strategy. We then verify the effectiveness of the RAM module. Settings 5, 6, and 7 indicate feature modulation using concatenation, channel attention, or channel self-attention, respectively, after upsampling fea-

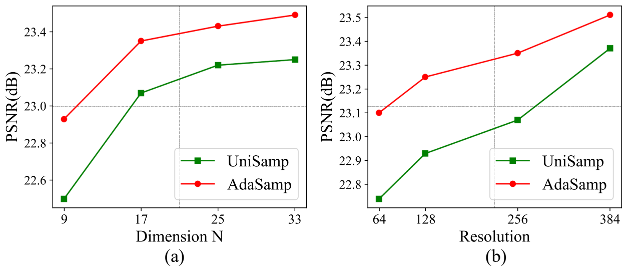


Figure 10. Ablation study of adaptive sampling with (a) 3D LUTs of different dimensions (N), and (b) different sampling resolutions.

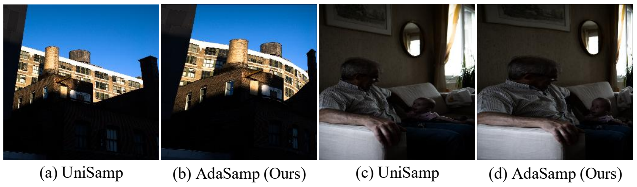

(a) UniSamp (b) AdaSamp (Ours) (c) UniSamp (d) AdaSamp (Ours)

Figure 11. Visualization of our adaptive sampling.

tures to the same resolution. As can be seen, our RAM module provides better results, probably because direct cross-resolution interaction avoids ambiguity compared to naive upsampling, and self-attention can model correlations better. As can be seen in Figure 9, with the help of adaptive sampling and collaborative transformations, our full model yields more visually pleasing results with better local contrast. These results consistently demonstrate the effectiveness of our method.

Analysis of adaptive sampling. We perform ablation studies to analysis the effect of adaptive sampling. For a fair comparison, we take the original 3D LUT with uniform sampling as a baseline. **First**, we evaluate the performance of adaptive sampling under different LUT sizes N. As shown in Figure 10(a), performance goes up as N increases, and our method consistently improves baseline under all N settings. **Second**, we analyze the effect at different sampling resolutions. From Figure 10(b), the performance improves as the resolution increases, which shows the importance of color information for LUT learning. While adaptive sampling boosts the performance at a given size, which proves that adaptive sampling retains more color information. **Finally**, as illustrated in Figure 11, our method can densely sample colorful regions and sparsely sample flat regions. Note that our adaptive sampling requires only a slight increase in computation. For example, at a sample size of 256 × 256, our method adds only 0.02G FLOPs, which is almost negligible.

4.4. Extension and Discussion

Ultra-High-Definition (UHD) Images. We further extend our CoTF to UHD images, which is a more challenging.

Method	PSNR	SSIM	Time(s)
3D LUT	18.11	0.6194	0.0023
CoTF(Ours)	**19.09**	**0.6390**	0.0431

Table 4. Quantitative results and runtime of our method on UHD images (3840 × 2160 resolution).

Setting	Train	Test	PSNR	SSIM
1	×	×	23.07	0.8298
2	✓	×	23.33	0.8324
3	✓	✓	23.35	0.8343

Table 5. Investigation of adaptive sampling as a data augmentation strategy (i.e. Setting 2).

We use the original resolution version of SICE [2], which contains 4K-5K resolution images. Most methods fail to process UHD images due to out-of-memory. In contrast, our method can still process UHD images efficiently on an NVIDIA TITAN V GPU, as shown in Table 4. Despite slower than 3D LUTs, our method has substantially improved performance and still meets real-time requirements. More results are in the supplementary material.

Adaptive sampling as data augmentation. We further investigate adaptive sampling as a data augmentation strategy. As shown in Table 5, Setting 1 does not use adaptive sampling. Setting 2 utilizes adaptive sampling as a data augmentation strategy and deactivates it during testing. In setting 3, adaptive sampling is a network module. It can be seen that using adaptive sampling to augment the samples also improves the performance, suggesting that the higher quality data provided by adaptive sampling can facilitate the 3D LUT learning.

5. Conclusion

In this paper, we present a collaborative transformations framework (CoTF) for real-time exposure correction that efficiently integrates global and pixel-wise transformations. To efficiently combine these two kinds of transformations, we design a relation-aware modulation module (RAM) to complement the global transformation results with local context information. In addition, to further improve the learning of 3D LUTs, we propose an adaptive sampling strategy to preserve more color information and thus provide higher quality input data. Extensive experiments demonstrate the superiority of our method over the previous methods in terms of performance and efficiency.

Acknowledgments. This work was partially supported by the National Key R&D Program of China 2022YFC3301000 and Knowledge Innovation Program of Wuhan-Shuguang Project under Grant 2023010201020226.

References

[1] Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9157–9167, 2021. 1, 3, 5, 6, 7

[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. 5, 8

[3] Meng Cao, Long Chen, Mike Zheng Shou, Can Zhang, and Yuexian Zou. On pursuit of designing multi-modal transformer for video grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9810–9823, 2021. 1

[4] Meng Cao, Haozhi Huang, Hao Wang, Xuan Wang, Li Shen, Sheng Wang, Linchao Bao, Zhifeng Li, and Jiebo Luo. Unifacegan: a unified framework for temporally consistent facial video editing. *IEEE Transactions on Image Processing*, 30: 6107–6116, 2021.

[5] Meng Cao, Ji Jiang, Long Chen, and Yuexian Zou. Correspondence matters for video referring expression comprehension. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4967–4976, 2022.

[6] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022.

[7] Meng Cao, Can Zhang, Long Chen, Mike Zheng Shou, and Yuexian Zou. Deep motion prior for weakly-supervised temporal action localization. *IEEE Transactions on Image Processing*, 31:5203–5213, 2022.

[8] Meng Cao, Fangyun Wei, Can Xu, Xiubo Geng, Long Chen, Can Zhang, Yuexian Zou, Tao Shen, and Daxin Jiang. Iterative proposal refinement for weakly-supervised video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6524–6534, 2023. 1

[9] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018. 1, 3, 6, 7

[10] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2782–2790, 2016. 2, 6, 7

[11] Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 22252–22261, 2023. 6, 7

[12] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1777–1786, 2020. 3, 6, 7

[13] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017. 2, 6, 7

[14] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[15] Jie Huang, Yajing Liu, Xueyang Fu, Man Zhou, Yang Wang, Feng Zhao, and Zhiwei Xiong. Exposure normalization and compensation for multiple-exposure correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6043–6052, 2022. 1, 3, 4, 5, 6, 7

[16] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. In *Proceedings of the European Conference on Computer Vision*, pages 163–180. Springer, 2022. 1, 3, 6, 7

[17] Jie Huang, Man Zhou, Yajing Liu, Mingde Yao, Feng Zhao, and Zhiwei Xiong. Exposure-consistency representation learning for exposure correction. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6309–6317, 2022. 3

[18] Jie Huang, Feng Zhao, Man Zhou, Jie Xiao, Naishan Zheng, Kaiwen Zheng, and Zhiwei Xiong. Learning sample relationship for exposure correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9904–9913, 2023. 1, 3, 6, 7

[19] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] Chongyi Li, Chunle Guo, Ruicheng Feng, Shangchen Zhou, and Chen Change Loy. CuDi: Curve distillation for efficient and controllable exposure adjustment. *arXiv preprint arXiv:2207.14273*, 2022. 3

[22] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4225–4238, 2022. 3, 6, 7

[23] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6):2828–2841, 2018. 2

[24] Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8094–8103, 2023. 3, 6, 7

[25] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. 4D LUT: learnable context-aware 4D lookup table for image

enhancement. *IEEE Transactions on Image Processing*, 32: 4742–4756, 2023. 3

[26] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10556–10565, 2021. 3, 6, 7

[27] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022. 6, 7

[28] Ntumba Elie Nsampi, Zhongyun Hu, and Qing Wang. Learning exposure correction via consistency modeling. In *Proceedings of the British Machine Vision Conference*, 2021. 3

[29] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 1987. 2, 6, 7

[30] Haoyuan Wang, Ke Xu, and Rynson WH Lau. Local color distributions prior for image enhancement. In *Proceedings of the European Conference on Computer Vision*, pages 343–359. Springer, 2022. 3, 5, 6, 7

[31] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6842–6850, 2019. 3

[32] 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. 2

[33] Tao Wang, Yong Li, Jingyang Peng, Yipeng Ma, Xian Wang, Fenglong Song, and Youliang Yan. Real-time image enhancer via learnable spatial-aware 3D lookup tables. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2471–2480, 2021. 2, 3

[34] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2654–2662, 2023. 3

[35] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2604–2612, 2022. 3

[36] Yang Wang, Long Peng, Liang Li, Yang Cao, and Zheng-Jun Zha. Decoupling-and-aggregating for image exposure correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18115–18124, 2023. 1, 3

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[38] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 3, 6, 7

[39] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2022. 3, 6, 7

[40] Ke Xu, Xin Yang, Baocai Yin, and Rynson W.H. Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2278–2287, 2020. 3

[41] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. SNR-aware low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17693–17703, 2022. 3

[42] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17522–17531, 2022. 2, 3

[43] Canqian Yang, Meiguang Jin, Yi Xu, Rui Zhang, Ying Chen, and Huaida Liu. SepLUT: Separable image-adaptive lookup tables for real-time image enhancement. In *Proceedings of the European Conference on Computer Vision*, pages 201–217. Springer, 2022. 3

[44] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3060–3069, 2020. 3, 6, 7

[45] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality. *IEEE Transactions on Image Processing*, 30:3461–3473, 2021. 3

[46] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021. 3

[47] Lu Yuan and Jian Sun. Automatic exposure correction of consumer photographs. In *Proceedings of the European Conference on Computer Vision*, pages 771–785. Springer, 2012. 2

[48] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 5

[49] 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*, 44(4):2058–2073, 2022. 2, 3, 4, 5

[50] Fengyi Zhang, Hui Zeng, Tianjun Zhang, and Lin Zhang. CLUT-Net: Learning adaptively compressed representations

of 3DLUTs for lightweight image enhancement. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6493–6501, 2022. 3

[51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[52] 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*, pages 1632–1640, 2019. 3

[53] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *International Journal of Computer Vision*, 129:1013–1037, 2021. 3

[54] Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, and Meng Wang. Deep color consistent network for low-light image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1889–1898, 2022. 3

[55] Anqi Zhu, Lin Zhang, Ying Shen, Yong Ma, Shengjie Zhao, and Yicong Zhou. Zero-shot restoration of underexposed images via robust retinex decomposition. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6, 2020. 3

[56] Karel Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics gems*, 1994. 2, 6, 7