RGB and LUT based Cross Attention Network for Image Enhancement

Tengfei Shi[1,3,4]
tengfeishi_bh@163.com

Chenglizhao Chen[✉2]
cclz123@163.com

Yuanbo He[1]
heluxixue@163.com

Wenfeng Song[5]
songwenfenga@163.com

Aimin Hao[1]
ham@buaa.edu.cn

[1] Beihang University, Beijing, China

[2] China University of Petroleum (East China), Qingdao, China

[3] Qingdao Research Institute, Qingdao, China

[4] Peng Cheng Laboratory, Shenzhen, China

[5] Beijing Information Science and Technology University, Beijing, China

Abstract

Image enhancement aims to improve the quality of images by adjusting their color and is widely used in professional digital photography. Deep learning-based 3 **D**imensional **L**ook**U**p-**T**able (3D LUT) of RGB color transformation has achieved promising performance in terms of speed and precision. However, the focus has mainly been on building an adaptive enhancer by only learning the global color adjusting weights from the image, which ignores the significant relationship between the intrinsic semantic information of the image and LUT that is relevant to photographers. In this paper, we propose the **C**ross **A**ttention **N**etwork (CANet), a new framework that formulates image enhancement as a parallel learning process based on the image and LUT features. To better learn the adjustment weights for both global color and intrinsic semantics, we propose a cross attention architecture that connects low-level (color, edge and outline) and high-level (semantic) features of the image and color transform LUT features to generate appropriate adjustment weights. Meanwhile, we employ a **L**UT-**A**ware **M**odule (LAM) to construct the channels and spatial attention for refining the LUT features. Since these modules have a more powerful representational capacity, they can better capture the intrinsic relationship between image semantics and LUT features. The extensive evaluations on standard benchmarks, including FiveK and HDR datasets, show that CANet achieves better performance compared to state-of-the-art methods.

1 Introduction

Computer vision has achieved great success in processing professional photography images and videos. However, distorted colors and low-light conditions in the real world present

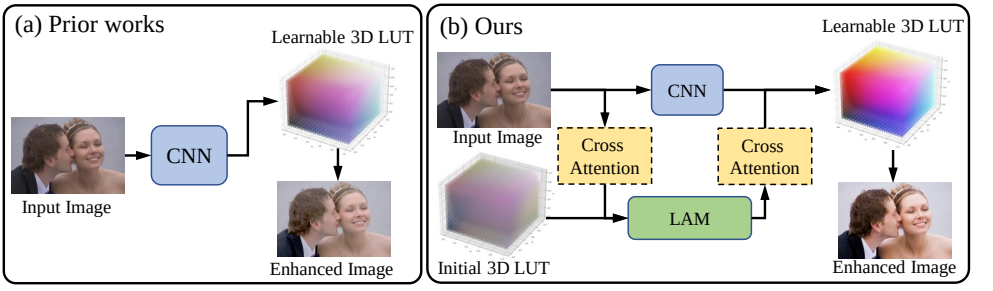

Figure 1: Comparsion between previous works and our method. Different from prior works (a) that emply CNN to extract the features of image directly, our method takes image and LUT features into consider simultaneously. And we adopt cross attention architecture and **LUT-a**ware **m**odule LAM to construct the fine-grained LUT on the learning process.

challenges for both human visual perception and downstream vision tasks. Various image enhancement methods [3, 13, 14, 15, 24, 26] have been proposed to adapt to these difficult lighting conditions and distorted colors. Based on the pattern of image enhancement, these methods can be roughly divided into two types: the **G**enerative **A**dversarial **N**etwork-based (GAN-based) methods and **C**olor **T**ransform-based (CT-based) methods.

GAN-based methods [3, 14] directly generate the enhanced image in a dense manner, resulting in precise results. However, their heavy computational cost and unstable learning process limit the practical applications. To overcome these problems, recent works [7, 13, 19, 20] have extended GAN-based methods into CT-based methods, which learn the color transformations from the CNN model in a more efficient way. Compared with GAN-based methods, CT-based methods employ CNNs on a low-resolution [24] input image to predict image-adaptive parameters of color transform.

The adaptive lookup-table proposed by 3D LUT [24] has achieved promising performance in terms of speed and precision. However, two challenges have arisen for 3D LUT since the learning scene become more complex. For example, as shown in Fig. 1-(a), the paradigm of 3D LUT learns the color transform function through a serial schema and models the color transform information only from single image. Therefore, it is difficult for them to model the deep relationship between semantic and color transform style [21]. Moreover, the specific scene with different color styles includes different semantics, which also increases the difficulty of image enhancement by 3D LUT [25].

To address above problems, we propose a novel method called **C**ross **A**ttention **N**etwork (CANet). In contrast to Fig. 1-(a) by inputing an image for enhancement, our method learns and fuses semantic and contextual information from image and LUT feature with the parallelize way in Fig. 1-(b). First, we propose to add the initial 3D LUT feature as another input and design a cross attention architecture instead of only learning from the image, so that the relationship between image semantics and LUT features can be learned by fusing low-level and high-level information. Furthermore, we design a **LUT-A**ware **M**odule (LAM) to learn the relationship of multi-channel and spatial features in LUT features. These can construct the channel and spatial relationship of semantic-level LUT features and help better integrate the semantic information of the image and the color transform function of LUT features.

Our contribution could be summarised as follows:

- We have proposed a novel framework, **C**ross **A**ttention **N**etwork (CANet), to adapt to fine-grained semantics in the diverse scene, which uses the cross attention architecture to fuse image and LUT feature in a parallelize way.
- We have degined a **L**UT-**A**ware **M**odule (LAM) that can effectively fuse multi-channel and spatial attention features, thereby enhancing the effect of color transform by LUT.
- We conduct comprehensive experiments on two benchmarks, FiveK and HDR, and the results show that our model outperforms recent state-of-the-art methods.

2 Related Work

Image enhancement. Image enhancement is widely used to further improve the perceptual aesthetic quality of photos captured by digital cameras. Existing learning-based image enhancement methods can be roughly categorized into two paradigms, *i.e.*, the GAN-based methods [3, 14, 15], and the color transform-based methods [13, 24, 26]. Chen *et al*. [3] proposed adaptive weighting WGAN to learn with unpaired dataset. However, this method was still far from practical applications due to their heavy computational burdens and limited feasible input resolutions. Moran *et al*. [13] introduced learnable parameterisaion of filters for spatially localised image enhancement. Zhang *et al*. [26] presented lightweight transformer model with small memory usage. Specifically, these methods employ CNNs or transformer on a low-resolution, fixed-size version of the input image to predict image-adaptive parameters of specific color transform functions.

Learnable image LUTs. The high efficiency and wide usage of the LUTs in ISP attract efforts in deep learning-based image enhancement to learning more powerful LUTs via data-driven approaches. Previous works [4, 7, 19] mainly focus on learning 1D LUTs to mimic the color adjustment curves in popular image enhancement. Gao *et al*. [4] employed salieny prior to improve the CNN architecture and model the tone mapping in luminance and affine transform in chrominance. However, this methods usually suffer from the lack of correlation between RGB color channels. Recently, [22, 24] extended those 1D LUT-based methods into using 3D LUTs. [24] predicted 3D LUT with adaption to different image contents by learning several image-independent basis 3D LUT and combining them using image-dependent weights. [22] extended the adaptive 3D LUTs by image-adaptive sampling intervals for learning 3D LUT layout. However, the capacity of the 3D LUT model can only consider the pixels of image, ignore the high-level semantics making these methods suffer from less flexibility and precision.

3 Proposed Method

3.1 3D LUT and trilinear interpolation

The 3D LUT is an effective and widely applied technique for image enhancement. It involves color transformation using a 3-dimensional lattice constructed from RGB color channels.

The 3D LUT primarily involves two key operations: lookup and trilinear interpolation. In Fig. 2 (a), the 3D LUT defines a 3-dimensional lattice containing M^3 elements $\{P_{(i,j,k)}\}_{i,j,k=0,\cdots,M-1}$, where M represents the number of bins for each color channel. Each element $P_{(i,j,k)}$ defines an RGB indexed color table and its transformed output RGB color

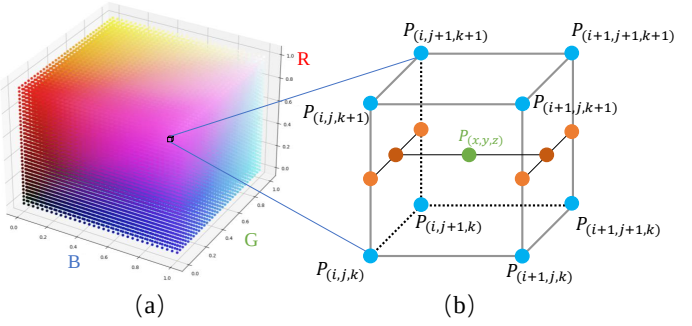

Figure 2: Illustration of (a) a 3D LUT containing 36^3 elements and (b) the trilinear interpolation of the input coordinates

value. Given M typically set to 33 and the RGB color range between 0 and 255, an approximation of values becomes necessary, requiring the application of trilinear interpolation. Specifically, as depicted in Fig. 2 (b), for an input RGB color $\{r^I_{(i,j,k)}, g^I_{(i,j,k)}, b^I_{(i,j,k)}\}$, the output RGB value $P_{(x,y,z)}$ is interpolated using the eight closest neighboring elements through trilinear interpolation. The detailed process of trilinear interpolation can be referred to [24].

The input image is defined as I, and the converted image is O through the 3D LUT, then the conversion process can be represented by the formula:

$$O_{(x,y,z)} = \psi(I^R_{(i,j,k)}, I^G_{(i,j,k)}, I^B_{(i,j,k)}), \tag{1}$$

where (i, j, k) are the pixel coordinates of the image in RGB space, and $\psi(.)$ is defined as the learnable LUT weight which transform input I to output color value O.

However, simply training a CNN to classify scenes and adjust overall color is difficult to achieve the desired relationship between high-level semantic and LUT weight. Inspired by [25], we designed a parallel learning network that takes both image and LUT features as input and integrates them through cross attention architecture. This enables us to obtain fused semantic LUT features, leading to improved results in image enhancement.

3.2 Model Structure

Overview: Given a low-light condition input image $X \in \mathbb{R}^{H\times W\times 3}$, where $H \times W$ represents the height and width of image and 3 corresponds to the RGB channel. As shown in Fig. 3, we present an overview of the proposed image enhancement framework. Our framework follows the parallel paradigm. Specifically, we parallelize the input image X and the initial 3D LUT to learn and adopt a cross attention module to connect high-level semantics and LUT features in an interactive manner. Subsequently, the **L**UT-**A**ware **M**odule (LAM) is utilized as a lightweight feature fusion network to learn channel and spatial LUT-aware relationships that will serve as a guide to generate weights. Finally, the input image interpolates the learnable 3D LUT that is adjusted by weights to output the enhanced image.

Cross Attention Scheme: CANet parallelizes the input image and initial 3D LUT, and connects them by cross attention (see Fig. 3). Inspired by [2], we propose a cross-attention module for image enhancement. We take an image as input ($X \in \mathbb{R}^{H\times W\times 3}$) and applies mobilenet v2 [18] to extract local features. Meanwhile, we take learnable parameters LUT as input, denoted as $\mathcal{Z} \in \mathbb{R}^{M\times d}$ where M and d are the number and dimension of LUT's feature,

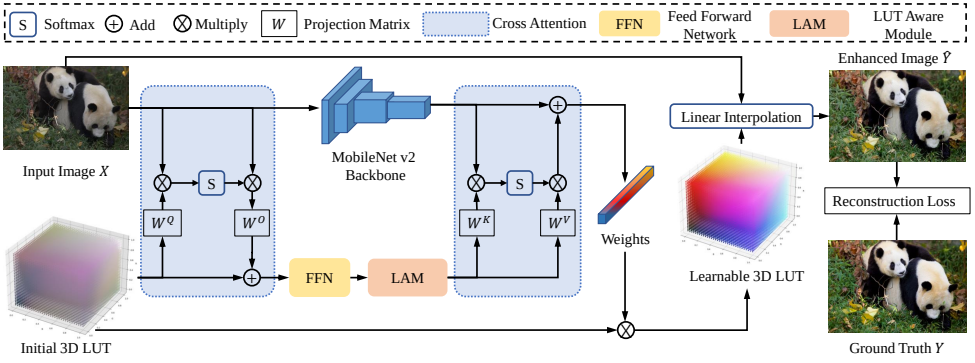

Figure 3: An overview of our method. First, the image X and initial 3D LUT are processed by the **C**ross **A**ttention **N**etwork (CANet), which learn the weights based on fusing high-level and low-level semantic and LUT feature. The **L**UT-**a**ware **m**odule (LAM) is then used to connect the spatial and channel features of LUT. Finally, the features of image are extracted from backbone and fed LUT features together into cross attention module to predict weights and obtain the enhanced image $\hat{Y}$ by interpolation. Best view in color.

respectively. These LUT are specifically initialized. Different from DualBLN [25], which requires the bilinear pooling module to fuse the image and 3D LUT feature, our approach directly learns image and 3D LUT features through projection, where each element represents the attention of color transform with LUT and the image. Notably, this architecture significantly reduces computational costs.

Two-way Interactive Bridge: The image and LUT interact through a two-way bridge where LUT and global semantic features are fused bidirectionally. We denote the two directions as image→LUT and image←LUT, respectively. For modeling this interaction, we utilize a lightweight cross attention module. Notably, only the projections (W^Q, W^K, W^V) are retained on the LUT branch. The cross attention module is computed at the low-level feature and high-level feature of the MobileNet v2 backbone, where the number of channels is different. Specifically, the lightweight cross attention module from the image feature map X to the LUT feature L is computed as follows:

$$\mathcal{A}_{X\to L} = [Attn(\widetilde{l}_i W_i^Q, \widetilde{x}_i, \widetilde{x}_i)]_{i=1:h} W^O, \tag{2}$$

where the image feature X and global LUT L are split into h heads as $X = [\widetilde{x}_1 ... \widetilde{x}_h], L = [\widetilde{l}_1 ... \widetilde{l}_h]$ for multi-head attention. The split for the i^{th} head $\widetilde{l}_i \in \mathcal{R}^{M\times \frac{d}{h}}$ is different from the i^{th} token $l_i \in \mathcal{R}^d$. W_i^Q is the query projection matrix for the i^{th} head. W^O is used to combine multiple heads together. $Attn(Q,K,V)$ is the standard attention function over query Q, key K, and value V as $softmax(\frac{QK^T}{\sqrt{d_k}})V$. $[.]_{1:h}$ denotes the concatenation of h elements. Note that the projection matrices for the key and value are removed from the MobileNet branch, while the projection matrix W_i^Q for the query is kept on the LUT branch. Similarly, the cross attention from the image to the LUT is computed as:

$$\mathcal{A}_{L\to X} = [Attn(\widetilde{x}_i, \widetilde{l}_i W_i^K, \widetilde{z}_i W^V)]_{i=1:h}, \tag{3}$$

where W_i^K and W_i^V are the projection matrices for the key and value on the LUT branch. The projection matrix of the query is removed from the LUT branch.

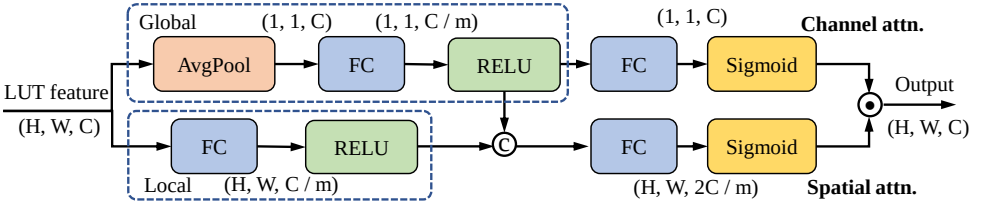

Figure 4: Our proposed LUT attention feature aggregation on LUT-aware Module. This module is learned in parallel by two branches. Inspired by extracted and squeeze [8], in which the channel attention branch acquires channel relationships by learning global features, and the spatial attention branch learns LUT local relationships to obtain more refined features at the channel and spatial levels.

3.3 LUT Attention Module

The feed-forward network (FFN) is a essential component in the transformer block for feature extraction. However, in light-weight models, the limited channel dimensions restrict their performance due to reduced computation costs.

To address this issue, we propose a **L**UT **A**ttention **M**odule (LAM) inspired by [9], which explicitly captures feature relationships in lightweight CNNs. The LAM captures spatial and channel dependencies and refines LUT features through two branches. As illustrated in Fig. 4, the module consists of two branches: the channel attention branch and the spatial attention branch. The channel attention branch aggregates global representations by averaging input features across the spatial dimension and computes channel attention using linear transformation. For spatial attention, we concatenate global representations with local feature for modeling pixel-wise relations. Our proposed LUT attention module can be easily integrated into existing transformer structures, enhancing the representation power of FFN with only a slight increase in computation cost.

3.4 Objective Function

As described in [24], the final loss function $\mathcal{L}_{total}$ can be decomposed into three components: the pixel mean squared error (MSE) loss $\mathcal{L}_{pixel}$, the smooth regularization loss $\mathcal{L}_s$, and the monotonicity regularization loss $\mathcal{L}_m$. These terms constitute the primary loss functions and can be defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{pixel} + \alpha_s \mathcal{L}_s + \alpha_m \mathcal{L}_m, \tag{4}$$

where α_s and α_m are coefficients to adjust the effects of the smooth and monotonicity regularization terms, we follow [24] and set $\alpha_s = 0.0001$, $\alpha_m = 10$. To ensure that the enhanced images produced during the training process exhibit a more natural appearance, it is imperative to accurately adjust the relative brightness and color saturation of the images. In this regard, the MSE loss $\mathcal{L}_{pixel}$ is employed as a training constraint to ensure that the color of the enhanced images is consistent with that of the expert-labeled results. Additionally, smoothness regularization $\mathcal{L}_s$ and monotonicity regularization $\mathcal{L}_m$ are utilized to ensure the smoothness and monotonicity of the LUT weight learning process during training, thereby mitigating the occurrence of color distortion and dispersion in the enhanced images.

4 Experiments

4.1 Experimental Sets

We evaluate our CANet method on two benchmark datasets: The publicly available MIT-Adobe FiveK [1] and HDR [6] datasets are adopted to evaluate the proposed method. The FiveK dataset contains 5,000 RAW images with five manually retouched ground truths (A/B/C/D/E). Version C is selected in our experiments. We use the commonly used settings [24] to split the dataset into 4,500 image pairs for training and the remaining 500 image pairs for testing. The HDR dataset is a burst photography dataset collected by Google camera group for research of high dynamic range (HDR) and low-light imaging on mobile cameras. We use this dataset to verify the generalizability of the model.

Implementation Details. We build our CANet using PyTorch [17], and all operations used in the enhancer have been efficiently and differentiably implemented. As for the LUT generators, inspired by [24], we initialize the 3D LUT generator to predict an identity mapping at the early training stage to speed up the training convergence. The mean square error (MSE) loss is adopted to train the proposed method in an end-to-end manner. The standard Adam optimizer [10] is adopted to train the proposed method, with the mini-batch size set to 1 and 16 on FiveK and HDR, respectively. All models are trained for 300 epochs with a fixed learning rate of 1e-4 on an NVIDIA 3090 GPU.

Evaluation metrics. During the training phase of our implementation, we utilize PSNR as the important metric to select the model with the highest score at the epoch for testing. The metrics of ΔE_{ab} and SSIM are computed during the testing phase. Typically, a higher PSNR score corresponds to better ΔE_{ab} and SSIM scores, indicating a superior learning outcome of the model and a higher quality of the augmented image.

4.2 Overall Comparison Results

We compare our proposed CANet with the existing state-of-the-art methods, including the transformer-based method STAR-DC [26] and the baseline method 3D-LUT [24]. Furthermore, to demonstrate the effectiveness of exploring our model with cross attention architecture for parallelize learning image and LUT feature, we also compare our model with recent image enhancement methods, including Dis-Rec [16], HDRNet [5], DeepLPF [13], CSRNet [7], and AdaInt [22]. The results on the FiveK and HDR datasets are reported in Table 1 and Table 2, respectively.

In Table. 3, we conduct a comparative analysis of five methods proposed in 2022 and 2023. Based on evaluation metrics PSNR and SSIM, our approach demonstrates comparable performance on the MIT FiveK dataset. Additionally, we present experimental results include model parameters and run time. Comparative analysis of these results reveal that our method leverages a parallel learning process based on the image and LUT features and cross-attention networks, requires a higher number of model parameters and cost relatively longer run time. However, it remains within the real-time constraints, ensuring its practical applicability.

Based on the results of comparisons with these methods, we have made the following observations: (1) Our proposed method outperforms related image enhancement methods, demonstrating the effectiveness of our novel architecture with cross attention learning scheme. Specifically, our method surpasses AdaInt [22] by 0.21 dB on FiveK and 3D LUT [24] by 0.28 dB on HDR in terms of the PSNR metric. (2) Our proposed model

Table 1: Quantitative comparisons state-of-the-art methods on the FiveK.

Method	*PSNR* ↑	ΔE_{ab} ↓	*SSIM* ↑
Dis-Rec [16]	21.98	10.42	0.856
HDRNet [6]	24.32	8.49	0.912
DeepLPF [13]	24.73	7.99	0.916
CSRNet [7]	25.17	7.75	0.924
3D LUT [24]	25.21	7.61	0.922
STAR-DCE [26]	24.50	-	0.893
AdaInt [22]	<u>25.28</u>	<u>7.48</u>	**0.925**
CANet (Ours)	**25.49**	**7.25**	**0.925**

Table 2: Quantitative comparisons state-of-the-art methods on the HDR.

Method	*PSNR* ↑	ΔE_{ab} ↓	*SSIM* ↑
Camera Raw	19.86	14.98	0.791
UPE [6]	21.21	13.05	0.816
DPE [13]	22.56	10.45	0.872
HDRNet [7]	23.04	8.97	0.879
3D LUT [24]	<u>23.54</u>	<u>7.93</u>	<u>0.885</u>
CANet (Ours)	**23.82**	**7.85**	**0.890**

Table 3: Effectiveness comparisons of lastest methods on the MIT FiveK.

Method	Params↓	PSNR↑	SSIM↑	Runtime↓
3D LUT [24]	593.5K	25.21	0.922	**1.99ms**
AdaInt [22]	619.7K	**25.49**	**0.926**	2.56ms
SepLUT [23]	**119.8K**	<u>25.47</u>	0.921	<u>2.25ms</u>
DualNet [25]	11.25M	25.42	0.917	56.12ms
4D LUT [12]	924.4K	24.96	0.924	5.75ms
FlexiCurve [11]	<u>130K</u>	24.74	0.920	2.82ms
CANet (Ours)	1.52M	**25.49**	<u>0.925</u>	8.55ms

also shows significant superiority compared to the conventional transformer method STAR-DCE [26], indicating that learning intrinsic semantic and LUT features is helpful for image enhancement in parallel interactive manner. (3) The results on FiveK show better performance compared to the results on HDR. A probable reason is that FiveK is the most commonly used dataset for image enhancement tasks, and its labels are marked by experts, which are more consistent with user study and beneficial to semantic and LUT feature modeling.

4.3 Ablation Study

To demonstrate the effectiveness of our CANet method, we conduct further ablation experiments on essential components. As shown in Table 4, we report the experimental results on different model structures.

Table 4: Ablation Study of each component in our method.

Method	*PSNR* ↑	ΔE_{ab} ↓	*SSIM* ↑
Baseline	24.28	8.42	0.891
Baseline + Transfomer	24.32	8.49	0.912
Baseline + LAM	25.18	7.78	0.916
Baseline + Cross Attn.	25.23	7.52	0.918
CANet (Ours)	**25.49**	**7.25**	**0.925**

Effectiveness of Model Components. We adopt 3D LUT [24] as the baseline model with image and LUT as input for learning the weighted LUT. We verify four ablated models and keep the same experimental parameters: (1) Baseline: We employ Convolutional Neural Network (CNN) to extract image and LUT features for learning the weighted LUT. (2) Baseline + Transformer: The feature extraction is replaced from CNN to Transformer. (3) Baseline + LAM: We change the conventional multi-head attention to LUT-aware module. (4) Baseline + Cross Attn.: The parallelize learning image and LUTs feature with cross attention scheme.

Based on the results presented in Table 4, we make the following observations:
1) Using only CNN or Transformer is insufficient to effectively improve the precision of image enhancement. This suggests that learning and fusing image and LUT features in a serial scheme may limit the representational capacity of the model. Furthermore, comparing (1) and (2) of our ablated model, we find that the effectiveness of the module's representation ability can lead to more improvements.
2) The LAM demonstrates superior advantages in reasoning the LUT feature. According to the experimental results of (2) and (3), employing the LUT-aware module to replace the multi-head attention enables the module to learn the multi-channel and spatial relationship of the LUT feature, which improves the interactive effects of the LUT feature for different channels and locations.
3) The parallelized learning manner and cross attention module is effective for image enhancement tasks. Compared with the experimental results of (2) and (4), adopting the cross-attention module can improve the *PSNR* metric by 0.91 dB. In a parallelized learning scheme, the cross attention module can effectively enhance the model's ability to learn intrinsic semantic features by interacting with the image and LUT, thereby improving the model's representation of image enhancement.

4.4 Qualitative Analysis

To demonstrate the effectness of image enhancement method of the proposed CANet, we present two examples from the FiveK dataset, as depicted in Fig. 5. The visual results demonstrate that our CANet produces the color transformation of the image enhancement that is closest to the ground-truth. Additionally, in the second row of Fig. 5, it can be observed that our method predicts an enhanced result with a colder heatmap in the facial and clothing regions of the person, compared to the input image and the 3D LUT method. This indicates that our method predicts the color deviation from the ground truth in the enhanced result is smaller, and the color adjustment is more in line with the color retouching of human experts in terms of brightness and color saturation.

However, we have observed some errors in the predicted results. Specifically, our model incorrectly predicts the edge pixels of objects in the input images. We hypothesize that this is due to the significant difference in edge color information, which may require more fine-grained contextual features to model. This is a problem that we aim to explore in the future.

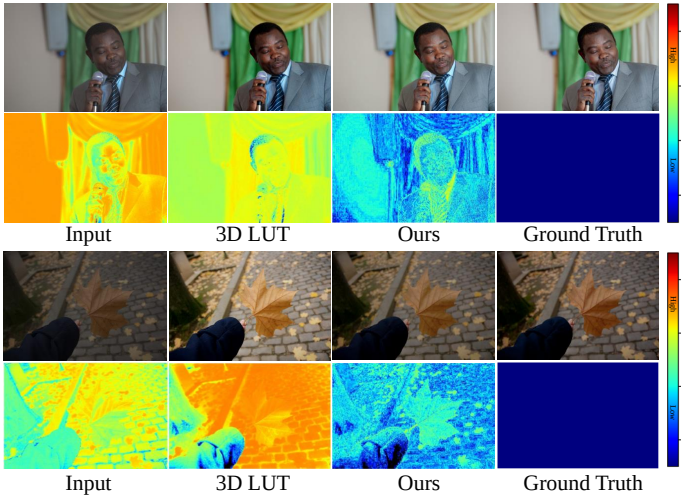

Figure 5: Qualitative comparisons with absolute error heatmaps on the FiveK dataset for image enhancement. Colder regions indicate lower error, while warmer regions indicate larger error between the predicted and ground truth color value. Best viewed in color.

5 Limition and Conclusion

From Table. 1, it can be observed that although our method achieves the comparable results in terms of PSNR and SSIM metrics, its application potential is limited due to the higher model complexity and longer run time caused by the additional LUT feature as inputs and cross-attention networks. In the future, we intend to address the issue of high model complexity by employing model pruning techniques to improve model efficiency. Moreover, we will explore the incorporation of temporal information in video enhancement.

We present a novel framework for image enhancement, called the **C**ross **A**ttention **N**etwork (CANet). The proposed framework employs a parallelized learning scheme for modeling images and LUTs, which is facilitated by a cross attention module that generates LUTs considering the semantic of images. Furthermore, a **L**UT-**A**ware **M**odule (LAM) is designed to enhance the features of LUTs, which enables the model to effectively construct channel and spatial relation information and improve the fine-grained weight LUT. The experimental results on the FiveK and HDR datasets demonstrate the effectiveness of our proposed CANet method. In the future, we plan to explore the idea of saliency-based masks as fine-grained prior information to explicitly guide the construction of the relationship between LUT color mapping and semantics in image enhancement.

Acknowledgement: This work is supported in part by Beijing Advanced Innovation Center for Biomedical Engineering under Grant No. ZF138G1714, CAMS Innovation Fund for Medical Sciences (CIFMS) under Grant No. 2019-I2M-5-016, the National Natural Science Foundation of China under Grant No. 62172246, and the Youth Innovation and Technology Support Plan of Colleges and Universities in Shandong Province under Grant No. 2021KJ062.

References

[1] Connelly Barnes, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. The generalized patchmatch correspondence algorithm. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III 11*, pages 29–43. Springer, 2010.

[2] Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5270–5279, 2022.

[3] 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 *CVPR*, pages 6306–6314, 2018.

[4] Qifan Gao and Xiaolin Wu. Real-time deep image retouching based on learnt semantics dependent global transforms. *IEEE Transactions on Image Processing*, 30:7378–7390, 2021.

[5] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM TOG*, 36(4): 1–12, 2017.

[6] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.

[7] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 679–695. Springer, 2020.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[9] Tao Huang, Lang Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Lightvit: Towards light-weight convolution-free vision transformers. *arXiv preprint arXiv:2207.05557*, 2022.

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

[11] Chongyi Li, Chun-Le Guo, Qiming Ai, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Flexible piecewise curves estimation for photo enhancement. In *CVPRW*, 2023.

[12] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. 4d lut: learnable context-aware 4d lookup table for image enhancement. *arXiv:2209.01749*, 2022.

[13] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *CVPR*, pages 12826–12835, 2020.

[14] Zhangkai Ni, Wenhan Yang, Shiqi Wang, Lin Ma, and Sam Kwong. Unpaired image enhancement with quality-attention generative adversarial network. In *ACM MM*, pages 1697–1705, 2020.

[15] Zhaoqing Pan, Feng Yuan, Jianjun Lei, Wanqing Li, Nam Ling, and Sam Kwong. Miegan: Mobile image enhancement via a multi-module cascade neural network. *IEEE Transactions on Multimedia*, 24:519–533, 2021.

[16] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *CVPR*, pages 5928–5936, 2018.

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[19] Yuda Song, Hui Qian, and Xin Du. Starenhancer: Learning real-time and style-aware image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4126–4135, 2021.

[20] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, pages 6849–6857, 2019.

[21] 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/CVF International Conference on Computer Vision*, pages 2471–2480, 2021.

[22] 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 *CVPR*, pages 17522–17531, 2022.

[23] 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 *ECCV*, pages 201–217. Springer, 2022.

[24] 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. *TPAMI*, 2020.

[25] Xiang Zhang, Chengzhe Lu, Dawei Yan, Wei Dong, and Qingsen Yan. Dualbln: Dual branch lut-aware network for real-time image retouching. In *ACCV*, pages 2139–2155, 2022.

[26] Zhaoyang Zhang, Yitong Jiang, Jun Jiang, Xiaogang Wang, Ping Luo, and Jinwei Gu. Star: A structure-aware lightweight transformer for real-time image enhancement. In *ICCV*, pages 4106–4115, 2021.