ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10445905

ATTENTIONLUT: ATTENTION FUSION-BASED CANONICAL POLYADIC LUT FOR REAL-TIME IMAGE ENHANCEMENT

Kang Fu[1], *Yicong Peng*[1], *Zicheng Zhang*[1], *Qihang Xu*[2], *Xiaohong Liu*[3]*, *Jia Wang*[1]*, *Guangtao Zhai*[1]

[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China
[2] Transsion, China
[3] John Hopcroft Center, Shanghai Jiao Tong University, China

ABSTRACT

Recently, many algorithms have employed image-adaptive lookup tables (LUTs) to achieve real-time image enhancement. Nonetheless, a prevailing trend among existing methods has been the employment of linear combinations of basic LUTs to formulate image-adaptive LUTs, which limits the generalization ability of these methods. To address this limitation, we propose a novel framework named AttentionLut for real-time image enhancement, which utilizes the attention mechanism to generate image-adaptive LUTs. Our proposed framework consists of three lightweight modules. We begin by employing the global image context feature module to extract image-adaptive features. Subsequently, the attention fusion module integrates the image feature with the priori attention feature obtained during training to generate image-adaptive canonical polyadic tensors. Finally, the canonical polyadic reconstruction module is deployed to reconstruct image-adaptive residual 3DLUT, which is subsequently utilized for enhancing input images. Experiments on the benchmark MIT-Adobe FiveK dataset demonstrate that the proposed method achieves better enhancement performance quantitatively and qualitatively than the state-of-the-art methods.

Index Terms— Image enhancement, Photo retouching, 3D lookup table, Attention mechanism.

1. INTRODUCTION

We very likely capture low-quality photos with advanced cameras or cell phones, which is because the quality of final photos is affected by external factors such as ambient light and temperature. To procure visually pleasing images, image enhancement[1, 2, 3, 4, 5, 6, 7] technology becomes indispensable for the refinement of these low-quality photographs. The realm of image enhancement has witnessed remarkable advancements through deep learning-based methodologies, consistently achieving state-of-the-art results. HDRNet [8] uses a low-resolution vision of the input image to predict a set of affine transformations in bilateral space and applies the upsampled vision of these transformations to enhance the input image. CSRNet [9] employs a conditional network to extract global features and utilizes 1×1 convolutions to enact pixel-independent transformations, which simulate global brightness adjustments and other enhancement operations. Zeng *et al.* [10] introduced a framework that predicts an image-adaptive 3DLUT, representing pixel-independent enhancement transformations, for real-time enhancing input images. Subsequent refinements in this domain include SA-3DLUT [11] and AdaInt [12], which respectively incorporate spatial information and image-adaptive sampling to improve the image-adaptive 3DLUT.

However, it's important to note that these methods all rely on a uniform approach to generate image-adaptive 3DLUT. This approach entails the utilization of a lightweight Convolutional Neural Network (CNN) for predicting coefficients used in linear combinations of basic 3DLUTs, ultimately yielding the image-adaptive 3DLUT through these linear combinations. This ordinary linear combination of basic LUTs limits the ability to represent complex transformations of image-adaptive 3DLUT, which leads to mediocre enhancement results. To address these issues, we proposed a novel framework to predict a more precise image-adaptive 3DLUT. We adopt an attention mechanism-based fusion method instead of linear fusion and utilize the Canonical Polyadic (CP) decomposition to decompose 3DLUT into multi one-dimension tensors. Our contributions are summarized as follows:

- We employ an attention fusion module to fuse the global image context feature and the priori attention feature to obtain image-adaptive CP tensors.
- We design a canonical polyadic reconstruction module to convert the image-adaptive CP tensors to image-adaptive 3DLUT for image enhancement. This way can reduce computational complexity and parameters of the model.
- Experiments conducted on benchmark demonstrate that our method presents better image enhancement perfor-

* Corresponding Authors. E-mail:{xiaohongliu, jiawang}@sjtu.edu.cn.
This work was supported in part by STCSM under Grant 22DZ2229005, the National Natural Science Foundation of China under Grant 62301310, the Shanghai Pujiang Program under Grant 22PJ1406800 and Transsion, China.

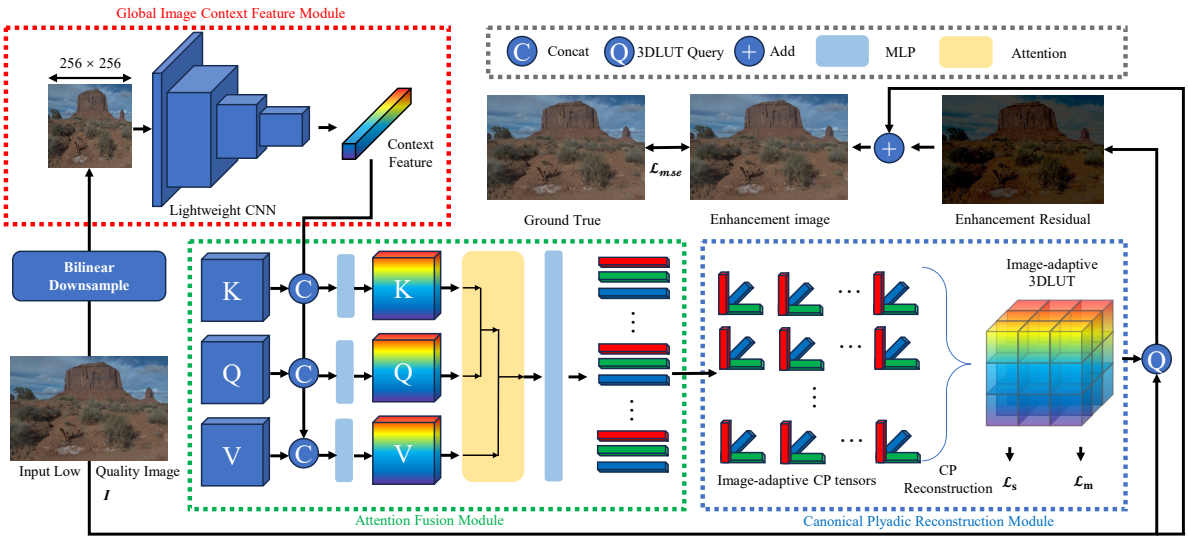


Fig. 1. The framework of our proposed method. It consists of three modules: GICF module, AF module, and CPR module. The functionality of these modules are detailed in 2.3, 2.4, and 2.5 respectively. The entire framework is trained in an end-to-end manner, and the loss functions are detailed in 2.6.

mance than the SOTA models.

2. METHODOLOGY

2.1. Preliminary: 3D Lookup Tables

Traditional 3DLUT is a widely used technology for real-time color mapping or correcting. A 3DLUT is commonly represented by a 3D cubic grid $\mathbf{\Psi} \in \mathbb{R}^{3\times N\times N\times N}$ (N is the number of grid in each dimension) and each grid stores the corresponding output values $\mathbf{\Psi}(x, y, z) \in \mathbb{R}^3$, where $x, y, z = 1, \ldots, N$. Assuming an input pixel value is (r_i, g_i, b_i), the outputs (r_o, g_o, b_o) can be obtained in two steps: (1) Find the nearest eight grids with input pixel values as coordinate points. (2) Calculate the output through trilinear interpolation for these eight grids. This process is intuitively illustrated in the Fig 2. Given an input image I, we can get the enhanced image $O = \mathbf{\Psi}(I)$ by querying pixel by pixel.

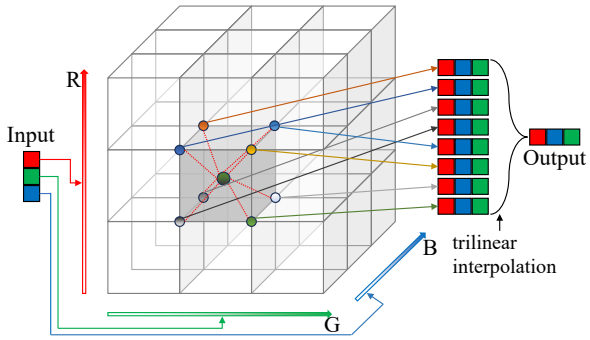


Fig. 2. Schematic diagram of pixel mapping process of 3DLUT. Given a pixel, the output is obtained by trilinear interpolation through the nearest eight grid of input pixels.

2.2. Overall Framework

As outlined in Fig 1, Our proposed framework consists of three parts: global image context feature (GICF) module, attention fusion (AF) module, and canonical polyadic reconstruction (CPR) module. Firstly, the GICF module employs a lightweight CNN to extract the global image context feature, which is used to predict image-adaptive 3DLUT. Next, the AF module converts the priori attention feature and global image context feature to image-adaptive CP tensors. Then, the CPR Module calculates image-adaptive 3DLUT by the outer products of above CP tensors. Finally, the enhancement image is the addition of the input image and enhancement residual image which is obtained by image-adaptive residual 3DLUT $\mathbf{\Psi}_r$ query.

2.3. Global Image Context Feature Module

The image enhancement algorithm using traditional 3DLUT manually selects one from a series of preset 3DLUTs for color mapping. This results in the different images being only able to use preset color mapping, which lacks generalization and flexibility. To solve this shortcoming, the image feature needs to be used in the calculation of image-adaptive 3DLUT. Since 3DLUT performs global color mapping on images, we chose to use a lightweight CNN and a low-resolution version of the input image to predict global image context features as in the previous work[10]. In order to process an arbitrarily sized image in real-time, The GICF module first resizes the input

low-quality image to 256×256 using bilinear interpolation and then extracts global image context feature from it through a lightweight CNN $\mathcal{G}$. This CNN consists of 5 convolutional layers with a stride of 2 and an average pooling layer. In conclusion, the GICF module can be described as follows:

$$f_I = \mathcal{G}(\mathcal{D}(I)), \tag{1}$$

where I is the input image, $\mathcal{D}$ is the bilinear downsample and $f_I \in \mathbb{R}^{1\times 128}$ denotes the global image context feature.

2.4. Attention Fusion Module

It is significant to predict precise image-adaptive 3DLUT because every pixel value of the enhancement image is stored in its grid. Previous works [10, 11, 12, 13, 14, 15] are all converting image features to the linear combination coefficients of basic LUTs and representing image-adaptive 3DLUT as the linear combination of basic LUTs. This fusion approach leads to image-adaptive 3DLUT with weak generalization and limited representation of complex color mappings. Therefore, we opt to obtain better enhancement by changing the fusion method of image-adaptive 3DLUT.

Inspired by [16], we use the multi-head attention mechanism to calculate image-adaptive 3DLUT by global image context feature and priori information. To simplify calculation, we utilize **Key**, **Query**, and **Value** (abbreviated as KQV) to store prior information instead of basic LUTs. We first convert priori KQV and global image context feature to image-adaptive KQV, and then calculate image-adaptive CP tensors through image-adaptive KQV. This process can be formulated as:

$$\mathcal{A}_I = \mathcal{P}_{\mathcal{A}}(\mathcal{A}, f_I) \quad \mathcal{A} \in \{K, Q, V\}, \tag{2}$$

where $\mathcal{A}$ and $\mathcal{A}_I$ are respectively priori KQV and image-adaptive KQV, $P_{\mathcal{A}}$ is an ordinary linear layer, which convert $\mathcal{A}$ to $\mathcal{A}_I$.

$$\mathcal{C} = \mathcal{Z}(softmax(\frac{Q_I K_I^{\mathsf{T}}}{\sqrt{d_{K_I}}})V_I), \tag{3}$$

where K_I, Q_I, V_I are respectively image-adaptive K, Q and V, d_{K_I} is the last dimension of K_I and Q_I, $\mathcal{Z}$ is fully-connected layer to compress intermediate feature dimensions and $\mathcal{C} \in \mathbb{R}^{9\times X\times N}$ denotes image-adaptive CP tensors.

2.5. Canonical Polyadic Reconstruction Module

The parameters of 3DLUT are proportional to N^3, which leads to large model size and calculation. Inspired by [17], we apply classical Canonical Polyadic decomposition [18] to factorize the image enhancement 3DLUT into X Rank-1 tensor (shown in Fig 3). This process can be formulated as :

$$\mathbf{\Psi}_r = \sum_{x=1}^{X} R_x \otimes G_x \otimes B_x, \tag{4}$$

where $R_x, G_x, B_x \in \mathbb{R}^{3\times 1\times N}$ are respectively the R, G, and B channels in image-adaptive CP tensors, $\otimes$ denotes the outer product.

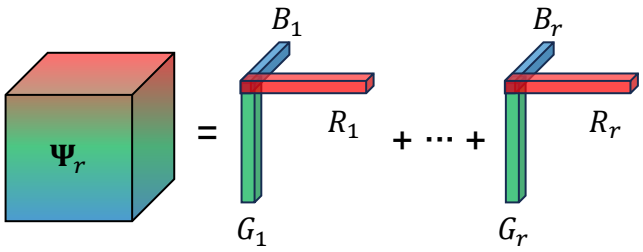


Fig. 3. Illustration of canonical polyadic decomposition. It factorizes a tensor as a sum of vector outer products.

2.6. Loss Functions

Similar to previous work [10], The training loss function consists of three basic losses: the MSE loss $\mathcal{L}_{\mathbf{mse}}$, the smoothness regularization loss $\mathcal{L}_{\mathbf{s}}$ and the monotonicity regularization loss $\mathcal{L}_{\mathbf{m}}$. The final loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathbf{mse}} + \lambda_s \mathcal{L}_{\mathbf{s}} + \lambda_m \mathcal{L}_{\mathbf{m}}, \tag{5}$$

where λ_s and λ_m are trade-off coefficients, which are set to 1×10^{-4} and 10. The $\mathcal{L}_{\mathbf{mse}}$ is used to optimize the performance of enhancement, The $\mathcal{L}_{\mathbf{s}}$ and $\mathcal{L}_{\mathbf{m}}$ are employed for ensuring a more natural enhancement results. Since the image-adaptive residual 3DLUT stores the pixel value of the enhancement residual, the $\mathcal{L}_{\mathbf{s}}$ and $\mathcal{L}_{\mathbf{m}}$ are calculated through the 3DLUT $\mathbf{\Psi}_l = \mathbf{\Psi}_r + \mathbf{\Psi}_f$, where $\mathbf{\Psi}_f$ is a 3DLUT where the value stored at grid (i, j, k) is (i, j, k).

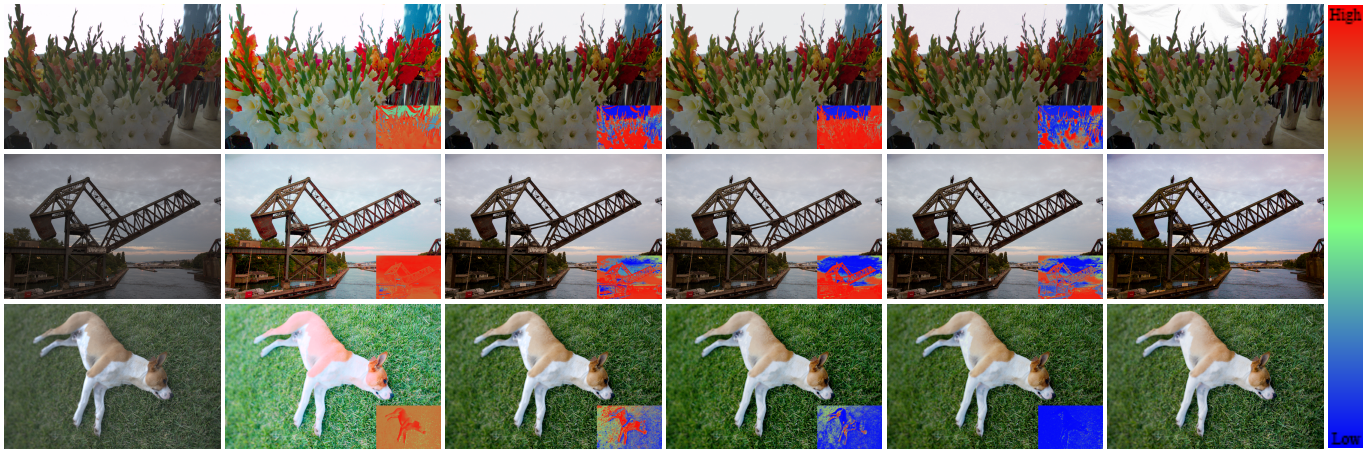
3. EXPERIMENTS

3.1. Experiment settings

Datasets The MIT-Adobe-5K dataset[19] is a large image enhancement dataset with manually retouched ground truth, which contains 5,000 sets of images, each with a RAW image and retouched images by five human experts. Following the common practice in recent works [9, 10], we use the images retouched by expert C as the ground truth in subsequent experiments and divide the dataset into 4,500 training image pairs and 500 testing pairs.

Evaluation metrics In order to evaluate the enhancement performance of different methods, we use PSNR, SSIM [20], $\triangle E$ as experimental evaluation metrics. $\triangle E$ is the mean L2-distance between the predicted image and ground truth in CIELAB color space.

Implementation Details We conduct all the experiments on an NVIDIA TITAN Xp GPU with Pytorch [21] framework and use the standard Adam [22] optimizer to minimize the loss function in Equation 5. The mini-batch size is set to 1. All our models are trained in 400 epochs with a fixed learning



(a) Input (b) CSRNet (c) 3DLUT (d) SepLUT (e) Ours (f) GT

Fig. 4. Qualitative comparisons with corresponding error maps on the FiveK dataset for image enhancement. Blue indicates good effect and red indicates large differences. Best viewed on screen.

rate 1×10^{-4}. In order to achieve a trade-off between enhancement effect and parameter quantity, the number of CP tensors X and the number of grids in each dimension N are set to 15 and 33 respectively. d_{K_I} is set to 128.

3.2. Comparison with SOTA and Results

We compare proposed method with a series of state-of-the-art methods including CSRNet [9], 3DLUT [10], SALUT [11], DHFN [23] and SepLUT [24].

Quantitative Comparisons As shown in Table 1, our proposed method effectively improves the enhancement performance while ensuring the size of parameters, demonstrating the effectiveness of the AF module and CPR module.

Qualitative Comparisons Fig 4 also shows that our method produces more visually pleasing results than other methods. The error maps indicate that our results are closer to ground truth, as the AF module makes the framework predict more precise image-adaptive 3DLUT, which represents a more precise color mapping of enhancement.

3.3. Ablation Study

In this section, we conduct a series of ablation experiments on the AF module and CPR Module to verify the effectiveness of these proposed modules. Specifically, we train two different models: (b) Using linear combination fusion instead of attention fusion, and (c) Using complete 3DLUT instead of CP decomposition. At the same time, we also choose two models to compare: (a) the original image-adaptive 3DLUT, and (d) the complete version of our method. As shown in Table 2. the AF module significantly increases the performance because it helps predict a more precise and generalized image-adaptive 3DLUT. By comparing (c) and (d), we can conclude that the CPR module effectively reduced the model size while slightly improving performance.

Table 1. Quantitative comparison with state-of-the-art methods. Best in bold.

Method	PSNR↑	SSIM↑	Δ E ↓	Param.(k)↓
CSRNet [9]	24.70	0.881	9.69	**37**
3DLUT [10]	25.23	0.912	7.60	592
SALUT [11]	25.40	0.925	**7.46**	4,155
DHFN [23]	25.46	0.898	9.16	332
SepLUT [24]	25.47	0.925	7.54	120
Ours	**25.56**	**0.926**	7.53	593

Table 2. Ablation study of the different modules in our framework. The best results are boldface.

Method	PSNR ↑	SSIM↑	Δ E ↓	Param.(k)↓
(a)	25.23	0.912	7.60	592
(b)	25.42	0.915	7.58	**294**
(c)	25.47	0.918	7.57	3239
(d)	**25.56**	**0.926**	**7.53**	593

4. CONCLUSION

Since the existing deep learning work on image enhancement uses the linear combination of basic LUTs to represent the image-adaptive 3DLUT, which limits the expression ability of 3DLUT, this paper proposed a novel real-time image enhancement framework called AttentionLUT, which consists of three modules: GICF module, AF module, and CPR module. The GICF module is used to extract the global image context feature from the low-resolution version of the input image and the AF module converts image feature and priori attention feature to image-adaptive CP tensors. The CPR module uses CP tensors to reconstruct image-adaptive 3DLUT to enhance the input image. Extensive experiments demonstrate that our method achieves great performance in both quantitative measures and visual qualities thanks to the AF and CPR modules.

5. REFERENCES

[1] Wenhao Li et al., "Fastllve: Real-time low-light video enhancement with intensity-aware look-up table," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8134–8144.

[2] Xiaohong Liu et al., "Griddehazenet+: An enhanced multi-scale network with intra-task knowledge transfer for single image dehazing," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 870–884, 2022.

[3] Xiaohong Liu et al., "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7314–7323.

[4] Xiangyu Yin et al., "Fmsnet: Underwater image restoration by learning from a synthesized dataset," in *Proceedings of International Conference on Artificial Neural Networks*. Springer, 2021, pp. 421–432.

[5] Shan Huang et al., "Transmrsr: Transformer-based self-distilled generative prior for brain mri super-resolution," *arXiv preprint arXiv:2306.06669*, 2023.

[6] Wenyi Wang et al., "Single image super resolution based on multi-scale structure and non-local smoothing," *EURASIP Journal on Image and Video Processing*, vol. 2021, no. 1, pp. 16, 2021.

[7] Zicheng Zhang et al., "A no-reference evaluation metric for low-light image enhancement," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[8] Michaël Gharbi et al., "Deep bilateral learning for real-time image enhancement," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[9] Jingwen He et al., "Conditional sequential modulation for efficient global image retouching," *arXiv preprint arXiv:2009.10390*, 2020.

[10] Hui Zeng et al., "Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2058–2073, 2020.

[11] Tao Wang et al., "Real-time image enhancer via learnable spatial-aware 3d lookup tables," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2471–2480.

[12] Canqian Yang et al., "Adaint: Learning adaptive intervals for 3d lookup tables on real-time image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17522–17531.

[13] Zicheng Zhang et al., "Subjective and objective quality assessment for in-the-wild computer graphics images," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 4, pp. 1–22, 2023.

[14] Zicheng Zhang et al., "No-reference quality assessment for 3d colored point cloud and mesh models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7618–7631, 2022.

[15] Zicheng Zhang et al., "Perceptual quality assessment for digital human heads," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] Ashish Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Anpei Chen et al., "Tensorf: Tensorial radiance fields," in *European Conference on Computer Vision (ECCV)*, 2022.

[18] J Douglas Carroll et al., "Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[19] Vladimir Bychkovsky et al., "Learning photographic global tonal adjustment with a database of input/output image pairs," in *CVPR 2011*. IEEE, 2011, pp. 97–104.

[20] Zhou Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[21] Adam Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[22] Diederik P Kingma et al., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] Yuhong Zhang et al., "Dual-head fusion network for image enhancement," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] Canqian Yang et al., "Seplut: Separable image-adaptive lookup tables for real-time image enhancement," in *European Conference on Computer Vision*. Springer, 2022, pp. 201–217.