文章编号:1674-2974(2025)08-0014-09 DOI:10.16339/j.cnki.hdxbzkb.2025279

基于多模态学习的高动态范围图像色调映射

岳焕景 ,何长安 ,杨敬钰 †
(天津大学 电气自动化与信息工程学院, 天津 300072)

摘 要:针对现有色调映射技术在实际应用中所面临的若干关键挑战,诸如映射结果的稳定性不足、难以兼顾图像的自然美感以及对复杂光照环境与多元场景类型的适应性有限等问题,本文提出了一种基于多模态学习的色调映射方法,旨在通过共享文本与图像的语义空间,获取跨模态监督信息,以期实现更为精准、自然且具有普适性的色调映射.借助对比文本图像大模型的文本图像匹配信息辅助模型进行无监督训练,可以有效抑制欠曝区域和过曝区域的产生,避免了生成对抗方法和对比学习中存在的训练不稳定和训练复杂等问题.实验表明所提出的色调映射方法在多个开放基准数据集上都展现出较好的效果.与现有主流色调映射算法相比,该方法在保持图像整体光照氛围的同时,能更有效地抑制过曝区域,提升欠曝区域,保留丰富的色彩细节,增强视觉层次感,且对各类光照条件及场景类型的适应性更强.同时,本文的研究工作也证实了多模态学习在底层视觉任务中的巨大潜力.

关键词:色调映射;图像增强;无监督学习
中图分类号:TP37 **文献标志码**:A

High Dynamic Range Image Tone Mapping Based on Multimodal Learning

YUE Huanjing, HE Chang'an, YANG Jingyu†
(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: In response to several key challenges faced by the existing tone mapping techniques in practical applications, such as insufficient stability of mapping results, difficulty in balancing the natural aesthetics of images, and limited adaptability to complex lighting environments and diverse scene types, this paper proposes a tone mapping method based on multimodal learning. The goal is to acquire cross-modal supervisory information through the shared semantic space of text and images, aiming to achieve more accurate, natural, and universally applicable tone mapping. By leveraging the text-image matching information from large text-image models to assist in unsupervised training, the method effectively suppresses the occurrence of underexposed and overexposed areas, avoiding the training instability and complexity issues present in generative adversarial methods and contrastive learning. Experiments demonstrate that the proposed tone mapping method displays superior performance across multiple open benchmark datasets. Compared with the existing mainstream tone mapping algorithms, this method not only maintains the overall lighting atmosphere of images but also more effectively suppresses overexposed areas, enhances underexposed areas, retains rich color details, and enhances visual hierarchy, with stronger adaptability

* **收稿日期**:2024-02-04
基金项目:国家自然科学基金资助项目(62072331), National Natural Science Foundation of China(62072331)
作者简介:岳焕景(1987—),女,山东济宁人,天津大学副教授,博士
† 通信联系人,E-mail: yjy@tju.edu.cn

to various lighting conditions and scene types. Moreover, this work also confirms the significant potential of multimodal learning in foundational vision tasks.

Key words: tone mapping; image enhancement; unsupervised learning

随着科技发展,高动态范围(high dynamic range, HDR)成像技术在摄影、电影制作、游戏开发以及虚拟现实(VR)等领域被广泛应用.HDR成像能够捕捉到远超传统低动态范围(low dynamic range, LDR)设备所能再现的宽广亮度范围,提供更接近人眼所见的真实世界视觉效果.尽管如此,但常见显示设备并非全部支持HDR,大量手机、电视和电脑显示器仍然是LDR规格,无法直接呈现HDR图像.随着电子消费品市场对HDR内容呈现需求的增长,作为实现LDR显示屏适配HDR图像的关键手段,色调映射技术受到学术界与工业界的广泛关注与高度重视.

色调映射任务旨在通过压缩HDR图像场景亮度范围尽可能还原人眼观察到的真实场景效果.传统的色调映射方法可以分为全局算子和局部算子两类.全局算子[1-2]在所有图像上应用相同的映射关系.尽管这类算子非常简单高效,但往往导致局部出现大幅度的对比度衰减,造成局部细节丢失严重.局部算子[3-6]通过设计基于局部细节的自适应算子解决这个问题.这些研究成果显著提高了HDR图像的色调映射效果,同时保留了它们的局部对比度,并最大限度地减少了边缘相关的伪影.然而,由于不同局部区域对比度差异较大,高压缩水平仍然可能导致伪影,影响观感.因此,本文希望利用深度学习技术解决这一问题.

伴随着深度学习技术在图像处理领域的快速发展,许多学者尝试利用深度神经网络完成HDR色调映射.Hou等人[7]利用VGG感知损失函数[8],使网络学习从HDR图像到LDR图像的映射关系.然而他们的网络必须针对每个输入图像进行专门的训练.Gharbi等人[9]受双边网格的启发设计网络结构,通过学习给定输入和其增强结果之间的映射关系,致力于利用神经网络完成实时的对应图像增强操作.最近的一些工作[10-14]利用条件生成对抗网络[15](conditional generative adversarial network, CGAN)学习色调映射任务.然而CGAN训练要求配对的HDR图像和其色调映射后的LDR图像.这些工作的不同之处在于获取配对数据的方式.Rana等人[11]、Patel等人[13]和Cao等人[14]在HDR图像上应用了一系列传统色调映射算法并选择TMQI[16](tone mapped quality index)指标最高的结果作为训练标签.Zhang等人[12]利用三个专业人士为HDR图像提供色调映射后图像,并选择TMQI指标最高的作为训练标签.尽管上述方法已经取得了优于传统方法的结果,但通常导致在训练场景中过拟合,在其他场景中泛化能力较差.

色调映射任务真值具有很强的主观性,采用人工标注结果进行训练往往无法得到最佳结果.因此,以无监督学习为基础的训练方法更适合用于训练色调映射任务.Vinker等人[17]将基于非配对学习的生成对抗网络引入色调映射任务中,取得了很好的结果.然而该方法输出稳定性较差,在局部区域经常生成不自然的黑点.这本质上是因为判别器对场景亮度范围判别能力有限.Cao等人[18]引入了对比学习,在特征域引导网络进行学习,然而对比学习训练较为困难.

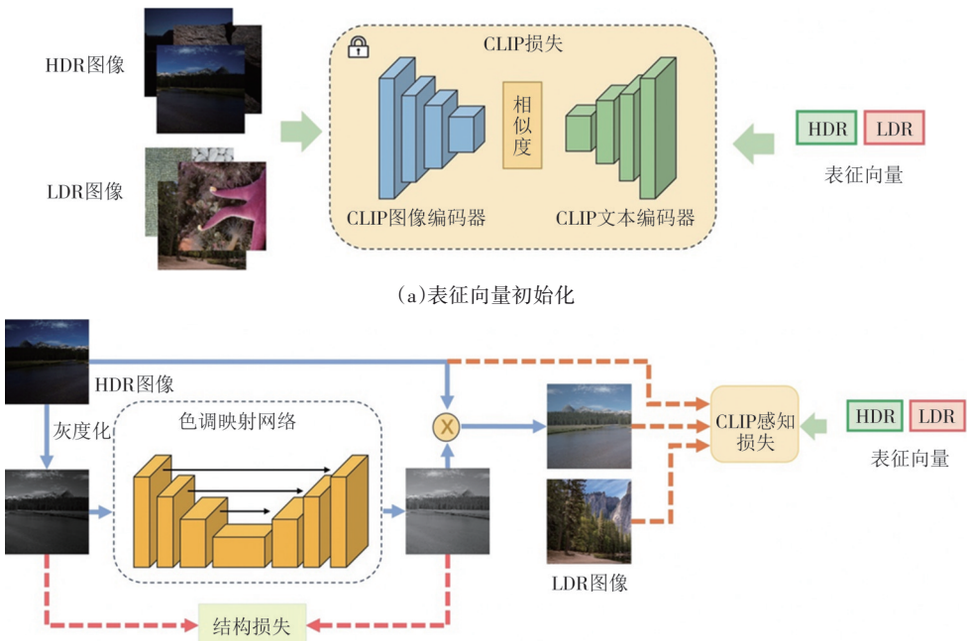
与上述方法不同,本文引入文本图像预训练模型辅助提升色调映射任务,旨在借助跨模态信息提升模型对亮度动态范围的理解能力,进而提升色调映射结果的稳定性.

1　基于多模态学习的色调映射

本文采用的是跨模态信息辅助的非配对学习方法.如图1所示,本文所提方法主要包括两个阶段:第一阶段是设计多模态损失函数,如图1(a)所示,旨在通过共享的文本和图像语义空间构建HDR和LDR图像对应的文本表达;第二阶段是多模态辅助色调映射,如图1(b)所示,旨在利用跨模态监督信息优化色调映射网络输出的美观和自然程度.

1.1　多模态损失设计

本文方法的第一阶段在于设计多模态损失,并



(a)表征向量初始化

(b)多模态辅助学习

图 1 模型整体结构

Fig.1 Architecture of the proposed model

利用非配对数据构建HDR和LDR图像对应的文本表达.本文使用CLIP模型[19]整合跨模态信息,即通过对比学习,使匹配的图像-文本对在特征空间中相对接近,而不匹配的图像-文本对则相对远离.这一机制使得模型能够有效关联文本特征与对应图像特征.基于这一机制,我们在网络训练第一阶段利用真实的HDR和LDR图像来训练并获取对应的文本表征向量.如图1(a)所示,本文使用预训练的CLIP模型[19]计算特征相似度.为不改变CLIP模型特征,CLIP图像编码器和文本编码器的权重在本文训练过程中都固定不变,不受网络训练影响.首先随机初始化HDR表征向量$\boldsymbol{P}_{\mathrm{H}} \in \mathbb{R}^{N\times 512}$和LDR表征向量$\boldsymbol{P}_{\mathrm{L}} \in \mathbb{R}^{N\times 512}$,其中$N$表示文本表征向量中的语素个数.然后将它们输入至训练好的CLIP文本编码器Φ_{test}中获取对应的潜在特征.同时,本文将HDR图像和LDR图像$I_{\mathrm{H}}, I_{\mathrm{L}} \in \mathbb{R}^{N\times W\times 3}$分别输入训练好的CLIP图像编码器$\Phi_{\mathrm{image}}$中获取对应的潜在特征.受已有研究工作[20]启发,本文基于CLIP特征空间中语义相同的文本图像之间特征相似的前提条件,对上述文本图像潜在特征计算余弦相似度,并利用二分类交叉熵损失函数迭代优化HDR表征向量$\boldsymbol{P}_{\mathrm{H}}$和LDR表征向量$\boldsymbol{P}_{\mathrm{L}}$,公式可表示为:

$$L_{\mathrm{initial}} = -\left(y \times \log\left(\hat{y}\right) + \left(1-y\right) \times \log\left(1-\hat{y}\right)\right) \tag{1}$$

式中:y代表当前输入图像标签;$\hat{y}$表示当前输入图像与HDR表征向量的相对相似概率,公式可表示如下.

$$\hat{y} = \frac{\mathrm{e}^{\cos\left(\Phi_{\mathrm{image}}(I),\, \Phi_{\mathrm{text}}(\boldsymbol{P}_{\mathrm{H}})\right)}}{\sum_{i\in\{\mathrm{H},\mathrm{L}\}} \mathrm{e}^{\cos\left(\Phi_{\mathrm{image}}(I_i),\, \Phi_{\mathrm{text}}(\boldsymbol{P}_i)\right)}} \tag{2}$$

式中:I代表当前输入图像;I_{H}对应标签y为1;I_{L}对应标签y为0.通过本阶段训练,优化表征向量与对应的图像在CLIP特征空间潜在特征相似,使$\boldsymbol{P}_{\mathrm{H}}$和$\boldsymbol{P}_{\mathrm{L}}$可以分别表示与HDR图像和LDR图像语义相同的文本信息.

1.2 多模态辅助色调映射

本文方法的第二阶段训练致力于优化色调映射网络.借助CLIP的共享图像文本特征空间,第一阶段训练得到的文本表征向量包含了不同类型图像对应的色调、亮度和其他视觉特征信息.为利用这一点,在网络训练第二阶段,本文利用表征向量构建CLIP损失函数,提升色调映射网络输出结果与LDR文本向量在上述共享特征空间的特征相似度,同时降低其与HDR文本向量的特征相似度,使得色调映射网络输出结果更为真实自然.具体而言,如图1(b)所示,本阶段将上一阶段训练得到的HDR和LDR表征向量与CLIP文本图像编码器组合形成CLIP感知损失.本文选用最常见的UNet[21]网络结构作为色调映射网络.与常见色调映射方法相同,本文网络只对

图像亮度层(灰度图)进行处理,致力于在保持色彩的前提下,压缩图像亮度动态范围,即本文以HDR图像亮度层$I_i \in \mathbb{R}^{H\times W\times 1}$为输入,以映射后LDR图像亮度层$I_o \in \mathbb{R}^{H\times W\times 1}$为输出.参考以往研究工作[18,22],本文使用HDR图像I_{H}和网络输出I_o生成LDR图像$\hat{I}_{\mathrm{L}}$,HDR图像中每个颜色通道被独立映射到估计的LDR图像中,公式表示如下:

$$\hat{I}_{\mathrm{L}}^{c}=\left(\frac{I_{\mathrm{H}}^{c}}{I_i}\right)^{s}\times I_o,\ \forall c\in\{r,g,b\} \tag{3}$$

式中:s为色调饱和参数,本文使用默认值0.5.为使网络输出与真实LDR图像相近,本文希望网络生成LDR图像$\hat{I}_{\mathrm{L}}$与第一阶段训练得到的LDR表征向量$\boldsymbol{P}_{\mathrm{L}}$在CLIP潜在特征空间中具有相似性,由此得到CLIP损失函数,公式表示如下:

$$L_{\mathrm{CLIP}}=\frac{\mathrm{e}^{\cos\left(\Phi_{\mathrm{image}}\left(\hat{I}_{\mathrm{L}}\right),\Phi_{\mathrm{text}}\left(\boldsymbol{P}_{\mathrm{L}}\right)\right)}}{\sum\limits_{i\in\{\mathrm{H,L}\}}\mathrm{e}^{\cos\left(\Phi_{\mathrm{image}}\left(\hat{I}_{\mathrm{L}}\right),\Phi_{\mathrm{text}}\left(\boldsymbol{P}_i\right)\right)}} \tag{4}$$

同时,色调映射任务应该只对图像亮度进行处理,而不改变其他信息.因此,本文使用结构损失函数L_{struct}来保证映射后LDR图像与网络输入HDR图像场景一致.由于结构相似性指标[23](structure similarity index measure, SSIM)对图像亮度和对比度变化敏感,不适用于约束色调映射任务,因此本文采用皮尔逊相关系数(Pearson correlation coefficient)来衡量两幅图像之间的结构相似度,公式表示如下:

$$\rho\left(I_1,I_2\right)=\frac{1}{n_p}\sum_{P_{I_1},P_{I_2}}\frac{\mathrm{cov}\left(P_{I_1},P_{I_2}\right)}{\sigma\left(P_{I_1}\right)\sigma\left(P_{I_2}\right)} \tag{5}$$

式中:P_{I_1}和P_{I_2}表示图像I_1和I_2的相同位置的图像块;$\mathrm{cov}(\cdot,\cdot)$和$\sigma(\cdot)$表示在这些图像块上计算的协方差和标准差;$n_p$表示图像块数量.由此定义结构损失,公式表示如下:

$$L_{\mathrm{struct}}=\sum_{k\in\{0,1,2\}}\rho\left(\downarrow^{k}I_{\mathrm{H}},\ \downarrow^{k}I_{\mathrm{L}}\right) \tag{6}$$

式中:$\downarrow^{k}$表示k次1/2下采样操作,当k为0时表示不进行下采样操作.结构损失函数L_{struct}通过约束多尺度输入输出的局部相似来保证场景信息的一致性.同时,本文利用CLIP图像编码器设计感知损失函数,约束输入输出图像在CLIP图像特征空间的特征相似,在保证场景一致性的前提下,提升网络输出观感,公式表示如下:

$$L_{\mathrm{perc}}=\sum_{l=0}^{4}\alpha_l\cdot\left\|\Phi_{\mathrm{image}}^{l}\left(I_i\right)-\Phi_{\mathrm{image}}^{l}\left(I_o\right)\right\|_2 \tag{7}$$

式中:α_l表示CLIP图像编码器第l层输出特征的权重.最终色调映射网络损失可表示如下:

$$L_{\mathrm{tone}}=L_{\mathrm{struct}}+\lambda_{\mathrm{CLIP}}\cdot L_{\mathrm{CLIP}}+\lambda_{\mathrm{perc}}\cdot L_{\mathrm{perc}} \tag{8}$$

式中:λ_{CLIP}和λ_{perc}分别表示CLIP损失权重和感知损失权重.

在网络训练的第一阶段,由于只使用了真实的HDR和LDR图像进行学习,表征向量只能模糊表征两类图像.而随着色调映射网络的不断优化,网络输出与上述图像存在较大差异,第一阶段训练得到的表征向量逐渐无法指导网络训练.所以本文在第二阶段借助色调映射网络输出对表征向量进行微调.本文希望在第一阶段表征训练的基础上,提升表征对各类图像的识别能力,拉近网络输出与LDR表征相似度,拉远网络输出与HDR表征相似度,所以定义表征损失函数如下:

$$\begin{aligned}L_{\mathrm{prompt}}=&\sum_{k\in\{0,1,2\}}\left(\max\left(0,S\left(\downarrow^{k}I_{\mathrm{L}}\right)-S\ \downarrow^{k}I_{\mathrm{H}}+m_0\right)\right)+\\&\left(\max\left(0,S\left(\downarrow^{k}\hat{I}_{\mathrm{L}}\right)-S\left(\downarrow^{k}I_{\mathrm{H}}\right)+m_0\right)\right)+\\&\left(\max\left(0,S\left(\downarrow^{k}I_{\mathrm{L}}\right)-S\left(\downarrow^{k}\hat{I}_{\mathrm{L}}\right)+m_1\right)\right)\end{aligned} \tag{9}$$

式中:$I_{\mathrm{H}},I_{\mathrm{L}},\hat{L}_{\mathrm{L}}$分别表示输入的HDR图像、输入的LDR图像、色调映射网络输出的LDR图像;$\downarrow^{k}$表示k次1/2下采样操作.通过多尺度训练,表征向量可以更准确地感知不同亮度范围的图像特征,提升图像辨别能力.$S(\cdot)$表示输入图像与HDR表征相似的分数,公式表示如下:

$$S(I)=\frac{\mathrm{e}^{\cos\left(\Phi_{\mathrm{image}}(I),\Phi_{\mathrm{text}}\left(\boldsymbol{P}_{\mathrm{H}}\right)\right)}}{\sum\limits_{i\in\{\mathrm{H,L}\}}\mathrm{e}^{\cos\left(\Phi_{\mathrm{image}}(I),\Phi_{\mathrm{text}}\left(\boldsymbol{P}_i\right)\right)}} \tag{10}$$

当图像I被表征向量认为是HDR图像时,$S(I)=1$;当图像I被表征向量认为是LDR图像时,$S(I)=0$,因此$S(\cdot)$可被认为是根据表征向量判断输入图像为HDR图像的概率.本文实验中,m_0取0.9,表示要求的文本特征对真实LDR图像和真实HDR图像的分类概率差异下限,也表示网络输出和真实HDR图像的分类概率差异下限,希望文本表征可以在保证对真实图像正确判断的前提下有效分类网络输出结果.同时,m_1取0.2,表示要求的文本表征对真实LDR图像和网络输出的分类差异下限.在微调文本表征中,并不希望文本表征对真实LDR图像和网络输出的分类概率完全一致,否则在后续优化网络时, CLIP损失将趋于0,失去优化空间.本文希望利用CLIP损失和prompt损失交替训练色调映射网络和文本表征,

进而提升文本表征对HDR和LDR图像的特征判断的准确性,同时指导色调映射网络输出更优结果.

2 实验结果与分析

2.1 整体训练流程

本文训练主要分为两个阶段.第一阶段是表征向量初始化训练阶段,如图1(a)所示.第二阶段是色调映射网络与表征向量交替训练阶段,如图1(b)所示.在第一阶段,本文使用固定长度的字符串作为初始化文本,通过文本编码器生成相应的文本特征,并通过梯度回传不断调整这些文本特征,使其能够有效表示对应的图像信息.在第二阶段,本文交替训练色调映射网络和文本表征向量.首先利用第一阶段训练好的文本表征向量构建CLIP损失函数,通过最小化L_{tone}训练色调映射网络,然后通过最小化L_{prompt}优化两类表征向量.两种优化交替进行,直至网络收敛.

2.2 训练细节及参数设置

参考以往工作[17],本文使用HDR+数据集[24]中的1 000张HDR图像和DIV2K数据集[25]中的1 000张曝光良好的LDR图像作为色调映射任务的训练数据.在输入色调映射网络前,本文利用裁剪和缩放将每张训练图像的图像大小都调整为256×256.本文在训练过程中色调映射网络和表征向量均采用Adam优化器.在表征向量初始化训练阶段,本文采用固定学习率为2×10^{-4},训练周期为150,输入图像批次大小为64.在色调映射网络与表征向量交替训练阶段,初始学习率均设定为2×10^{-4},并且通过余弦退火算法最后衰减至5×10^{-5},训练周期为200,输入图像批次大小为24.色调映射网络中CLIP损失权重λ_{CLIP}和感知损失权重λ_{perc}分别设定为32和54.在皮尔逊相关系数计算中,滑窗大小设定为5×5,并且滑窗步长大小设定为1.在感知损失函数L_{perc}中,通过本文观察,深层特征通常与图像色彩信息关系紧密,而色调映射任务对图像色调深浅有较大影响,所以设定各层输出特征损失权重为$[\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5]=[1,1,1,1,0.5]$.本文所提方法基于PyTorch框架实现,并且利用一块NVIDIA 3090显卡进行训练.

2.3 对比实验

为了展示本文所提方法的先进性,本文与领域内5种先进方法进行了实验对比.为了验证所提方法的泛化性能,本文采用了三个训练阶段未见的数据集用于测试,即LVZ-HDR数据集[22],HDRI Haven数据集[26]和HDR Survey数据集[27].本文用于检验HDR图像色调映射性能的指标包括TMQI[16]和BTMQI[28].TMQI[16]以HDR图像和映射后LDR图像为输入,从保真度和自然度两方面衡量色调映射质量.BTMQI[28]在不依赖于原始HDR数据或任何参考图像的情况下,对经过色调映射处理生成的LDR图像进行质量评价.在比较方法中,CLIP-LIT[20]被提出用于背光图像增强,同样利用多模态损失和感知损失处理非配对数据.但与本文不同的是,CLIP-LIT以RGB三通道图像作为输入,且并未考虑结构损失和多尺度信息,更加关注图像语义信息.TMONet[22]是基于监督学习的色调映射方法,由于其只向外界提供了模型在LVZ-HDR数据集上的测试结果,未开源工程代码,因此本文只在LVZ-HDR数据集上与TMONet进行比较.TMOCAN[29]是基于无监督学习的色调映射方法,只需要使用HDR数据即可训练.由于其未提供训练权重文件,为了进行公平对比,本文按照其所述方法在HDR+数据集上对TMOCAN进行了训练.UnpairedTMO[17]和UnCLTMO[18]是基于非配对学习的色调映射方法,与本文所提方法相同,利用非配对的HDR图像和LDR图像进行训练.

不同方法在三个HDR数据集上的定量比较结果分别如表1、表2和表3所示.不难发现,本文所提方法在各个测试数据集中BTMQI指标都取得了最好的结果,TMQI指标也仅次于UnCLTMO的结果.UnCLTMO方法在训练过程中,利用TMQI指标指导对比学习,对结果的TMQI指标有较好的提升作用,但在BTMQI指标中表现不及本文.

表1 LVZ-HDR数据集对比实验

Tab.1 Comparisons on LVZ-HDR dataset

模型	TMQI ↑	BTMQI ↓
CLIP-LIT[20]	0.825	3.83
TMONet[22]	0.873	3.44
TMOCAN[29]	0.877	3.32
UnpairedTMO[17]	0.883	3.04
UnCLTMO[18]	**0.899**	2.91
本文模型	0.896	**2.82**

可以观察到,在同样引入多模态信息的CLIP-LIT的实验结果中,TMQI和BTMQI都表现较差.原因在于,色调映射任务与低照度图像增强任务不同,HDR图像与LDR图像存在更加显著的亮度差异,而

表 2　HDRI Haven 数据集对比实验

Tab.2　Comparisons on HDRI Haven dataset

模型	TMQI ↑	BTMQI ↓
CLIP-LIT[20]	0.734	4.42
TMOCAN[29]	0.796	4.16
UnpairedTMO[17]	0.902	3.24
UnCLTMO[18]	**0.916**	3.13
本文模型	0.906	**3.12**

表 3　HDR Survey 数据集对比实验

Tab.3　Comparisons on HDR Survey dataset

模型	TMQI ↑	BTMQI ↓
CLIP-LIT[20]	0.782	4.31
TMOCAN[29]	0.899	2.94
UnpairedTMO[17]	0.919	2.89
UnCLTMO[18]	**0.932**	2.83
本文模型	0.930	**2.74**

感知损失相对结构损失对亮度信息更加敏感,只利用感知损失约束色调映射网络输入输出一致,极大地限制了网络对图像亮度的调整,同时抑制了跨模态信息的辅助作用.而本文方法只在亮度层上进行处理,同时引入结构损失,并在结构损失和表征损失中都采用多尺度输入,在保证结构一致的前提下,给予网络更大的亮度调整空间,并利用跨模态信息监督,可以很好地完成色调映射任务.

图 2、图 3 和图 4 分别展示了不同方法在三个 HDR 数据集上的输出结果对比.其中,CLIP-LIT 输出结果过暗,原因如上文所述.TMONet 输出图像色彩表现较差且不自然.TMOCAN 方法复杂度较低,但输出结果过饱和严重,场景色彩出现了明显偏差.UnpairedTMO 和 UnCLTMO 方法表现较好,但对场景的理解能力较差,对同一物体经常输出差异的亮度信息,导致物体出现不自然的亮度差异,如图 2 中的垃圾桶、房屋,图 3 中的天空等.此外,其倾向于在高光区域产生褪色效应,即一定程度的过曝,如图 2 中的南瓜、图 4 中的牛形雕塑等.而本文所提方法在视觉效果上可以取得较好的结果,说明了在多模态模型的辅助下,本文方法可有效提升被摄物体的自然性以及亮度一致性,可以更真实地还原观看体验.本文在表 4 中给出了不同方法的复杂度、参数量、推理时间和显存占用的比较.本文方法相较于 UnpairedTMO 和 UnCLTMO 具有更低的复杂度和参数量,同时具有更快的推理速度,内存占用也略低.这说明了本文所提方法具有更高的计算效率.



(a)CLIP-LIT　(b)TMONet　(c)TMOCAN　(d)UnpairedTMO　(e)UnCLTMO　(f)本文模型

图 2　不同方法在 LVZ-HDR 数据集上的测试结果

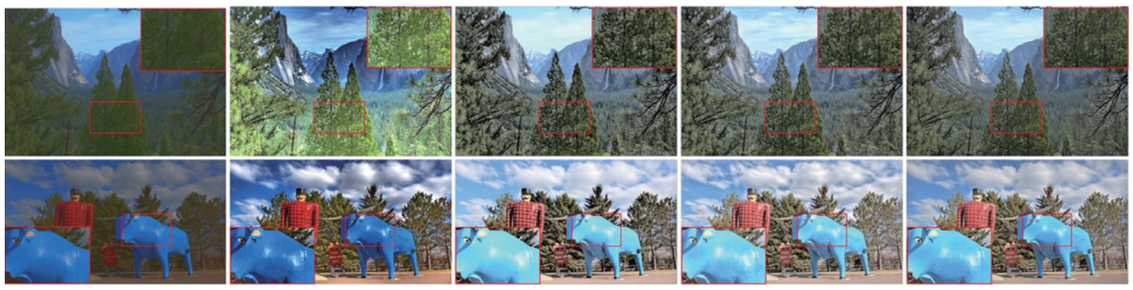
Fig.2　Test results of different methods on LVZ-HDR dataset



(a)CLIP-LIT　(b)TMOCAN　(c)UnpairedTMO　(d)UnCLTMO　(e)本文模型

图 3　不同方法在 HDRI Haven 数据集上的测试结果

Fig.3　Test results of different methods on HDR Haven dataset



(a)CLIP-LIT (b)TMOCAN (c)UnpairedTMO (d)UnCLTMO (e)本文模型

图4 不同方法在HDR Survey数据集上的测试结果

Fig.4 Test results of different methods on HDR Survey dataset

表4 复杂度比较

Tab.4 Computing complexity comparison

模型	MACs /G	参数量/M	推理时间/ms	内存占用/GB
CLIP-LIT[20]	18.25	0.28	6.78	1.31
TMOCAN[29]	3.27	0.07	2.12	1.22
UnpairedTMO[17]	10.06	4.49	12.68	1.37
UnCLTMO[18]	10.11	4.89	12.84	1.49
本文模型	8.26	4.05	5.40	1.29

2.4 消融实验

为了体现所提方法的有效性,本文在HDR Survey数据集上进行了损失函数消融实验,实验结果如表5所示,效果对比如图5所示.首先,将不引入跨模态信息辅助的模型定为基准模型,即表5中的"w/o clip loss"模型.其次,在引入共享的文本和图像语义空间后,对微调表征向量进行消融,即表5中的"w/o prompt loss"模型.



(a)w/o clip loss (b)w/o prompt loss (c)本文模型

图5 在HDR Survey数据集上的消融实验结果

Fig.5 Results of ablation experiments on HDR Survey dataset

表5 损失函数消融实验

Tab.5 Ablation experiment on the loss function

模型	TMQI ↑	BTMQI ↓
w/o clip loss	0.896	3.85
w/o prompt loss	0.907	3.44
本文模型	0.930	2.73

实验结果显示,在仅采用结构损失的情况下,模型能够产生一个初步的色调映射结果,但性能有限.此时模型缺少对场景的理解能力,对复杂光照调节能力不足,常常导致部分区域欠曝或过曝.在引入多模态损失函数并且保持预初始化的表征向量不变的情况下,性能增益并不显著.原因在于,初始化表征向量仅通过真实的HDR图像和LDR图像进行训练,无法对网络训练中的输出结果进行有效判断.而利用L_{prompt}微调文本表征后,可以很好借助文本表征指导色调映射网络输出更优结果.

本文在HDR Survey数据集上进行了损失函数权重消融实验,实验结果如表6所示.CLIP损失函数旨在提升生成图像的自然程度,而感知损失旨在增强图像在场景信息上的保真度.通过权重参数设定,平衡二者之间关系,使模型输出达到最佳效果.

另外,本文也尝试在网络训练第一阶段使用不同的字符串作为初始变量进行实验,但最后经过梯度更新后会收敛到类似的结果.这表明我们的方法

表6　损失函数权重消融实验

Tab.6　Ablations experiment on the weights of loss functions

CLIP损失权重	感知损失权重	TMQI↑	BTMQI↓
32	27	0.910	3.03
64	54	0.912	2.98
32	54	0.930	2.73

对不同的初始化具有一定鲁棒性.

2.5　用户主观评测

为更好地说明所提方法相对其他方法的先进性,本文开展了用户研究,各方法偏好率如表7所示.为减少用户评测压力,本文在上述三个测试集中随机抽取40组图像进行比较,用户被要求挑选最喜欢的结果.参与调研的用户人数为25位.结果显示,有33%的用户更喜欢我们方法的输出结果,高于第二名的29%.这些结果表明,本文的色调映射结果更符合用户的主观偏好.

表7　用户主观评测结果

Tab.7　Results of user study

模型	喜爱率/%
TMONet[22]	7
TMOCAN[29]	3
UnpairedTMO[17]	28
UnCLTMO[18]	29
本文模型	33

3　结 论

本文针对色调映射问题,提出了一种创新性的多模态学习方法,旨在通过深度融合文本与图像的跨模态信息,提升图像色调映射的准确性和自然度,同时增强其适应复杂光照条件和场景变化的能力.本文通过引入多模态信息先验知识,辅助色调映射网络优化映射结果,相较于生成对抗和对比学习方法,所提方法训练更容易,且可以实现更加稳定自然的增强效果.视觉对比结果显示了本文方法的优越性,充分说明了多模态学习的有效性,也为后续更多基于多模态学习的底层视觉技术研究提供了重要的参考价值.

参考文献

[1] KALANTARI N K, RAMAMOORTHI R. Deep high dynamic range imaging of dynamic scenes [J]. ACM Transactions on Graphics, 2017, 36(4): 1-12.

[2] DRAGO F, MYSZKOWSKI K, ANNEN T, et al. Adaptive logarithmic mapping for displaying high contrast scenes [J]. Computer Graphics Forum, 2003, 22(3): 419-426.

[3] FATTAL R, LISCHINSKI D, WERMAN M. Gradient domain high dynamic range compression [M]//Seminal graphics papers: pushing the boundaries: volume 2. New York, NY, USA: ACM, 2023: 671-678.

[4] REINHARD E, STARK M, SHIRLEY P, et al. Photographic tone reproduction for digital images [M]//Seminal graphics papers: pushing the boundaries: volume 2. New York, NY, USA: ACM, 2023: 661-670.

[5] LIANG Z T, XU J, ZHANG D, et al. A hybrid l1-l0 layer decomposition model for tone mapping [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 4758-4766.

[6] SHIBATA T, TANAKA M, OKUTOMI M. Gradient-domain image reconstruction framework with intensity-range and base-structure constraints [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 2745-2753.

[7] HOU X X, DUAN J, QIU G P. Deep feature consistent deep image transformations: downscaling, decolorization and HDR tone mapping [EB/OL]. (2017-09-11) [2024-01-10] 2017: 1707.09482. https://arxiv. org/abs/1707.09482v2.

[8] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution [M]//Computer vision-ECCV 2016. Cham: Springer International Publishing, 2016: 694-711.

[9] GHARBI M, CHEN J W, BARRON J T, et al. Deep bilateral learning for real-time image enhancement[J]. ACM Transactions on Graphics, 2017, 36(4): 1-12.

[10] MONTULET R, BRIASSOULI A. Deep learning for robust end-to-end tone mapping [C]//30th British Machine Vision Conference. BMVA Press, 2020: 160373.

[11] RANA A, SINGH P, VALENZISE G, et al. Deep tone mapping operator for high dynamic range images[J]. IEEE Transactions on Image Processing, 2019, 29: 1285-1298.

[12] ZHANG N, WANG C, ZHAO Y, et al. Deep tone mapping network in HSV color space[C]//2019 IEEE Visual Communications and Image Processing (VCIP). Sydney, Australia. IEEE, 2019: 1-4.

[13] PATEL V A, SHAH P, RAMAN S. A generative adversarial network for tone mapping HDR images [M]//Computer vision, pattern recognition, image processing, and graphics. Singapore: Springer Singapore, 2018: 220-231.

[14] CAO X, LAI K, YANUSHKEVICH S N, et al. Adversarial and adaptive tone mapping operator for high dynamic range images [C]//2020 IEEE Symposium Series on Computational Intelligence (SSCI). Canberra, ACT, Australia. IEEE, 2020: 1814-1821.

[15] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Honolulu, HI, USA. IEEE, 2017: 5967-5976.

[16] YEGANEH H, WANG Z. Objective quality assessment of tone-mapped images[J]. IEEE Transactions on Image Processing, 2013, 22(2): 657-667.

[17] VINKER Y, HUBERMAN-SPIEGELGLAS I, FATTAL R. Unpaired learning for high dynamic range image tone mapping[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada. IEEE, 2021: 14637-14646.

[18] CAO C, YUE H J, LIU X, et al. Unsupervised HDR image and video tone mapping *via* contrastive learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(2): 786-798.

[19] RADFORD A, KIM J W, HALLACYal C, et al. Learning transferable visual models from natural language supervision[C]// International conference on machine learning. PMLR, 2021: 8748-8763.

[20] LIANG Z X, LI C Y, ZHOU S C, et al. Iterative prompt learning for unsupervised backlit image enhancement[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France. IEEE, 2023: 8060-8069.

[21] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[M]//Medical image computing and computer-assisted intervention-MICCAI 2015. Cham: Springer International Publishing, 2015: 234-241.

[22] PANETTA K, KEZEBOU L, OLUDARE V, et al. TMO-net: a parameter-free tone mapping operator using generative adversarial network, and performance benchmarking on large scale HDR dataset[J]. IEEE Access, 2021, 9: 39500-39517.

[23] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[24] HASINOFF S W, SHARLET D, GEISS R, et al. Burst photography for high dynamic range and low-light imaging on mobile cameras[J]. ACM Transactions on Graphics, 2016, 35(6): 1-12.

[25] AGUSTSSON E, TIMOFTE R. NTIRE 2017 challenge on single image super-resolution: dataset and study[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu, HI, USA. IEEE, 2017: 1122-1131.

[26] Hdri-haven[EB/OL]. (2024-11-3)[2025-05-13]. https://hdri-haven.com/#google_vignette.

[27] FAIRCHILD M D. The HDR photographic survey[J]. Color and Imaging Conference, 2007, 15(1): 233-238.

[28] GU K, WANG S Q, ZHAI G T, et al. Blind quality assessment of tone-mapped images *via* analysis of information, naturalness, and structure[J]. IEEE Transactions on Multimedia, 2016, 18(3): 432-443.

[29] LE C Y, YAN J B, FANG Y M, et al. Perceptually optimized deep high-dynamic-range image tone mapping[EB/OL]. (2021-09-11)[2024-01-10]. https://arxiv.org/abs/2109.00180v3.