低光照环境下基于 CNN-Transformer 并行架构的图像增强方法

顾睿[1]

（1. 兰州大学信息科学与工程学院，甘肃兰州 730000）

摘要：本研究提出了一种用于弱光图像增强的 PACUT 算法。它通过融合 CNN 分支和 Transformer 分支的并行结构，有效提取图像的长距离和短距离特征，进而使得增强图像接近真实值。具体而言，CNN 分支采用改进的 U 型网络架构，包括编码器和解码器子网络，而 Transformer 分支以视觉变换器为核心，利用自注意力机制提取全局特征。实验结果表明 PACUT 在 LOL 和 SCIE 等数据集上取得了较好性能，尤其在 PSNR 和 SSIM 方面相对于主流模型达到最佳，并且我们的模型具有一定的鲁棒性，在非训练所属的数据集上也表现出不错的效果。

关键词：低光图像增强；LLIE；Transformer；并行网络结构

1 引言

弱光图像增强是图像处理领域的一个关键任务，旨在提升低光环境下拍摄图像的感知质量。该领域的最新进展主要由深度学习技术推动，涵盖了多样的学习策略、网络架构、损失函数和训练数据集。低光图像增强技术在视觉监控、自动驾驶和计算摄影等多个领域有着广泛应用。特别是在智能手机摄影领域，由于相机光圈大小、实时处理需求和存储限制，低光环境下的图像捕获尤为具有挑战性。

传统的低光增强方法主要包括基于直方图均衡化 [1,2]，基于 Retinex 理论 [3–5] 和基于去雾理论 [6,7] 的技术。直方图均衡化方法能有效提升图像亮度和对比度，但由于其全局性增益特性，容易导致噪声放大和伪影 [8,9] 产生。而基于 Retinex 理论的方法虽能改善照度和对比度，但同样存在忽视噪声、保留或放大噪声的问题 [10]。此外，这类方法中有效先验或正则化的确定具有挑战性，不准确的先验或正则化可能导致增强结果中的过度增强和颜色偏差 [11,12]。由于其复杂的优化过程，这些方法的运行时间也相对较长。基于去雾理论的方法利用低照度图像的逆图像进行去雾操作增强，但存在的问题与上述方法一致，增强操作会导致增强图像中的噪声。

近年来，基于深度学习的低光图像增强（LLIE）技术取得了显著成就，它使用神经网络来学习从弱光图像到自然光图像的映射。与传统方法相比，基于深度学习的解决方案在准确性、鲁棒性和处理速度方面表现更优，因此受到了广泛关注。特别是卷积神经网络（CNN）在多个计算机视觉任务中展现出卓越的性能。CNN 通过利用注意力机制和 [13,14] 上下文信息，能够从原始图像中有效提取多尺度特征 [15,16]。在这些成果的推动下，基于 CNN 的低光图像增强方法得到了持续发展。例如，一种基于 CNN 的自适应低光图像增强框架 [17] 显著提升了图像的对比度、颜色和细节信息。然而，现有的基于 CNN 的方法大多集中于图像亮度、纹理和颜色的恢复，对于局部光照不均匀、颜色信息和细节信息的丢失问题，仍存在过增强或增强不足的挑战。

自 Transformer 架构 [18] 在图像处理领域的应用以来 [19]，其在捕获全局特征方面的能力已经引起了行业的广泛关注。然而，在图像复原任务中，由于 Transformer 基于自注意力机制的计算特性，其面临着较高的计算复杂度问题。这一挑战通常需要通过改进前馈神经网络（FFN）来解决，以便更有效地捕获关键信息，从而增强模型的非线性建模能力 [20]。为了缓解计算负担，部分研究工作尝试采用局部窗口策略来限制自注意力机制的计算范围，并借鉴 U-Net 架构，引入跳过连接机制以增强特征传递 [21]。尽管这种方法在降低计算成本方面取得了一定成效，但它可能会削弱 Transformer 架构在捕获图像中长距离特征的能力，进而影响其在图像复原任务中的整体性能。为了克服这一限制，提出的解决方案 [22] 主要集中在窗口间的交互和与卷积神经网络（CNN）模块的耦合上。这种方法旨在将 CNN 的平移不变性和局部性归纳偏差融入 Transformer 架构中，从而实现两者优势的互补。我们将这类解决方案称为 CNN 与 Transformer 的串联策略。

本研究基于 Conformer 架构[23]，该架构巧妙地结合了卷积神经网络（CNN）在捕获短距离特征[24–26]方面的优势与 Transformer 自注意力模块在捕获长距离特征[27]依赖关系方面的能力。然而，Transformer 的自注意力模块在捕获远距离特征的同时，可能会忽视局部特征的细节。为了克服这一局限性，Conformer 结构融合了卷积运算和自注意力机制，以增强表示学习的能力。在此基础上，我们提出了一种新颖的并行深度学习架构，命名为结合 U 型网络和 Transformer 模型的并行架构（PACUT），专门针对低光照条件下的图像增强问题。PACUT 模型融合了经典的 U 型网络和 Transformer 深度学习方法，旨在提升弱光图像增强任务中恢复图像的可视性和质量，同时保留关键的细节和纹理信息。我们的 PACUT 模型主要展现了以下创新点：

(1) 将并行架构成功应用于弱光图像增强任务，并在实验中展示了其优越的性能。

(2) 提出了一种基于 U 型网络的 CNN 分支，专门用于处理短距离特征。

(3) 设计了一种创新的特征耦合单元，不仅与原有方法不同，而且能够有效地融合 CNN 分支和 Transformer 分支的特征。该单元实现了 CNN 特征图和 Transformer 的嵌入块（Patch embeddings）之间的无缝转换，从而实现了短距离和长距离特征的有效耦合。

2 相关工作

2.1 基于 CNN 的弱光图像增强方法

目前，基于深度学习的弱光图像增强方法主要注重通过优化输出与地面真实值之间的外观重建误差来提升图像质量[28,29]。与传统方法相比，基于深度学习的方法，如 CNN，能够更全面地捕获图像特征，因此在弱光条件下更适用于图像增强。先前的研究，如 Lore 等人的工作[30,31]，采用稀疏去噪编码器来增强闪烁图像，展示了深度学习在构建闪烁增强模型方面的潜力。Wang 等人[12]提出了一种多级弱光图像增强方法，通过在潜空间分解输入图像为两个低耦合特征分量。Guo 等人[29]则提出了 ZeroDCE 算法，通过学习输入图像的曲线函数实现弱光增强。近期研究者尝试多种方法来实现低光图像增强，致力于提高在复杂黑暗背景下图像细节的可见性。Jiang 等人[32]采用非参考策略设计了一种启发式 GAN 用于弱光增强。Zhang 等人[33]提出了自监督的弱光图像增强方法，以提高图像对比度并降低噪声，避免生成图像的模糊。Fu 等人[34]设计了一种自适应的弱光原始图像增强网络。Zhang 等人[35]引入了有效的洗牌注意 (Shuffle Attention, SA) 模块，通过通道分组和洗牌单元描述特征的空间和通道依赖关系。另外，Zhang 等人[36]提出了一种新的图像重建方法，利用周围环境信息对过曝光或过饱和的纹理信息进行图像细节重建，并应用边缘感知图像分解进行图像增强。

2.2 基于 Transformer 的弱光图像增强方法

Transformer[18]最初被引入于自然语言处理领域，随后，视觉变换器[19]在计算机视觉序列中被提出，并首次用于图像分类问题。视觉变换器巧妙地结合了计算机视觉和自然语言处理领域的知识，其处理方式包括将图像进行分割、平面化成序列，并将其输入到原始 Transformer 模型的 Encoder 部分，最终通过访问完全连接的层对图像进行分类。SETR（Spatially Enhanced Transformer）[37]模型将全局上下文模块嵌入到变压器的每一层，该变压器的编码器可与简单的解码器相结合，从而形成强大的分割模型。此外，IPT 模型通过在 ImageNet 上进行训练，实现了一个能够处理降噪、超分割和去雨化的图像预训练模型。借鉴 Swin Transformer[38]的思想，SwinIR 模型针对图像恢复提出了一种包括浅层特征提取、深层特征提取和图像重建等三个部分的解决方案。Uformer[39]利用 Transformer 模块构建了一种分层编解码网络，其引入了局部增强窗口和跳过连接机制，使得其具备图像降噪和弱光图像增强能力。Transformer 在图像修复中应用的难点在于计算复杂度高，因此，一些改进方法[20]主要通过调整多头自注意力高度和宽度轴的大小来缩小复杂度。

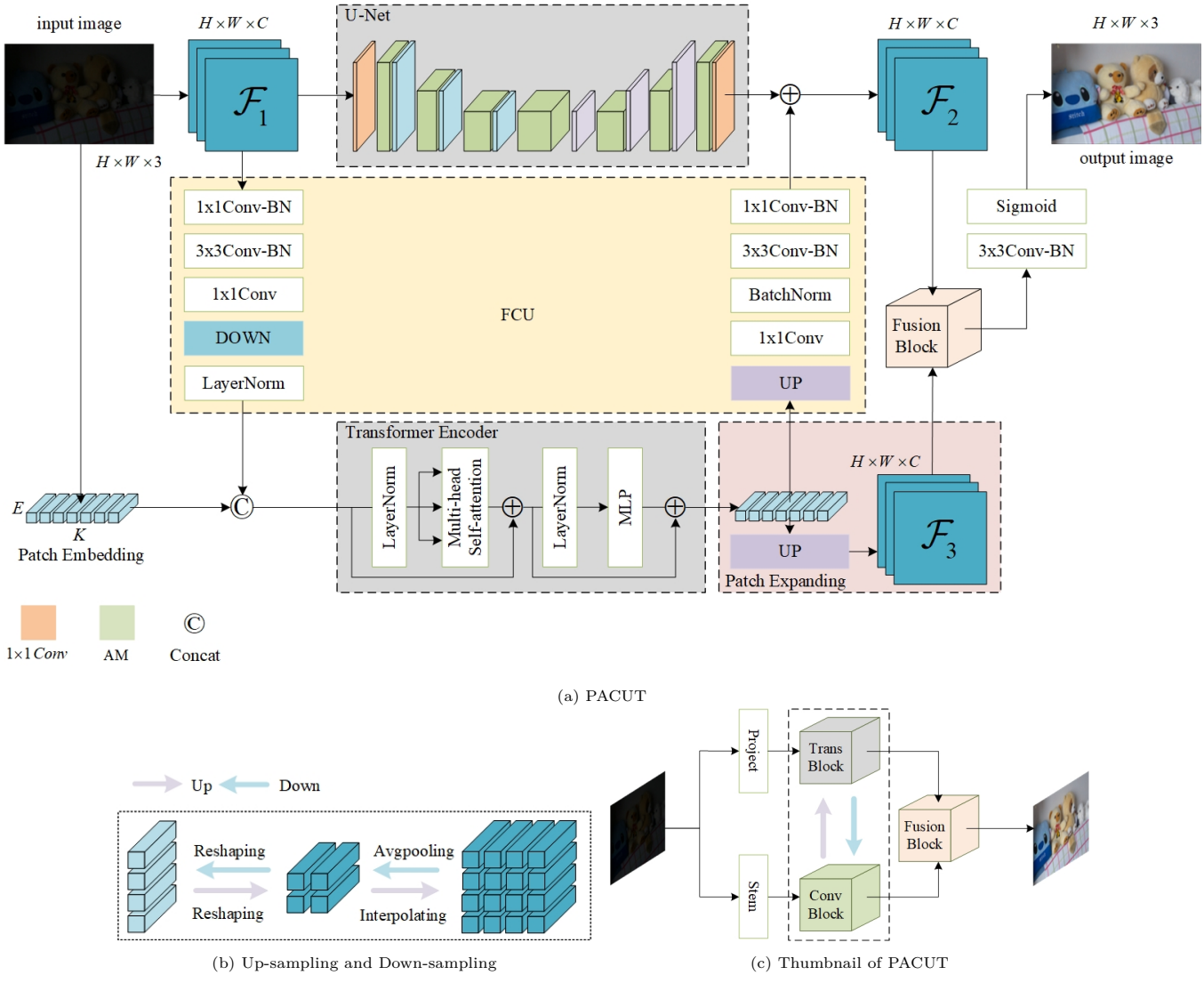
3 方法

3.1 PACUT 结构

在计算机视觉领域，图像特征的研究通常聚焦于两个主要方面：局部特征和全局特征。局部特征，也称为短距离特征，指的是对图像中微小区域的紧凑向量描述，这些特征在众多计算机视觉算法中扮演着基础角色[24–26]。相对而言，全局特征，或长距离特征，涵盖了更广泛的范围，包括但不限于长距离轮廓[27]、形状描述符以及不同对象类别的识别。

在深度学习的框架下，卷积神经网络（CNN）通过分层的卷积操作，逐步提取并加工图像的局部特征。这些特征随后被保留在特征图中，以供后续处理和分析。另一方面，视觉变换器（Vision Transformers）则采用一系列自注意力模块，这些模块能够以更加灵活的方式聚合这些特征，从而形成对整体图像的全面和全局性理解。这种方法在处理图像的全局特征时显示出了显著的优势，尤其是在捕获长距离依赖关系和整体图像结构方面。

为了最大限度地利用局部特征和全局表征的互补优势，本研究设计了一种创新的并行网络架构，命名为 PACUT（Parallel Architecture Combining U-Net and Transformer），如图1c所示。PACUT 的设计理念基于对 CNN 和 Transformer 模型特性的深入理解和互补性分析。在此架构中，Transformer 分支的全局上下文信息被连续地整合到特征图中，以此增强 CNN 分支对全局信息的感知能力。同时，CNN 分支中提取的局部特征被逐步融入到 Patch Embedding 中，以丰富 Transformer 分支的局部细节表现。



(a) PACUT

(b) Up-sampling and Down-sampling

(c) Thumbnail of PACUT

图 1. PACUT 模型结构。

在这种并行结构中，CNN 分支和 Transformer 分支分别专注于最大程度地保留局部特征和全局表征。FCU 作为一种桥接模块，其作用是融合 CNN 分支中的局部特征和 Transformer 分支中的全局表征，如图1a 所示。FCU 从第二步开始应用，因为两个分支的初始特征虽相同，但表征形式不同。在整个网络中，FCU 通过双向交互融合特征图和 Patch Embeddings。

3.2 CNN 分支

在 PACUT 架构中，CNN 分支基于 U-Net 网络结构，这是一种原本为图像分割任务设计的网络架构，但其在图像增强领域同样展现出卓越的性能。U-Net 的显著特点在于其对称的编码器-解码器结构和编码器与解码器之间的跳跃连接。在本研究中，CNN 分支的目标是从输入的低光照图像（LLI）中提取丰富的特征，包括边缘、纹理、颜色和亮度信息，并最终生成增强的图像（EI）。

3.2.1 注意力残差多尺度块

注意力残差多尺度块 (ARMB, Attention Residual Merging Block) 是一种用于深度学习模型中的结构单元，特别是在处理图像相关任务时，它的主要作用是融合来自不同来源的特征，并通过注意力机制增强模型对关键信息的关注。其核心作用之一是融合来自模型不同层或不同分支的特征。这种融合有助于结合低级特征（如边缘和纹理）和高级特征（如对象部分和语义信息），从而提升模型的表现力。ARMB 中的注意力机制有助于 UNetEnhance 更加关注于图像中的重要区域或特征。通过加权重要特征，ARMB 可以提高模型对关键信息的敏感性。ARMB 通过残差连接来解决深度神经网络中的梯度消失问题。如图**??**c 所示，残差连接允许信息直接从早期层传递到后续层，从而保持信息流的连续性。

模块对输入特征 $\mathcal{F}$ 进行两次分支下采样，分别获取尺寸为 $\frac{1}{2}$ 和 $\frac{1}{4}$ 的特征。每个分支使用空间注意力 (Spatial Attention) 过滤噪声，通过 3×3 卷积层提取特征，并通过上采样恢复到原始尺寸。将两个特征图拼接后，通过 1×1 卷积层处理，再应用通道注意力 (Channel Attention) 进行特征加权，最后通过跳跃连接与原始输 $\mathcal{F}$ 合并，得到最终特征图。

假设 $\mathcal{F}_{in}$ 是输入特征图，$\mathcal{F}_{out}$ 是输出特征图，$\mathcal{F}'_i$ 表示对特征图进行裁剪或放大，见下式：

$$
\begin{aligned}
\mathcal{F}'_{\frac{1}{2}} &= \text{Resize}\left(\mathcal{F}_{in}, \left(\frac{H}{2}, \frac{W}{2}\right)\right), \\
\mathcal{F}'_{\frac{1}{4}} &= \text{Resize}\left(\mathcal{F}_{in}, \left(\frac{H}{4}, \frac{W}{4}\right)\right), \\
\mathcal{F}_{out} &= W_1 \mathcal{F}_{out} + W_2 \mathcal{C}\left(f^{1\times1}\left[\mathcal{F}'_2(f^{3\times3}(\mathcal{S}(\mathcal{F}'_{\frac{1}{2}}))), \mathcal{F}'_4(f^{3\times3}(\mathcal{S}(\mathcal{F}'_{\frac{1}{4}})))\right]\right).
\end{aligned}
\tag{1}
$$

其中 $\mathcal{S}$ 和 $\mathcal{C}$ 分别表示空间注意力函数和通道注意力函数，W_1 和 W_2 为对应的特征图权重。

在 ARMB 模块内部，我们引入了 SCConv 模块，这是一种专门设计用于特征冗余的空间和通道重构卷积。SCConv 模块由两个关键子模块组成：空间重构单元（SRU）和通道重构单元（CRU）。SRU 的主要作用是通过门控机制对特征进行选择性激活，从而增强模型对于关键特征的关注度。具体来说，SRU 通过学习特征的空间分布，实现对特征图中重要区域的强调和非重要区域的抑制。这种机制使得模型能够更加有效地处理图像中的空间信息，提高特征提取的精确性。

另一方面，CRU 负责特征的融合和重构。它通过分析和整合不同通道上的特征信息，优化特征表示的通道维度。CRU 的设计基于这样一个假设：不同通道的特征具有不同的重要性和贡献度。因此，CRU 通过学习不同通道间的关系，实现对特征的有效融合，从而提高了特征表示的丰富性和鲁棒性。这一过程不仅增强了模型对于通道信息的利用效率，还提升了整体网络对于复杂图像内容的理解能力。

综上所述，SCConv 模块通过 SRU 和 CRU 的协同作用，实现了对特征的空间和通道维度的有效重构，从

而优化了特征表示的质量和效率。这种设计在处理图像特征时，不仅提高了特征提取的精度，还增强了模型对于复杂图像内容的处理能力。

3.3 Transformer 分支

Transformer 分支的核心是基于视觉变换器（ViT, Vision Transformer）架构，其主要职责在于从输入图像中提炼出全局特征。这一分支专注于揭示图像中的长距离依赖关系，其目的是为 CNN 分支提供的局部特征提取过程提供补充和增强。ViT 架构通过将图像划分为一系列小块（patches），并将这些块转换为一维序列，从而使得模型能够利用自注意力机制来捕获这些块之间的复杂关系。

3.3.1 嵌入块

在深度学习和计算机视觉领域，嵌入块（Patch Embedding）是一种关键操作，它涉及将图像划分为较小的单元（称为 patches），随后将这些单元转换为一系列向量，这些向量随后作为模型的输入。此过程的核心在于将图像的空间特征转换为可以被深度学习模型处理的形式。

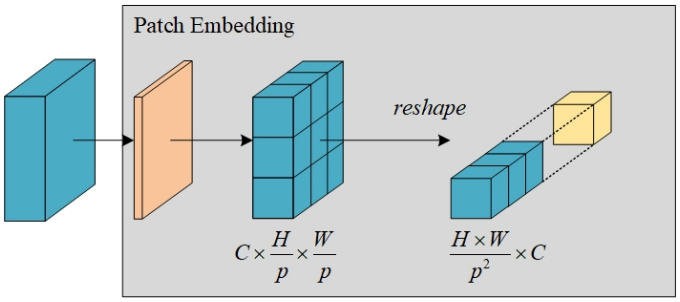


图 2. 嵌入块的一般性过程。

具体来说，对于原始的低光照图像 P_1，首先执行尺寸压缩操作，以适配预训练的视觉变换器（Vision Transformer, ViT）模型的权重。此操作将低光照图像（LLI）转换为一系列的 Patch Embeddings，生成 576 个 D 维的特征向量，其中每个向量对应于一个 16×16 的图像块（patch）。接下来，将 384×384 的特征向量重排列为 576 个 16×16 的图像块，每个图像块随后被展平（flatten）为 256 维的向量。考虑到图像的三通道特性，实际上每个通道被展平为 256 维，从而形成总计 256×3 维的向量 ℓ_1。

此过程的关键在于将二维图像数据转换为一维向量序列，这为后续的 Transformer 模型处理提供了适合的输入格式。通过这种方式，模型能够有效地处理图像数据，捕获重要的空间和颜色特征，从而为深度学习任务提供强大的数据基础。

为了进一步处理这些向量，我们采用一个线性投影（Linear Projection）操作，将 ℓ_1 转换为 D 维的向量。这一转换过程不仅保留了原始图像块的关键信息，而且使其适配于深度学习模型的输入要求。如图3所示，这一过程是将图像的空间特征转换为模型可以有效处理的形式的关键步骤。通过这种方式，模型能够更有效地处理和理解图像内容，从而在低光照图像增强任务中实现更优的性能。

嵌入块的具体过程可以被描述为下列过程：

- 输入图像 $X \in \mathbb{R}^{H \times W \times C}$，其中 H，W 和 C 分别代表图像的高度、宽度和通道数。
- 图像被分割成 N 个大小为 $P \times P$ 的块，每个块被展平并线性投影到 D 维空间，得到嵌入块 $E \in \mathbb{R}^{N \times D}$

图1a 展示了在 Patch Embedding 过程之后，特征 $\mathcal{F}_1$ 经过必要的变换处理，以便与向量 ℓ_1 进行拼接操作。这一步骤是必要的，因为未经处理的特征 $\mathcal{F}_1$ 无法直接与向量 ℓ_1 进行有效的拼接。

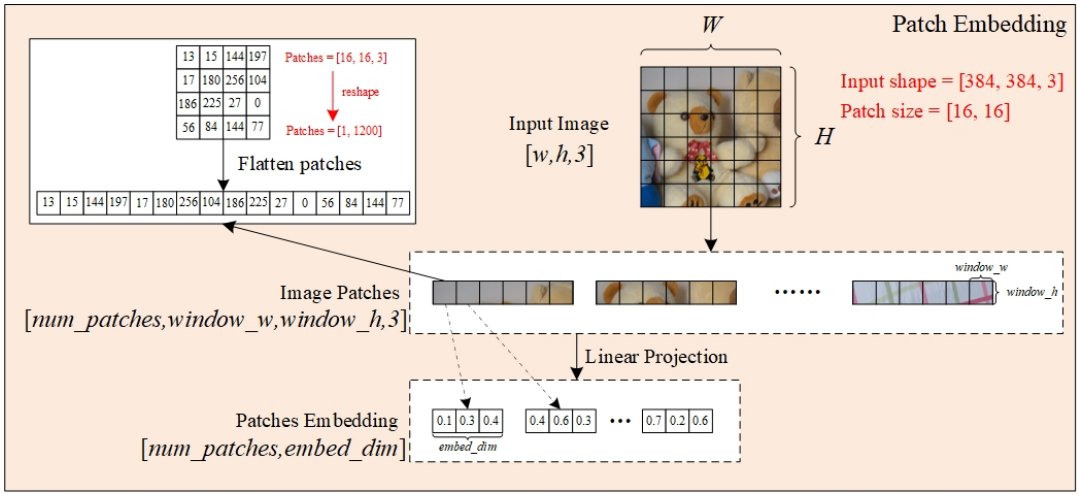


图 3. Transformer 分支中嵌入块的过程。

在本研究中，我们采用了 20×20 的 Patch 尺寸进行 Embedding。然而，选择此尺寸的适宜性依赖于输入图像的尺寸及其包含的细节的重要性。对于那些含有微小但关键特征的图像，较小的 Patch 尺寸可能更为适宜，因为它们能够更精确地捕捉这些细节。相反，较大的 Patch 尺寸可能更适合于捕获图像中的高层次抽象特征。因此，为了在细节捕捉和高层次特征表示之间找到最佳平衡点，我们的后续工作将包括对不同 Patch 尺寸的实验性探索，以确保模型能够有效地处理和理解图像中的关键信息。

3.3.2 位置编码

在 Transformer 中，由于缺乏像卷积神经网络 (CNN) 那样的固有空间结构，需要一种方法来告知模型不同块之间的相对或绝对位置。位置编码就是为了解决这个问题而设计的。

位置编码通常是通过将一个固定的或可学习的编码添加到每个块的 embedding 中来实现的。这个编码是根据块在原始图像中的位置计算得出的。一种常见的方法是使用正弦和余弦函数的组合来生成每个位置的编码。例如，对于位置 pos 和维度 i，编码可以被表示为式2

$$\begin{aligned} \mathrm{PE}(\mathrm{pos},\ 2i) &= \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{D}}}\right) \\ \mathrm{PE}(\mathrm{pos},\ 2i+1) &= \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{D}}}\right) \end{aligned} \tag{2}$$

其中，D 是 embedding 的维度。计算出的 Positional Encoding 被加到每个 patch 的 embedding 上。位置编码为 $PE \in \mathbb{R}^{N \times D}$，Positional Encoding 被描述为 $E_{pos} = E + PE$。这样，即使在多头自注意力机制中，模型也能够利用这些编码来理解和利用 patches 之间的相对位置信息。在 ViT 中，Positional Encoding 对于模型理解图像内容至关重要。它允许模型捕捉到图像中的空间层次结构和对象之间的相对位置关系，这对于图像理解任务来说是非常重要的。

我们采用 timm 库来创建预训练的 ViT 模型，并移除了原始的分类头，在 timm 库中创建的预训练 ViT 模型已经内置了 Positional Encoding 的过程。

3.3.3 Transformer 编码器

在视觉变换器 (ViT) 架构中，Transformer Encoder 是核心组件，负责处理图像的序列化表示。经过位置编码的 Patch Embeddings 被送入 Transformer 的编码器。如图1a 所示，Transformer 编码器主要设计两个主要步骤，包括多头自注意力机制 (Multi-head Self-Attention) 和多层感知器 (Multi-Layer Perceptron)。这些组

件共同工作，以捕获图像中的复杂特征和长距离依赖关系。

其中，多头自注意力机制允许模型在处理每个图像块时同时考虑其他所有块的信息。通过这种方式，模型能够捕获图像中不同区域之间的复杂关系和依赖性。在多头自注意力中，注意力机制被分割成多个“头”，每个头学习图像的不同方面，从而提高了模型的表达能力。MLP 由多个线性层和非线性激活函数组成，它为模型引入了必要的非线性处理能力。这种非线性是处理复杂数据（如图像）时不可或缺的，因为它允许模型学习更加复杂和抽象的特征表示。

对于每个 Transformer 层 l，输入 X_l 经过以下过程：

1) 多头自注意力 (MSA):

$$\mathrm{MSA}(X_l) = [\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_k]W^O \tag{3}$$

其中每个头 (head) 是 $\mathrm{head}_i = \mathrm{Attention}\left(X_l W_i^Q, X_l W_i^K, X_l W_i^V\right)$。Attention 机制通常定义为

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

其中 d_k 是 key 向量的维度。

2) 残差连接和层归一化:

$$X_l' = \mathrm{LayerNorm}\left(X_l + \mathrm{MSA}(X_l)\right) \tag{5}$$

3) 多层感知机: MLP 包含两个全连接层和一个激活函数，即 $\mathrm{MLP}(X_l') = \mathrm{FC}\left(\mathrm{GELU}\left(\mathrm{FC}(X_l')\right)\right)$

4) 第二个残差连接和层归一化：

$$X_{l+1} = \mathrm{LayerNorm}(X_l' + \mathrm{MLP}(X_l')) \tag{6}$$

经过一层 Transformer 编码器处理后，得到的输出 X_L 可以用于下一步的处理。

Transformer 编码器通过多层自注意力和 MLP 的堆叠，有效地处理图像数据，捕获长距离依赖关系，并学习复杂的特征表示。这种架构使得模型能够处理高维度的图像数据，并提取有用的特征，为后续的图像处理任务提供强大的特征支持。

3.3.4 特征耦合单元

在 PACUT 架构中，CNN 分支产生的特征映射与 Transformer 分支的 Patch Embedding 之间的有效融合是一个关键挑战。为了解决这一问题，我们引入了特征耦合单元（FCU），它通过交互式方法将局部特征与全局表示进行连续耦合，从而消除两者之间的不一致性。

为了实现 CNN 特征图和 Transformer Patch Embedding 之间的融合，我们首先通过 1×1 卷积调整 CNN 特征图的通道维度，以匹配 Patch Embedding 的维度。接着，使用下采样模块（如图1b 所示）调整空间尺寸。完成这些步骤后，特征图与 Patch Embedding 进行融合（如图1a 所示）。当特征从 Transformer 分支传输到 CNN 分支时，我们采用上采样（如图 1b 所示）来调整空间比例，并通过 1×1 卷积调整通道尺寸，以便与 CNN 特征图对齐。此外，LayerNorm 和 BatchNorm 被用于特征的正则化。

最后，考虑到特征映射和嵌入块之间存在的语义差异——前者源自局部卷积操作，后者则通过自注意力机制聚合——FCU 的应用至关重要，以填补这一语义鸿沟，确保两种特征的有效融合。

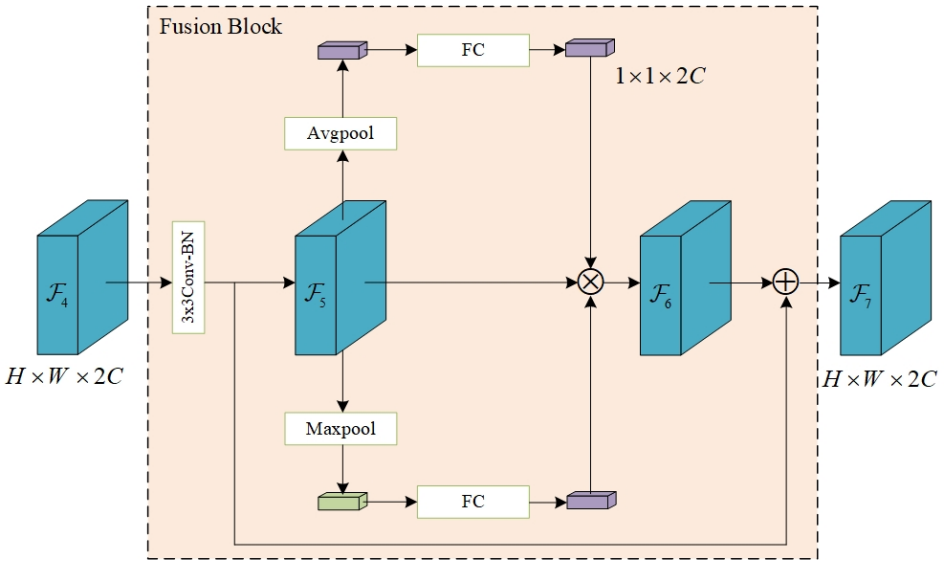


图 4. 融合模块的结构。

3.3.5 特征融合模块

在 U 型网络编码器结构中，浅层特征包含丰富的纹理信息和颜色信息，多层卷积得到的高层特征包含全局信息。对于弱光图像增强，需要尽可能的保留输出结果中的颜色信息。为此，PACUT 结构引入了图像通道来保留具由特定颜色特征的信息，从而更好的恢复暗区信息，防止亮区信息过度曝光。具体来说，在 U 型网络中设计了三个条约链接，将浅层特征发送到后头，以解决颜色信息丢失的问题。为了有效的融合视觉变换器和 U 型特征，提出一个如图4所示的特征融合模块。其输入由两部分组成，即 PACUT 中 CNN 分支得到的特征 $\mathcal{F}_2$ 和 Transformer 分支中块展开（Patch Expanding）后得到的特征 $\mathcal{F}_3$。该模块包含两个步骤。

第一步，将两个尺寸为 $H \times W \times C$ 的特征图连接起来，形成一个尺寸为 $H \times W \times 2C$ 的特征图 $\mathcal{F}_4$。以该特征图为输入，通过 3×3 的卷积核，步长为 1，进一步得到大小为 $H \times W \times 2C$ 的特征图 $\mathcal{F}_5$。这个过程可以有效的捕获颜色信息，表示为

$$\mathcal{F}_5 = \delta\left(f^{3\times3}(\mathcal{F}_4)\right) \tag{7}$$

在第二步中，通过为不同的通道分配适当的权重来重新校准特征映射 $\mathcal{F}_5$。为此，构建一个由两个加权支路和一个短连接组成的模块，如 Fig. 4所示。首先，通过全局平均池化将特征映射 $\mathcal{F}_5$ 压缩成 $1 \times 1 \times 2C$ 个向量；同时，特征映射 $\mathcal{F}_5$ 还需要进行全局最大池化，以尽可能的保留纹理细节。得到的 $1 \times 1 \times 2C$ 向量中的每一个元素表示通道信息。它们的公式为

$$\mathcal{Z}_{GAP} = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} \mathcal{F}_{5C}(i,j) \tag{8}$$

$$\mathcal{Z}_{GMP} = \max\left(\mathcal{F}_5\right) \tag{9}$$

其中 $\mathcal{Z}_{GAP}$ 和 $\mathcal{Z}_{GMP}$ 分别为全局平均池化和全局最大池化得到的特征值，$\mathcal{F}_{5C}$ 为 $\mathcal{F}_5$ 的第 C 个通道的特征。接下来，这两个压缩向量通过两个完全连接的层进行加权，以获得各自的权重向量。最后，通过特征向量乘以相应的权值，得到一个重新校准后，大小为 $H \times W \times 2C$ 的特征图 $\mathcal{F}_7$

$$\mathcal{F}_7 = \sigma\left(W_2\delta(W_1 Z_{GAP})\right) \cdot M \cdot \sigma\left(W_2\delta(W_1 Z_{GMP})\right) + \mathcal{F}_5 \tag{10}$$

其中 W_1 和 W_2 是两个完全连接层的权值。

3.4 损失函数

在本研究中，为了优化图像重建和内容生成的质量，我们提出了一个综合的联合损失函数，用于指导模型生成高质量的增强图像。该联合损失函数综合考虑了多个关键因素，以确保生成图像的质量和真实性。具体地，联合损失函数定义为：

$$\mathcal{L}_{all} = w_0\mathcal{L}_{hub} + w_1\mathcal{L}_{per} + w_2\mathcal{L}_{\text{SSIM}} \tag{11}$$

其中，$\mathcal{L}_{hub}$ 代表休伯（Huber）损失 [41]，$\mathcal{L}_{per}$ 代表感知损失 [42]，而 $\mathcal{L}_{\text{SSIM}}$ 代表结构相似性（SSIM）损失 [43]。这些损失函数的组合旨在平衡图像的像素级精度和感知质量。

Huber 损失（$\mathcal{L}_{hub}$）是一种结合了均方误差和绝对误差的损失函数，用于减少对异常值的敏感性，从而提高模型的鲁棒性。感知损失（$\mathcal{L}_{per}$）则关注于图像的高级特征和内容，通过比较深层网络中的特征表示来评估图像之间的感知差异。结构相似性指数（SSIM）损失（$\mathcal{L}_{\text{SSIM}}$）则用于评估图像的结构、亮度和对比度等方面的相似性，以确保生成图像在视觉上与原始图像保持一致性。

权重 w_0, w_1 和 w_2 分别对应于这三种损失的相对重要性，通过实验调整以达到最佳的图像增强效果。通过这种方式，联合损失函数能够有效地平衡像素级精度和感知质量，从而生成在视觉上更加令人满意的增强图像。

我们设 $w_0 = 1, w_1 = 0.006$ 和 $w_2 = 0.01$。

通过最小化 PACUT 的输出与正常照明图像之间的误差，利用休伯损失指导 PACUT 生成结构完整的增强图像。令 I_{GT} 为真实图像，I_{EI} 为增强图像。休伯损失可以表示为式12

$$\mathcal{L}_{hub}(\delta) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}\frac{1}{2}\|I_{\text{GT}} - I_{\text{EI}}\|_2^2, \|I_{GT} - I_{EI}\| < \delta, \\ \delta\left(\|I_{\text{GT}} - I_{\text{EI}}\|_1 - \frac{1}{2}\delta\right), \|I_{\text{GT}} - I_{\text{EI}}\| \geq \delta.\end{cases} \tag{12}$$

$\mathcal{L}_2$-loss 但容易受离群点的影响，$\mathcal{L}_1$-loss 对离群点更加健壮但是收敛慢，Huber Loss 则是一种将 MSE 与 MAE 结合起来，取两者优点的损失函数，也被称作 Smooth Mean Absolute Error Loss。其原理很简单，就是在误差接近 0 时使用 $\mathcal{L}_2$-loss，误差较大时使用 $\mathcal{L}_1$-loss

然而，休伯损失容易同时平滑增强图像的噪点和细节。为了保留图像的纹理信息，提高图像内容质量，我们引入感知损失，用于比较图像之间的高层次差异，感知损失可以表示为式13。

$$\mathcal{L}_{feat}^{\phi,j}(I_{GT}, I_{EI}) = \frac{1}{C_j H_j W_j}\|\phi_j(I_{GT}) - \phi_j(I_{EI})\|_2^2 \tag{13}$$

其中 I_{GT} 为输出图像，I_{EI} 为目标图像，ϕ 为损失网络。$\phi_j(x)$ 为处理图像 x 时损失网络 ϕ 的第 j 层的激活情况，如果 j 是一个卷积层，那么 $\phi_j(x)$ 将是形状 $C_j \times H_j \times W_j$ 的特征映射，特征重建损失是特征表示之间的欧式距离。

此外，为了使得恢复图像的亮度、对比度和结构与真实图像更加接近和同时细化恢复图像的细节，我们引入了结构相似性损失，结构相似性损失可以表示为式

每个像素 p 的SSIM被定义为式16

$$\begin{aligned}\mathrm{SSIM}(p) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \\ &= l(p) \cdot cs(p)\end{aligned} \tag{14}$$

其中省略了均值和标准偏差对像素 p 的依赖性，均值和标准差是用标准偏差为 σ_G, G_{σ_G}

$$\begin{aligned}&\varepsilon(p) = 1 - \mathrm{SSIM}(p) : \\ &\mathcal{L}^{\mathrm{SSIM}}(P) = \frac{1}{N}\sum_{p\in P} 1 - \mathrm{SSIM}(p).\end{aligned} \tag{15}$$

eq. 14表明 SSIM(p) 需要关注像素 p 的邻域，这个领域的大小取决于 G_{σ_G}，网络的卷积性质允许我们将 SSIM 损失写为

$$\mathcal{L}^{\mathrm{SSIM}}(P) = 1 - \mathrm{SSIM}(\tilde{p}). \tag{16}$$

其中 $\tilde{p}$ 是 P 的中心像素。

4 总结

本研究提出了 PACUT 算法，用于有效增强弱光图像。该算法采用并行结构，融合了 CNN 分支和 Transformer 分支，以有效提取特征图的长距离和短距离特征。具体而言，CNN 分支采用改进的 U 型网络架构，包括编码器子网络和解码器子网络。编码器子网络由多层卷积网络组成，每层集成了空间注意力机制和通道注意力机制。Transformer 分支以视觉变换器为核心，利用其自注意力机制提取图像的全局特征，为 CNN 分支提供补充和增强。这些设计不仅增强了弱光图像的信息，还有效避免了增强过度和不足。然而，模型可能过于偏向短距离特征，结构上 CNN 分支较为冗余，而 Transformer 分支相对瘦小。未来的工作中，将深入探索 Transformer 编码器和 CNN 结构的并行架构，借鉴 Conformer 模型中多个 Transformer 编码器块以捕获更多长距离特征。

参考文献

[1] Shih-Chia Huang, Fan-Chieh Cheng, and Yi-Sheng Chiu. Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE transactions on image processing*, 22(3):1032–1041, 2012.

[2] Yu Wang, Qian Chen, and Baeomin Zhang. Image enhancement based on equal area dualistic sub-image histogram equalization method. *IEEE transactions on Consumer Electronics*, 45(1):68–75, 1999.

[3] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. Properties and performance of a center/surround retinex. *IEEE transactions on image processing*, 6(3):451–462, 1997.

[4] Edwin H Land. The retinex. In *Ciba Foundation Symposium-Colour Vision: Physiology and Experimental Psychology*, pages 217–227. Wiley Online Library, 1965.

[5] Edwin H Land. The retinex theory of color vision. *Scientific american*, 237(6):108–129, 1977.

[6] Jin-Hwan Kim, Jae-Young Sim, and Chang-Su Kim. Single image dehazing based on contrast enhancement. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1273–1276. IEEE, 2011.

[7] Lin Li, Ronggang Wang, Wenmin Wang, and Wen Gao. A low-light image enhancement method for both denoising and contrast enlarging. In *2015 IEEE international conference on image processing (ICIP)*, pages 3730–3734. IEEE, 2015.

[8] T-L Ji, Malur K Sundareshan, and Hans Roehrig. Adaptive image contrast enhancement based on human visual properties. *IEEE transactions on medical imaging*, 13(4):573–586, 1994.

[9] Siu Fong Tan and Nor Ashidi Mat Isa. Exposure based multi-histogram equalization contrast enhancement for non-uniform illumination images. *Ieee Access*, 7:70842–70861, 2019.

[10] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2782–2790, 2016.

[11] Li Tao, Chuang Zhu, Guoqing Xiang, Yuan Li, Huizhu Jia, and Xiaodong Xie. Llcnn: A convolutional neural network for low-light image enhancement. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.

[12] Lei Wang, Guangtao Fu, Zhuqing Jiang, Guodong Ju, and Aidong Men. Low-light image enhancement with attention and multi-level feature fusion. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 276–281. IEEE, 2019.

[13] Chenglin Yang, Siyuan Qiao, Adam Kortylewski, and Alan Yuille. Locally enhanced self-attention: Combining self-attention and convolution as local and context terms. *arXiv preprint arXiv:2107.05637*, 2021.

[14] Cheng Zhang, Qingsen Yan, Yu Zhu, Xianjun Li, Jinqiu Sun, and Yanning Zhang. Attention-based network for low-light image enhancement. In *2020 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2020.

[15] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European conference on computer vision (ECCV)*, pages 517–532, 2018.

[16] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020.

[17] Xiaoxiao Li, Xiaopeng Guo, Liye Mei, Mingyu Shang, Jie Gao, Maojing Shu, and Xiang Wang. Visual perception model for rapid and adaptive low-light image enhancement. *arXiv preprint arXiv:2005.07343*, 2020.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[20] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method, 2022.

[21] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration, 2021.

[22] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration, 2023.

[23] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 367–376, 2021.

[24] Anil K Jain and Farshid Farrokhnia. Unsupervised texture segmentation using gabor filters. *Pattern recognition*, 24(12):1167–1186, 1991.

[25] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[26] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[27] Dimitri A Lisin, Marwan A Mattar, Matthew B Blaschko, Erik G Learned-Miller, and Mark C Benfield. Combining local and global image features for object class recognition. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)-Workshops*, pages 47–47. IEEE, 2005.

[28] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.

[29] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020.

[30] Bochang Moon, Jong Yun Jun, JongHyeob Lee, Kunho Kim, Toshiya Hachisuka, and Sung-Eui Yoon. Robust image denoising using a virtual flash image for monte carlo ray tracing. In *Computer Graphics Forum*, volume 32, pages 139–151. Wiley Online Library, 2013.

[31] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.

[32] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.

[33] Yu Zhang, Xiaoguang Di, Bin Zhang, Qingyan Li, Shiyu Yan, and Chunhui Wang. Self-supervised low light image enhancement and denoising. *arXiv preprint arXiv:2103.00832*, 2021.

[34] Qingxu Fu, Xiaoguang Di, and Yu Zhang. Learning an adaptive model for extreme low-light raw image processing. *IET Image Processing*, 14(14):3433–3443, 2020.

[35] Qing-Long Zhang and Yu-Bin Yang. Sa-net: Shuffle attention for deep convolutional neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2235–2239. IEEE, 2021.

[36] Yang Zhang and TO Aydın. Deep hdr estimation with generative detail reconstruction. In *Computer graphics forum*, volume 40, pages 179–190. Wiley Online Library, 2021.

[37] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[39] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022.

[40] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[41] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[42] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

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