Modeling and Simulation 建模与仿真, 2024, 13(3), 3491-3506
Published Online May 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.133318

Hans汉斯

基于Transformer的U型低照度图像增强算法

缪天恒[1]，梁　艳[2]，王　飞[1]，丁德锐[1]

[1]上海理工大学光电信息与计算机工程学院，上海
[2]上海理工大学管理学院，上海

收稿日期：2024年4月28日；录用日期：2024年5月22日；发布日期：2024年5月31日

摘　要

针对非均匀光照环境下照度自适应算法的固有缺陷以及基于CNN的图像增强模型固有的卷积运算导致的感受野受限、无法建立长距离的全局依赖等问题，本文提出一种融合CA-Transformer模块和全局注意力融合模块的U型网络RT-UNet。本研究设计了基于轴向多头自注意力机制的CA-Transformer模块作为特征提取和重建的基础模块。该模块在兼顾CNN与Transformer结构优点的同时极大地减少了计算复杂度。进而，为了建立不同尺度，不同分辨率特征图之间的信息交互与融合，搭建了全局注意力融合模块来替代之前的残差连接，帮助网络关注到更感兴趣的区域，方便其学习到更有用，更精细的特征。实验结果表明，本算法在主客观评价指标上相比于近几年一些主流图像增强算法均具有很强的竞争力。

关键词

低照度增强，Transformer，注意力机制，特征融合

U-Shaped Low-Illumination Image Enhancement Algorithm Based on Transformer

Tianheng Miao[1], Yan Liang[2], Fei Wang[1], Derui Ding[1]

[1]School of Optoelectronic Information and Computer Engineering, University of Shanghai for Science and Technology, Shanghai
[2]Business School, University of Shanghai for Science and Technology, Shanghai

Received: Apr. 28th, 2024; accepted: May 22nd, 2024; published: May 31st, 2024

Abstract

To solve the inherent shortcoming of adaptive algorithms on illuminance in non-uniform illumi-

文章引用: 缪天恒, 梁艳, 王飞, 丁德锐. 基于 Transformer 的 U 型低照度图像增强算法[J]. 建模与仿真, 2024, 13(3): 3491-3506. DOI: 10.12677/mos.2024.133318

nation environments, the limited receptive field and the inability to establish long-distance global dependence caused by the inherent convolution operation of the CNN-based image enhancement models, this paper proposes a U-shaped network RT-UNet that integrates both CA-Transformer modules and global attention fusion modules. Specifically, a CA-Transformer module based on the axial multi-head self-attention mechanism is designed as the basic module for feature extraction and reconstruction. This module takes into account the advantages of CNN and Transformer structures while greatly reducing the computational complexity. To establish the information interaction and fusion between feature maps of different scales and different resolutions, a global attention fusion module is constructed to replace the previous residual connection, Such a module makes the network pay attention to the region of more interest and facilitates it to learn more useful and refined features. Finally, experimental results show that the proposed algorithm has strong competitiveness compared with some mainstream image enhancement algorithms in recent years in terms of subjective and objective evaluation indexes.

Keywords

Low Illumination Enhancement, Transformer, Attention Mechanism, Feature Fusion

1. 引言

在当前社会中，图像已成为信息交流和媒体传播的主要载体之一，而其质量的优劣直接关系到信息的有效传递和理解。但是由于采集时间、环境和角度等多种因素的影响，图像会难以避免的出现亮度不足的情况。这类低照度图像通常可分为两种情形：一是整体亮度较低，处于相对黑暗的环境；二是在背光等复杂光照条件下，主体和背景的对比度较高，呈现出明显的亮度差异。尤其在背光等复杂光照环境下，图像往往呈现局部光照不均的现象，如图 1 所示。

Figure 1. Low illumination images with uneven lighting and their normal light images
图 1. 光照不均匀的低照度图像及其正常光图像

近年来，随着深度学习技术的迅猛发展，卷积神经网络(CNN)在处理和恢复整体亮度较低的图像方面取得显著进展。然而，针对非均匀光照环境的图像增强[1]仍是一个严峻的挑战。传统的图像增强算法主要面向相对均匀的光照条件，难以有效处理背光等复杂场景。在处理非均匀光照环境下的照片时，现有的大部分主流卷积神经网络模型也显现出一些困境。首先，CNN 的卷积操作受限于固有的局部感受野，仅能在有限像素区域内建立特征关联，难以捕捉到长距离的全局依赖特征。这使得其在局部光照不均的情况下容易产生误差和失真。其次，由于主体对象和背景区域在光照条件上的显著差异，CNN 很难有效地学习并平衡这些局部不均的光照情况，导致输出图像难以达到理想的视觉效果。

鉴于以上 CNN 在低照度图像增强领域的局限性，引入 Transformer 结构[2]为低照度图像增强任务提供了一种新的思路。Transformer 借助其自注意力层的全局信息捕捉机制，能够有效地处理图像中远距离特征的编码，并灵活地关注不同区域及其上下文信息。因此，相对于 CNN，Transformer 在处理非均匀光照环境下的低照度图像增强任务中，具备更强的全局感知和上下文依赖能力。结合 U 型网络[3]和 Transformer 结构，本研究提出了一种基于 Transformer 的 U 型低照度图像增强模型，主要创新点如下：

1) 通过融合 CNN 与 Transformer 结构，设计了一种新的低照度图像增强网络模型 RT-UNet，兼顾了二者的优点。在维持局部信息挖掘能力的同时，强化了全局信息捕捉和长距离特征编码能力；

2) 提出了新的交叉轴 Transformer 模块作为 U 型网络编码器解码器的核心组件，既强化了整体网络的全局感知能力，又极大程度的减小了多头自注意力机制的计算复杂度；

3) 设计了全局注意力融合模块，强化了 U 型网络编码器和解码器之间信息传递的能力，确保了 CNN 提取的高分辨率空间特征与 Transformer 捕获的全局上下文特征的有效整合。

2. 相关工作

近年来，深度学习在计算机视觉领域取得了巨大进展，并成功应用于多项任务，如图像分类、语义分割、目标识别和人体关键点检测。在低照度图像增强领域，也出现了众多基于深度学习的创新算法。Jiang 等人[4]提出了 LL-Net 模型，其利用下采样中的全部可用信息生成高分辨率的增强结果，并通过逐步细化从深层提取的全局特征和浅层卷积产生的局部特征来减少噪声，提高对比度。Wei 等人[5]收集整理了低照度－正常照度图像对的 LOL 数据集，同时基于 Retinex 理论提出了 Retinex-Net，该网络包括将图像分解为光照分量和反射分量的 Decom-Net 以及调整光照分量的 Enhance-Net。其在训练中使用一致反射率和光照的平滑性作为约束，以获得更好的增强效果。Guo 等人提出了 Zero-DCE 网络[6]，通过设计一组非参考损失函数，使得深度学习模型能够摆脱对数据集的依赖，从而提高了模型的泛化性。这一方法为图像增强领域的发展提供了新的思路。Shi 等人[7]提出了一种基于 UNet++ [8]的新型低光照图像增强算法，其通过嵌套跳跃连接和特定的基于 Instance Normalization 的残差块来提高图像亮度、减少色彩失真，并保留更多细节。同时该研究还引入了一种新的混合损失函数，使其在多个指标上实现更优的增强效果。

Zhang 等人提出了一种结构感知的双分支轻量级模型 STAR [9]，通过将特征序列沿通道维度分为两部分，并分别送入基于 CNN 和 Transformer 的分支进行处理，从而降低了整体计算量。Souibgui 等人[10]提出了一种基于视觉变换器的端到端文档图像增强方法，该方法通过直接在像素块上操作并利用自注意力机制捕捉全局依赖性，有效提升了退化文档图像的质量。这在文档图像增强领域是首次尝试实现一种纯粹的基于 Transformer 的编解码架构。Jiang 等人提出了一种纯 Transformer 的生成对抗网络 TransGAN [11]，通过网格化的自注意力机制和分阶段迭代提高分辨率的方式，建立了一个内存友好型的生成器，实现了对高质量样本图像的恢复和增强。

综上所述，深度学习技术，尤其是卷积神经网络(CNN)和 Transformer 架构，在低照度图像增强方面

具有巨大的应用价值和发展潜力。通过结合全局和局部特征、设计新颖的损失函数、以及利用注意力机制捕捉图像的细节信息，这些算法能够在不依赖于大量标记数据的情况下，有效地提高图像的亮度、对比度和细节，减少噪声和色彩失真。尤其是近几年 Transformer 结构的引入为图像质量的提升提供了新的思路和可能性。通过全局自注意力机制，Transformer 能够捕捉图像中的长距离依赖关系，这对于增强低照度图像中的纹理和细节至关重要。同时，Transformer 的并行化处理能力和灵活性也为设计高效、适应性强的图像增强模型提供了便利。随着研究的深入，可以预见 Transformer 结构将在低照度图像增强领域扮演越来越重要的角色，并推动相关技术的进一步发展。

3. 本文方法

3.1. 网络结构

3.1.1. 总体网络模型

为了综合 CNN 和 Transformer 结构二者在低照度图像增强领域上的优势，本研究设计了一种融合 U 型网络和交叉轴 Transformer 模块的新型图像增强模型，具体结构如下图 2 所示。该网络输入图像为 RGB 三通道的低照度图像，先送入串接的交叉轴 Transformer 模块(CA-Transformer Block)，进行简单的特征提取和信息固化，然后特征序列被送入改进的 U 型主网络，经过网络中四层编码器的特征提取和下采样，不同分辨率的特征在全局注意力融合模块(GAF Block)中进行信息交互后，分别送入四层解码器进行特征重构得到增强后的特征序列，最后经过 3 个交叉轴 Transformer 模块重建为增强后的图片。

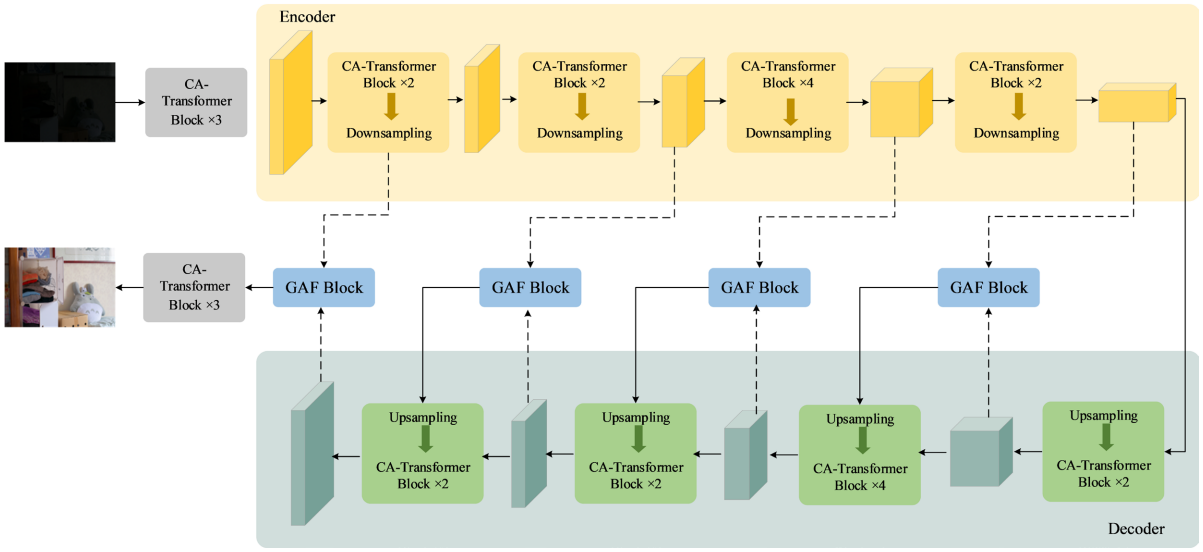

Figure 2. The network model of RT-UNet
图 2. RT-UNet 的网络模型图

具体来说，U 型主网络前后分别串接了三个交叉轴 Transformer Block，一方面使送入网络的图像的信息得到更好的固化保留，另一方面对输出图像中的噪声起到一定的抑制。考虑到 CNN 的平移不变性和捕捉长距离依赖能力的不足，本研究在普通 U 型网络的下采样与上采样模块的基础上串接了一定数量的交叉轴 Transformer Block，在维持了原先特征图尺寸的同时，加深了网络结构，强化了下采样过程中的特征提取和保持能力，使上采样过程的特征重构更细节、全面。同时为了补偿 Transformer 结构带来的特征分辨率减低的问题，这里在模块中部分保留了 CNN 的结构，确保 CNN 特征中的具有高分辨率的空间

信息和 Transformer 编码的全局上下文信息都得到更好地保留。交叉轴 Transformer Block 提取的自注意力特征经过下采样后，通过跳跃连接和上一层不同分辨率的自注意力特征一起送入全局注意力融合模块，来进行不同分辨率特征的融合，使浅层的局部信息和深层的语义知识可以充分交互学习。在解码器中，将所得到融合后的特征送入交叉轴 Transformer Block 进行逐层级联上采样来逐步还原图像的分辨率，并得到增强后的特征序列，最终经过 3 个串接的交叉轴 Transformer Block，恢复成增强后的图像。

3.1.2. 交叉轴 Transformer 模块

相较于 CNN 模型，Transformer 结构已经在大部分图像任务中被证实具备更强大的全局感知和上下文依赖能力。然而，其在局部细节的把握和计算成本上仍存在较大问题。针对 Transformer 带来的特征分辨率降低及局部细节的保持和恢复上的问题，设计了新的基于通道多头自注意力的交叉轴 Transformer (CrossAxis Transformer)模块，具体结构图见图 3。本模块采用双残差结构将浅层网络信息不断引入深层网络，来弥补特征分辨率的损失。首先，输入的特征图经过 1 × 1 卷积进行通道降维来节省计算资源。针对困扰 Transformer 结构在小模型上部署的最大障碍，本研究采用并行连接高度和宽度轴向多头自注意力模块并在多头自注意力相似性计算时交叉连接，确保建模的水平和垂直方向上的特征描述符可以交互学习，降低了特征图的计算成本，同时在两个多头注意力机制模块前分别用三个尺度的卷积操作提取了 x 轴或 y 轴上不同尺度的特征并融合，极大地提升了模型的表达能力，同时模型可以关注并学习不同尺度的特征，对后续细节，颜色，风格信息的恢复和重建提供针对性的特征指导。

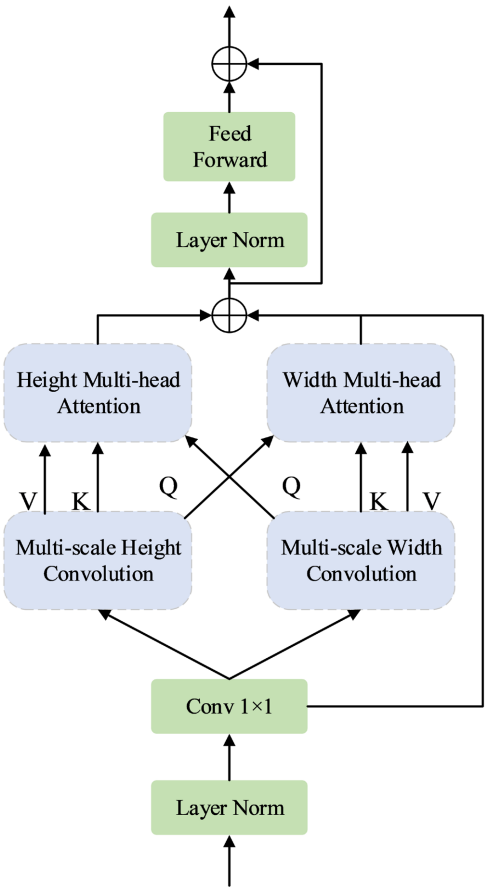

Figure 3. The structure of the CA-Transformer module
图 3. CA-Transformer 模块的结构图

具体来说，特征图经过 1 × 1 卷积进入交叉轴 Transformer 主模块后被分为两个平行分支，分别计算高度和宽度轴向关注。每个分支进入由三个不同大小核的一维卷积组成的多尺度高度卷积或多尺度宽度卷积，然后沿一个空间维度编码多尺度上下文信息，再沿另一个空间维度进行跨轴注意力聚集特征。这样设计的目标是利用卷积的力量来捕获多尺度特征表示。通过将多尺度特征融入轴向多头注意力中，使网络在特征提取和重建时关注图像中不同区域里大小不同的物体和细节信息。在低照度图像增强任务中捕获空间结构或形状信息至关重要。因此，本模型在两个空间维度之间建立双交叉多头注意力，来更好地利用从高度轴向注意力和宽度轴向注意力中提取的方向信息。

交叉轴 Transformer 模块中的轴向多头注意力被认为是多头自注意力的一种替代，它将自注意力分解为两部分，分别负责计算沿水平或垂直维度的自注意力。基于轴向的多头注意力可以沿着水平和垂直方向依次聚合特征，使捕获全局信息成为可能。给定大小为 $H \times W$ 的特征图作为输入，送入标准的自注意力模块进行运算，其相似性运算的计算成本为 $O\left(W^2H^2\right)$，而通过交叉并联的高度和宽度轴向自注意力模块，得到其相似性运算的计算成本为 $O\left(HW \times \left(H+W\right)\right)$。由此可以发现，通过交叉并联的高度和宽度轴向自注意力模块的方式进行相似性计算可以将注意力模块的计算成本大大削减一个量级，将计算复杂度从平方级难度转化为了线性级难度。因此，轴向注意力比自注意力更有效，计算也更简单。

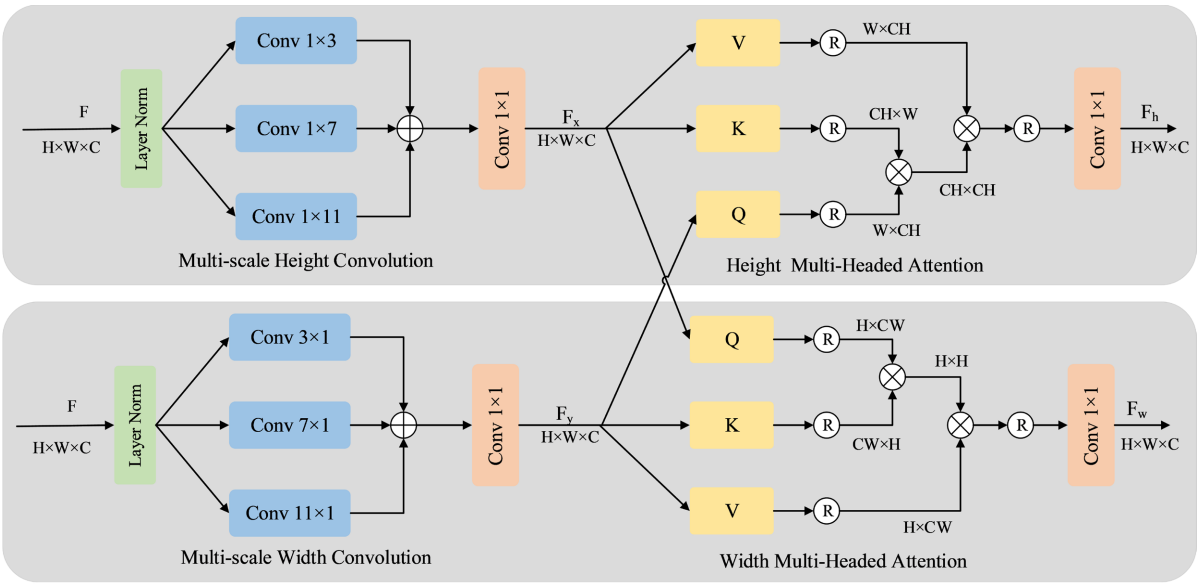

Figure 4. Expansion of the internal structure of the CA-Transformer
图 4. CA-Transformer 内部结构展开图

交叉轴 Transformer 的内部结构如图 4。由于高度轴向多头注意力模块和宽度轴向多头注意力模块很相似，这里选取图 4 上部的分支来简述一下交叉轴 Transformer 的工作原理。具体来说，给定输入特征 $F \in R^{H \times W \times C}$，首先使用 Layer Norm 来增强输入特征，然后经过多尺度高度卷积提取其多尺度上下文信息并融合得到 F_x，经过 1 × 1 卷积调整特征图维度后送入高度轴向多头注意力模块计算自注意力，最后经过尺度调整和 1 × 1 卷积后得到输出 F_h。其大致的计算流程如下：

$$F_x = Conv_{1\times1}\left(\sum_{i=1}^{3} Conv_{1\times i}\left(Norm\left(F\right)\right)\right) \tag{1}$$

$$F_y = Conv_{1\times1}\left(\sum_{i=1}^{3} Conv_{i\times1}\left(Norm\left(F\right)\right)\right) \tag{2}$$

其中 $Conv_{1\times1}$ 表示 1 × 1 卷积，$Norm$ 是层归一化，$Conv_{1\times i}$ 表示沿水平轴的一维卷积，$Conv_{i\times1}$ 表示沿垂直轴的一维卷积。

为了更好地利用两个空间方向上的多尺度卷积特征，本章需要计算 F_x 和 F_y 之间的交叉注意力。具体来说，先使用 3 × 3 卷积对 F_x 计算键矩阵 K_x 和值矩阵 V_x，对 F_y 计算查询矩阵 Q_x，y 轴方向上的交叉注意力同理可得。相关操作可描述为：

$$\begin{cases} Q_x = W_{3\times3}^{Q} F_y \\ K_x = W_{3\times3}^{K} F_x \\ V_x = W_{3\times3}^{V} F_x \end{cases} \tag{3}$$

$$\begin{cases} Q_y = W_{3\times3}^{Q} F_x \\ K_y = W_{3\times3}^{K} F_y \\ V_y = W_{3\times3}^{V} F_y \end{cases} \tag{4}$$

其中 $W_{3\times3}$ 表示 3 × 3 卷积。然后对查询向量 Q 和关键向量 K 进行重构并作点积处理，生成高度轴向多头注意力图 $A \in R^{CH\times CH}$。为了实现多头自注意，这里沿着特征通道维度分别将重构后的 $\hat{Q},\hat{k},\hat{v}$ 划分为 k 个部分：

$$\begin{cases} \hat{Q} = \left[\hat{q}_1,\cdots,\hat{q}_k\right], \\ \hat{K} = \left[\hat{k}_1,\cdots,\hat{k}_k\right], \\ \hat{V} = \left[\hat{v}_1,\cdots,\hat{v}_k\right], \end{cases} \tag{5}$$

其中每个头的维数 $d_k = C/k$。第 j 个头的高度轴多头自注意可表示为：

$$SA\left(\hat{q}_j,\hat{k}_j,\hat{v}_j\right) = \hat{v}_j softmax\left(\hat{q}_j\hat{k}_j/\partial\right) \tag{6}$$

$$F_{out} = Conv_{1\times1}\left[Concat_{j=0}^{k}\left(SA\left(\hat{q}_j,\hat{k}_j,\hat{v}_j\right)\right)\right] \tag{7}$$

其中，$\hat{q}_j$，$\hat{k}_j$，$\hat{v}_j$ 分别表示第 j 个头的 Q，K，V 的值，∂ 是一个比例因子，Concat 表示拼接操作。最终的输出特征 F_h 和 F_w 即公式 7 中的 $F_{out} \in R^{H\times W\times C}$。

3.1.3. 全局注意力融合模块

为了更好地将 CNN 捕获的细节特征与 Transformer 结构编码的全局上下文信息进行深度交互融合，同时优化 U 型网络架构中编码器与解码器层级间信息的传递效率，本研究设计了一种先进的全局注意力融合模块(Global Attention Fusion Block, GAF Block)，其详细结构如图 5 所示。

该模块采用全局平均池化策略，从两种不同类型的特征图中提炼出全局描述符，从而捕获图像的整体语义信息。随后，引入线性变换层对提取的全局特征进行降维处理，这一步骤不仅提升了模型的表达能力，还为特征的进一步融合创造了条件。两个经过线性变换的特征图通过像素求和操作合并为单一的新特征图，该新特征图经过 Sigmoid 激活函数生成一个权重矩阵，其值域限定在 0 到 1 之间。此权重矩阵用于对原始特征图进行加权，赋予模型对图像中关键细节区域以更高的关注度，而对背景或次要区域则分配较低的权重。

最终，通过对加权的特征图进行求和，并应用 ReLU 激活函数以实现非线性变换，确保了特征信息

的最大化保留和有效激活。该融合模块不仅极大地丰富了特征的内涵，有选择性地保留了重点关注区域的特征信息，而且为模型提供了更为精细的特征表示，为后续图像的恢复和增强打下了坚实的基础。该模型的公式表示大致如下：

$$\begin{cases} \hat{X} = MLP\left(GAP\left(X\right)\right) \\ \hat{Y} = MLP\left(GAP\left(Y\right)\right) \end{cases} \tag{8}$$

其中，GAP 为全局平均池化，MLP 表示多层感知器。进而，有：

$$\begin{cases} \tilde{X} = Sigmoid\left(\hat{X} \oplus \hat{Y}\right) \otimes X \\ \tilde{Y} = Sigmoid\left(\hat{X} \oplus \hat{Y}\right) \otimes Y \end{cases} \tag{9}$$

其中 $\oplus$ 表示逐像素加法，$\otimes$ 表示逐像素乘法。最后，模块输出为：

$$GAF\left(X,Y\right) = ReLu\left(\tilde{X} \oplus \tilde{Y}\right) \tag{10}$$

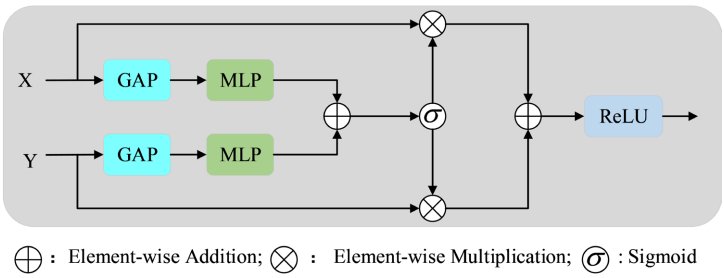

Figure 5. The structure of the global attention fusion module (GAF Block)
图 5. 全局注意力融合模块结构图

3.2. 损失函数

为了全面提升图像质量，本研究提出了一种创新的混合损失函数，该函数在定性和定量评估图像质量方面均表现出色。新提出的损失函数综合考量了图像的结构特性、感知质量、色彩保真度以及不同区域间的差异性，旨在实现对图像质量的全面优化。

通过这种多维度的损失函数设计，模型在训练过程中能够更加精准地调整和优化特征，从而在图像增强等任务中实现更高的图像质量。低照度图像 I_l 与其对应的正常光照图像 I_h 构成训练集 D_{train}。对于 $\forall\left(I_l,I_h\right) \in D_{train}$，本算法在输出端产生预测的增强图像 $\hat{I}$，该训练过程中的预测损失具体表示如下：

$$L = L_1 + L_{ssim} \times L_{psnr} + \lambda_p L_{perc} + \lambda_c L_{color} + \left(1 - \lambda_p - \lambda_c\right) L_{smo} \tag{11}$$

其中 L_{ssim} 表示结构相似度损失，L_{psnr} 表示峰值信噪比损失，L_{perc} 表示感知损失，L_{color} 表示颜色损失，L_{smo} 表示平滑损失。为了平衡网络训练时对结构感知项，感知颜色项和平滑项的关注程度，定义 λ_p 和 λ_c 作为两个取值为正的平衡系数，在本实验中分别取 $\lambda_p = 0.55, \lambda_c = 0.05$。

1) L_1 损失

L_1 损失函数[12]能较好的计算目标图像与预测图像之间的像素差值，对图像对比度以及纹理信息具有很好的约束能力，其具体公式为：

$$L_1 = \sqrt{\left\| I_h - \hat{I} \right\|^2 + \ell} \tag{12}$$

其中，ℓ 为正值常量，这里 $\ell=10^{-6}$。

2) 峰值信噪比损失 L_{psnr}

峰值信噪比[13]取值越大，则表示预测图像越接近于目标图像，因此取其倒数作为损失函数来指导网络模型的学习，从峰值信噪比角度减少预测图像与真值图像之间的差异性，提升模型性能：

$$MSE\left(\hat{I},I_h\right)=\frac{1}{m\times n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[\hat{I}\left(i,j\right)-I_h\left(i,j\right)\right]^2 \tag{13}$$

$$L_{PSNR}\left(\hat{I},I_h\right)=\left(10\times\log_{10}\left(\frac{MAX_{I_h}}{MSE\left(\hat{I},I_h\right)}\right)\right)^{-1}=\left(20\times\log_{10}\left(\frac{MAX_{I_h}}{\sqrt{MSE\left(\hat{I},I_h\right)}}\right)\right)^{-1} \tag{14}$$

其中，MAX_{I_h} 表示正常光图像中像素的最大值，对于 8 位的灰度图像通常取 $MAX_{I_h}=255$，$MSE\left(\hat{I},I_h\right)$ 为正常光图像和预测图像的均方差。

3) 结构相似度损失 L_{ssim}

结构相似度[14]代表两张图片的相似程度，二者呈正相关，其取值在 0 到 1 之间。该指标通过两张图像的像素统计特征分别计算出亮度，对比度和结构性信息，并使用这三个特征信息的组合来定义两幅图片的相似程度。其具体定义如下：

$$SSIM\left(x,y\right)=\frac{\left(2\mu_x\mu_y+c_1\right)\left(2\sigma_{xy}+c_2\right)}{\left(\mu_x^2+\mu_y^2+c_1\right)\left(\sigma_x^2+\sigma_y^2+c_2\right)} \tag{15}$$

$$L_{ssim}=1-SSIM \tag{16}$$

其中，μ_x 是 x 的均值，μ_y 是 y 的均值，σ_x 是 x 的方差，σ_y 是 y 的方差，σ_{xy} 是 x 和 y 的协方差，为避免分母为 0，引入 c_1、c_2 两个正常数，在实验中统一设置 $c_1=0.0001$、$c_2=0.0009$。x 和 y 代表输入的两张图像。

4) 感知损失 L_{perc}

为了确保正常光图像和预测图像的语义感知信息[15]尽可能一致，有必要在良好的感知空间中引入对图像对 $\left(\hat{I},I_h\right)$ 内容的感知差异进行约束。该损失可表示为：

$$L_{perc}\left(\hat{I},I_h\right)=\frac{1}{C_jH_jW_j}\left\|\varphi\left(I_h\right)-\varphi\left(\hat{I}\right)\right\|_1 \tag{17}$$

其中，$H_j\times W_j\times C_j$ 为特征图尺寸和通道数，$\varphi()$ 表示从 VGG16 网络中获取的第 j 层特征信息。

5) 颜色损失 L_{colour}

感知颜色损失[16]是通过计算正常光图像与预测图像在欧氏空间中的颜色色度值之间的差异进而来约束网络模型对图像颜色信息的恢复：

$$L_{colour}=\Delta E\left(I_h,\hat{I}\right) \tag{18}$$

其中，ΔE 表示色差计算。

6) 平滑性损失 L_{smo}

平滑性损失[17]定义为具有空间变形的 L_1 范数，用来惩罚预测图像相应区域的颜色差异。该损失是以真值图像 I_h 中每个像素邻域的颜色分布情况为指导，针对预测图像 $\hat{I}$ 对应的每个像素邻域颜色分布不一致得出的损失，其具体公式为：

$$L_{smo}\left(\hat{I},I_h\right)=\sum_{i=1}^{N}\sum_{j\in N_h(i)}\omega_{i,j}\left(I_h\right)\left\|\hat{I}_i-\hat{I}_j\right\|_1 \tag{19}$$

其中，N 表示真值图像和预测图像的像素总数，$N_h(i)$ 代表像素 i 的 $h\times h$ 邻域，$\omega_{i,j}\left(I_h\right)$ 是基于真值图像估计得到的成对像素 i 与 j 的权重系数，具体公式如下：

$$\omega_{i,j}\left(I_h\right)=\exp\left[-\frac{\sum_c\left(I_h^{i,c}-I_h^{j,c}\right)^2}{2\sigma^2}\right] \tag{20}$$

其中，c 表示 YUV 颜色空间中的通道序号，σ 为高斯核的标准差。本实验中统一设置 $h=5$，$c=0.1$。

4. 实验与分析

实验部分，本研究分别通过多个对比实验和消融实验来验证所提出的基于 CA-Transformer 的 U 型低照度图像增强网络的有效性。同时在 LOL 数据集和 LSRW 数据集上对分别不同网络低照度图像增强效果进行可视化展示，直观展示此方法的可行性和增强效果。

4.1. 实验数据集

1) LOL 数据集

Wei 等人[5]通过改变相机曝光时长和 ISO 的方式，从多个真实场景中采集到 500 对低照度－正常照度的图像数据集。其中图像均为 400 × 600 大小的便携式网络图形格式，提高了数据处理的效率。该数据集由 485 组图像对组成的训练集和 15 组图片对组成的测试集构成，每组图像对都包含一个含有噪声的低光照图像和其相对应的良好曝光下的参考图像组成。值得注意的是，该数据集中的图像大多数为室内场景。这为研究提供了一个专注于特定类型场景的数据集，有助于模型更好地理解和学习室内光照条件下的图像增强任务。

2) LSRW 数据集

Jiang Hai 等人[18]在 2021 年收集了第一个大规模的真实世界配对的图像数据集。数据集由 5650 张配对图片组成，其中包含了 5600 对图像构成的训练集和 50 张图像构成的验证集。数据集中的图像是通过尼康 D7500 相机和华为 P40 Pro 手机拍摄得到的，这增加了数据集在不同设备和传感器上拍摄图像的多样性，使得研究成果更具普适性。其拍摄场景既包括室内环境，也包括室外环境，可以提高模型在各种实际应用场景中的适应性。图片的大小为 960 × 720，这一分辨率既能够保证图像的细节丰富，又能够避免数据量过大导致的计算资源消耗。

4.2. 实验环境与参数设置

本实验采用基于 Python 3.8 的 PyTorch 深度神经网络框架，编译器为 Pycharm，实验环境为：NVIDIA RTX 3090Ti GPU，24GB 内存。CUDA 是 11.1 版本，运行于 Ubuntu 16.04 平台。为保持一致性，所有输入图像都被调整到了 256 × 256 像素的尺寸。在网络训练中，把预处理好的低照度图片和真实的标签送入 RT-UNet 网络训练。采用随机梯度下降(SGD)算法对网络进行迭代优化，batch size 设置为 10，patch size 设置为 16，初始学习率 lr 设置为 0.0005，权重衰减 weight decay 为 10^{-4}，共训练 200 个周期，动量参数 momentum 设置为 0.9。

4.3. 对比实验结果及分析

本研究选取了 RetinexNet，KinD，MIRNet [19]，UFormer [20]，Enlightengan [21]，R2RNet [18]，

Restormer [22]，LLFormer [23]以及 HWMNet [24]作为实验对比模型。其中，除了几个经典的和近三年的 CNN 架构的低照度图像增强网络，这里还选取了 UFormer，Restormer 等三个基于 Transformer 的低照度图像增强算法来针对性的衡量本算法的性能指标。由表 1 可以发现在 LOL 数据集上，RT-UNet 在结构相似度和峰值信噪比两个客观评价指标上都取得了不错的效果。鉴于 LOL 数据集的图像数量有限，而基于 Transformer 的模型通常需要大量的数据进行有效训练，因此在峰值信噪比这一性能指标上，本模型相较于 HWMNet 表现略有不足，但结构相似程度上基本持平。进一步来看，对比同为 Transformer 结构的 Restormer 和 LLFormer，本方法在两项实验指标上都实现了反超，分别比 LLFormer 高出了 0.13，0.024 dB。

Table 1. Comparison results of different network structures on the LOL dataset
表 1. 不同网络结构在 LOL 数据集上的对比结果

网络结构	PSNR↑	SSIM↑
RetinexNet (2018)	16.77	0.425
KinD (2019)	19.65	0.771
MIRNet (2020)	23.16	0.816
EnlightenGAN (2020)	17.48	0.578
UFormer (2021)	18.85	0.749
R2RNet (2021)	18.71	0.723
HWMNet (2022)	**24.24**	0.849
Restormer (2022)	22.56	0.827
LLFormer (2023)	23.65	0.833
RT-UNet (ours)	23.78	**0.857**

Table 2. Comparison results of different network structures on the LSRW dataset
表 2. 不同网络结构在 LSRW 数据集上的对比结果

网络结构	PSNR↑	SSIM↑
RetinexNet (2018)	15.49	0.347
KinD (2019)	16.47	0.492
MIRNet (2020)	18.30	0.612
UFormer (2020)	18.62	0.573
EnlightenGAN (2020)	16.31	0.470
R2RNet (2021)	18.08	0.550
HWMNet (2022)	19.17	0.612
Restormer (2022)	19.33	0.614
LLFormer (2023)	19.59	0.628
RT-UNet (ours)	**19.83**	**0.635**

相较于 LOL 数据集，LSRW 数据集收集了大量真实复杂场景下的低光照图像，图像数量庞大，对模型的泛化性能提出了很高的要求。同时该数据集涉及了大量复杂的室外场景和一些微光杂光下拍摄的室内场景，这些场景中的杂光，噪声等使拍摄主体更多的处于非均匀光照环境下，给低照度图像增强任务带来了更多挑战，对比表 1 和表 2 的数据也可以证明上述结论。在 LSRW 数据集上各个网络模型无论是结构相似度还是峰值信噪比等指标都大大逊色于在 LOL 数据集上得到的结果。

得益于 LSRW 数据集庞大的图片数量，本模型在经过充分的训练和学习之后，在两项性能指标上实现了对 HWMNet 的显著超越。由表 2 的实验数据可以发现，在 LSRW 数据集上，RT-UNet 无论是结构相似度还是峰值信噪比指标都在在第二名的基础上取得了一定程度的提升。

4.4. 消融实验

为了证明本算法中提出的各个模块的有效性，本研究在 LOL 数据集对模块的有无分别进行了消融实验。同时针对 RT-UNet 中各个位置交叉轴 Transformer 模块的堆叠数量和混合损失函数中各惩罚项的权重分布进行了相应的实验论证。

RT-UNet 在输入输出位置通过堆叠 CA-Transformer 模块使图像特征信息被更好地固化保留，同时对噪声起一定的抑制作用，相反输入图像直接送入 U 型主网络和特征重构后直接输出会损失大量特征信息。表 3 的实验也证明了在输入输出位置堆叠 CA-Transformer 模块对网络整体增强性能有很大贡献。同样地，表 4 的实验结果验证了在传统 U 型网络编码器解码器部分分别引入 CA-Transformer 模块的效果，其不仅可以增加网络深度，对下采样模块的特征提取和上采样位置的特征重构能力都有很大的提升作用。

Table 3. Experimental comparison results on the number of stacked CA Transformer modules at input and output positions on the LOL dataset
表 3. 在 LOL 数据集上对输入输出位置 CA-Transformer 模块堆叠数量的实验对比结果

CA-Transformer Block 的堆叠数量	PSNR	SSIM
1	19.09	0.768
2	19.27	0.783
3	**19.56**	**0.802**
4	19.52	0.798

Table 4. Comparison of experimental results on the stacking number of CA Transformer modules in encoder decoder on the LOL dataset
表 4. 在 LOL 数据集上对编码器解码器中的 CA-Transformer 模块堆叠数量的实验对比结果

每一层 CA-Transformer Block 的堆叠数量	PSNR	SSIM
[1, 1, 2, 1]	20.65	0.802
[2, 2, 4, 2]	21.18	**0.819**
[2, 2, 2, 2]	20.85	0.807
[4, 4, 8, 4]	**21.21**	0.815

通过对比不同堆叠数量CA-Transformer模块的RT-UNet在LOL数据集上的客观评价指标可以发现，最优的结构搭配是输入输出位置分别串接三个交叉轴Transformer模块，而在U型主网络内部，每一层的交叉轴Transformer模块堆叠数量分别为2，2，4，2。为了证明CA-Transformer模块和全局注意力融合模块对网络整体的贡献与否，分别进行了三组消融实验。通过表5的数据可以看出，本研究所提出的两个模块对最终网络的增强效果都有一定程度的提升。

Table 5. Evaluation of each module on the LOL dataset
表5. 在LOL数据集上各模块的评估

模型	PSNR	SSIM
w/o GAF and CA-Transformer	18.87	0.759
w/o GAF	21.67	0.824
w/o CA-Transformer	20.43	0.809
RT-UNet (ours)	**23.78**	**0.857**

最后，针对大小不同的混合损失函数各惩罚项系数对模型性能的影响展开消融实验，如表6所示，对其中数据进行分析可得，混合损失函数中的系数$\lambda_p=0.55$，$\lambda_c=0.05$时网络模型的整体效果最佳。

Table 6. Objective evaluation indicators under different loss function coefficients on the LOL dataset
表6. 在LOL数据集上不同损失函数系数下的客观评价指标

损失函数系数	PSNR	SSIM
$\lambda_p=0.55$，$\lambda_c=0.05$	**23.78**	**0.857**
$\lambda_p=0.65$，$\lambda_c=0.08$	23.55	0.841
$\lambda_p=0.45$，$\lambda_c=0.02$	23.43	0.829

4.5. 主观效果展示及分析

为了更直观地展现RT-UNet针对低光照图像的增强效果，本研究选取了RetinexNet，R2RNet，Uformer，Restormer和LLFormer等几个经典模型和近几年提出的基于Transformer的低照度图像增强模型作为对比，分别针对LOL和LSRW数据集上挑选出的两个场景的图像进行增强效果展示。

(a) 低光 (b) RetinexNet (c) R2RNet (d) UFormer

(e) Restormer (f) LLFormer (g) RT-UNet (h) 真值

Figure 6. Visualization results of different network structures on the LOL dataset
图 6. LOL 数据集上不同网络结构的可视化结果

由图 6 可以发现，对于 LOL 数据集上这两张光照相对均匀的图像，RT-UNet 和其他模型相比，无论是整体风格，色度的保持还是亮度的提升上都有一定的优势，呈现出了比较接近真值的视觉效果。而观察图 7 的两个室外场景对比图像可以看出，RT-UNet 增强后的图像不仅在视觉上更接近真实场景，而且在细节恢复和色彩保持方面也展现出了较高的质量，证明了其在多场景和多光照条件下的强大泛化能力和实用性。

(a) 低光 (b) RetinexNet (c) R2RNet (d) UFormer

(e) Restormer (f) LLFormer (g) RT-UNet (h) 真值

Figure 7. Visualization results of different network structures on the LSRW dataset
图 7. LSRW 数据集上不同网络结构的可视化结果

5. 总结

针对低照度图像增强领域中由光照分布不均，环境光源复杂导致的图像曝光不充分不均匀问题，本研究提出了一种融合 CA-Transformer 模块和全局注意力融合模块的 U 型网络 RT-UNet，它通过精心设计的网络架构显著提升了低照度图像的增强效果。RT-UNet 的核心优势在于保持传统 U 型网络对局部信息挖掘能力的同时还能够利用 Transformer 的自注意力机制来捕捉图像中的全局信息，这在处理光照不均匀的图像时尤为重要。在 LOL 和 LSRW 等数据集上的实验验证表明，RT-UNet 在客观评价指标和主观效果上都展现出了显著的优势，证明了其在复杂光照环境下图像增强任务中的有效性。

总体而言，RT-UNet 模型的提出为低照度图像增强领域提供了一种新的解决方案，特别是在非均匀光照环境下，其展现出的全局感知能力、计算效率等方面的优势，预示着 Transformer 结构在低照度图像增强领域的广阔应用前景。尽管 RT-UNet 在特定数据集上表现出色，但仍存在一些不足之处。未来的研究可以着重于提升模型的泛化性能，通过在更多样化的数据集上进行训练和测试，以适应更多的实际应用场景。同时，进一步优化计算效率，通过模型压缩和加速技术减少模型的计算资源需求，以满足实时图像增强的应用需求。

基金项目

国家自然科学基金项目(62373251)资助。

参考文献

[1] 刘鑫. 基于深度学习的低照度图像增强方法研究[D]: [硕士学位论文]. 天津: 天津大学, 2018.

[2] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *NIPS*'17: *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Red Hook, 4-9 December 2017, 6000-6010.

[3] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention-MICCAI* 2015: 18*th International Conference*, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[4] Jiang, L., Jing, Y., Hu, S., *et al*. (2018) Deep Refinement Network for Natural Low-Light Image Enhancement in Symmetric Pathways. *Symmetry*, **10**, Article No. 491. https://doi.org/10.3390/sym10100491

[5] Wei, C., Wang, W., Yang, W., *et al*. (2018) Deep Retinex Decomposition for Low-Light Enhancement.

[6] Guo, C., Li, C., Guo, J., *et al*. (2020) Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 1780-1789. https://doi.org/10.1109/CVPR42600.2020.00185

[7] Shi, P., Xu, X., Fan, X., *et al*. (2024) LL-UNet++: UNet++ Based Nested Skip Connections Network for Low-Light Image Enhancement. *IEEE Transactions on Computational Imaging*, **10**, 510-521. https://doi.org/10.1109/TCI.2024.3378091

[8] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., *et al*. (2018) Unet++: A Nested U-Net Architecture for Medical Image Segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*: 4*th International Workshop*, *DLMIA* 2018, *and* 8*th International Workshop*, *ML-CDS* 2018, *Held in Conjunction with MICCAI* 2018, Granada, 20 September 2018, 3-11. https://doi.org/10.1007/978-3-030-00889-5_1

[9] Zhang, Z., Jiang, Y., Jiang, J., *et al*. (2021) Star: A Structure-Aware Lightweight Transformer for Real-Time Image Enhancement. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11-17 October 2021, 4106-4115. https://doi.org/10.1109/ICCV48922.2021.00407

[10] Souibgui, M.A., Biswas, S., Jemni, S.K., *et al*. (2022) Docentr: An End-to-End Document Image Enhancement Transformer. 2022 26*th IEEE International Conference on Pattern Recognition* (*ICPR*), Montréal, 21-25 August 2022, 1699-1705. https://doi.org/10.1109/ICPR56361.2022.9956101

[11] Jiang, Y., Chang, S. and Wang, Z. (2021) TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale up. 35*th Conference on Neural Information Processing Systems* (*NeurIPS* 2021), 6-14 December 2021, 14745-14758.

[12] Lai, W.S., Huang, J.B., Ahuja, N., *et al*. (2018) Fast and Accurate Image Super-Resolution with Deep Laplacian Pyra-

mid Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 2599-2613. https://doi.org/10.1109/TPAMI.2018.2865304

[13] Brauers, J. and Aach, T. (2006) A Color Filter Array Based Multispectral Camera. 12. *Workshop Farbbildverarbeitung*, Ilmenau, 5-6 October 2006, 1-11.

[14] Wang, Z., Bovik, A.C., Sheikh, H.R., *et al*. (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, **13**, 600-612. https://doi.org/10.1109/TIP.2003.819861

[15] Johnson, J., Alahi, A. and Fei-Fei, L. (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *Computer Vision-ECCV* 2016: 14*th European Conference*, Amsterdam, 11-14 October 2016, 694-711. https://doi.org/10.1007/978-3-319-46475-6_43

[16] Wang, R., Zhang, Q., Fu, C.W., *et al*. (2019) Underexposed Photo Enhancement Using Deep Illumination Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 6849-6857. https://doi.org/10.1109/CVPR.2019.00701

[17] He, K., Zhang, X., Ren, S., *et al*. (2015) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

[18] Hai, J., Xuan, Z., Yang, R., *et al*. (2023) R2rnet: Low-Light Image Enhancement via Real-Low to Real-Normal Network. *Journal of Visual Communication and Image Representation*, **90**, Article ID: 103712. https://doi.org/10.1016/j.jvcir.2022.103712

[19] Zamir, S.W., Arora, A., Khan, S., *et al*. (2020) Learning Enriched Features for Real Image Restoration and Enhancement. *Computer Vision-ECCV* 2020: 16*th European Conference*, Glasgow, 23-28 August 2020, 492-511. https://doi.org/10.1007/978-3-030-58595-2_30

[20] Wang, Z., Cun, X., Bao, J., *et al*. (2022) Uformer: A General U-Shaped Transformer for Image Restoration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 17683-17693. https://doi.org/10.1109/CVPR52688.2022.01716

[21] Jiang, Y., Gong, X., Liu, D., *et al*. (2021) Enlightengan: Deep Light Enhancement without Paired Supervision. *IEEE Transactions on Image Processing*, **30**, 2340-2349. https://doi.org/10.1109/TIP.2021.3051462

[22] Zamir, S.W., Arora, A., Khan, S., *et al*. (2022) Restormer: Efficient Transformer for High-Resolution Image Restoration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 5728-5739. https://doi.org/10.1109/CVPR52688.2022.00564

[23] Wang, T., Zhang, K., Shen, T., *et al*. (2023) Ultra-High-Definition Low-Light Image Enhancement: A Benchmark and Transformer-Based Method. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 2654-2662. https://doi.org/10.1609/aaai.v37i3.25364

[24] Fan, C.M., Liu, T.J. and Liu, K.H. (2022) Half Wavelet Attention on M-Net+ for Low-Light Image Enhancement. 2022 *IEEE International Conference on Image Processing* (*ICIP*), Bordeaux, 16-19 October 2022, 3878-3882. https://doi.org/10.1109/ICIP46576.2022.9897503